Bayes-verkkojen mahdollisuudet

Petri Myllymäki
Henry Tirri

TEKNOLOGIAN KEHITTÄMISKESKUS

Teknologiakatsaus 58/98
Helsinki 1998

Kilpailukykyä teknologiasta

Teknologian kehittämiskeskus Tekes tarjoaa rahoitusta ja asiantuntijapalveluja kansainvälisesti kilpailukykyisten tuotteiden ja tuotantomenetelmien kehittämiseen. Tekesillä on vuosittain käytettävissä avustuksina ja lainoina noin puolitoista miljardia markkaa teknologian kehityshankkeisiin.

Teknologiaohjelmien avulla maahamme luodaan uutta teknologiaosaamista yritysten, tutkimuslaitosten ja korkeakoulujen yhteistyönä. Ohjelmien tavoitteena on nostaa teknologista kilpailukykyämme tulevaisuuden keskeisillä teollisuuden toimialoilla. Tällä hetkellä Tekesillä on käynnissä noin 50 teknologiaohjelmaa.

ISBN 951-53-1391-9
ISSN 0782-5420

Kansi: LM & CO
Paino: Paino-Center oy, Sipoo 1998

Esipuhe

Elektroniikan ja tietokonetekniikan kehityksen myötä taloudellisesti käytettävissä oleva laskenta- ja tietojenkäsittelykapasiteetti kasvaa jatkuvasti. Lisääntyvä laskentateho on tehnyt mahdolliseksi täysin uudenlaisen menetelmäkehityksen, jonka kohteena ovat entistä suuremmat tietomassat tai mutkikkuudeltaan ja päätösvaihtoehtomääriltään aiemmin mahdottomat tehtävät. Pelkän laskentatehon hyödyntämisen ohella uusilla menetelmillä tavoitellaan mallien ja järjestelmien kykyä oppia ja toimia kokemustietoon pohjautuen "älykkäästi" uusissa tilanteissa.

Menetelmätutkimus on suuntautunut moniin eri menetelmäperheisiin, jotka ovat osittain toisilleen vaihtoehtoisia mutta usein myös toisiaan täydentäviä. Tutkimusalueesta käytetään tavallisesti nimeä "Laskennallinen älykkyys" (Computational Intelligence). Monissa tapauksissa uusilla menetelmillä on voitu parantaa teollisuuden tuotteiden, prosessien ja järjestelmien ominaisuuksia tai toimintaa ratkaisevasti.

Tekes on julkaissut laskennallisen älykkyyden päämenetelmiin liittyviä, menetelmien soveltamimahdollisuuksiin keskittyviä katsauksia sumean logiikan, neuraalilaskennan ja geneettisten algoritmien alueelta. Lisääntynyt kiinnostus todennäköisyysteoreettiseen mallintamiseen ja erityisesti Bayesverkkojen monet menestyksekkäät sovellukset ovat luoneet tarpeen käsilläolevan raportin laadinnalle.

Raportin ovat Tekesin toimeksiannosta kirjoittaneet FT Henry Tirri ja FT Petri Myllymäki Helsingin yliopiston tietojenkäsittelytieteen laitokselta. Raportin alkusanat on laatinut professori Erkki Oja Teknillisestä korkeakoulusta. Raportin laadintaa ohjasi Tekesin "Oppivat ja älykkäät järjestelmät"-teknologiaohjelman johtoryhmä yhdyshenkilönään Matti Sihto Tekesistä.

Tekes kiittää lämpimästi raportin kirjoittajia sekä sen laadintaan eri tavoin osallistuneita henkilöitä. Tekes toivoo, että raportti osaltaan toimii innoituksen ja virikkeiden lähteenä yritysten arvioidessa uusien älykkäiden mallitus- ja laskentamenetelmien ja erityisesti bayesiläisen mallintamisen soveltuvuutta oman liiketoiminnan avainalueilla.

Helsingissä huhtikuussa 1998

Teknologian kehittämiskeskus Tekes

Alkusanat

Oppivien ja älykkäiden järjestelmien tutkimus on viime aikoina organisoitunut kokonaisuudeksi, josta käytetään myös nimiä *laskennallinen älykkyys* (computational intelligence) ja *soft computing*. Ajatuksena on, että vaikeissa päättelyongelmissa korvataan eksplisiittisten mallien ja ihmistietämyksen puute laskennallisilla (numeerisilla) malleilla, jotka sovitetaan tutkittavaan kohteeseen suurelta osin mittaustietoja hyväksikäyttäen. Tämä tietysti vain silloin, kun paremman puutteessa on pakko turvautua dataohjattuihin empiirisiin malleihin. Asiantuntijalla on kyllä näkemys kohteensa (esimerkiksi prosessi tai laite) toiminnasta, mutta tarkkojen ja ennustusvoimaisten mallien aikaansaamiseksi voi olla pakko täyttää puuttuvia osia soft computing-malleilla. Lopputulos on silloin monista osista koostuva *hybridimalli*.

Soft computing-alan valtavirrat ovat neuroverkot ja sumeat järjestelmät. Niillä on eniten sovelluksia niin Suomessa kuin muuallakin, alat ovat jo kypsyneet valmiiden ohjelmistopakettien ja hyvätasoisten oppikirjojen tasolle, ja runsaasti käyttökokemusta on olemassa neuroverkkojen tai sumeiden sääntöjen — tai molempien — soveltamisesta laajaan kirjoon käytännön ongelmia. Näitä kahta alaa sovelluksineen on selostettu TEKESin raporteissa “Sumean logiikan mahdollisuudet” (1993) ja “Neurolaskennan mahdollisuudet” (1994). Laskennalliseen älykkyyteen kuuluu kuitenkin muitakin metodiikkoja. Tärkeimmät ovat *evoluutiolaskenta*, etenkin geneettiset optimointialgoritmit, sekä *Bayes-verkot*, jota nyt käsillä oleva raportti käsittelee. Evoluutiolaskentaa tarkastellaan TEKESin raportissa “Geneettisten algoritmien mahdollisuudet” (1998). Yhdessä nämä neljä raporttia kattavat soft computing-alueen nykyiset valtavirtaukset ja tarjoavat yrityksille ja tutkijoille näkymiä eri menetelmien soveltamismahdollisuuksista.

Voi tietenkin kysyä, onko yleensä mielekästä puhua tieteenalasta nimeltä laskennallinen älykkyys tai soft computing, joka määritellään tiettynä suljettuna menetelmäjoukkona. Varmaa ainakin on, että ala on voimakkaassa muutostilassa ja uusia teknologioita ja menetelmiä epäilemättä syntyy lähitulevaisuudessa, joten kuva on koko ajan muutostilassa. Eri menetelmien lähtökohdat ja sovellusmahdollisuudet ovat myös osittain erilaisia. Mielestäni — ja monen muunkin alan tutkijan mielestä — on järkevää niputtaa nämä uudet alat omaksi metodologiakseen tai tieteenalakseen, koska niillä on kuitenkin joukko yhteisiä piirteitä, jotka erottavat ne perinteisestä komputoinnista. Ensinnäkin niiden välillä on yhteyksiä: esimerkiksi neuroverkkoja käytetään virittämään sumeita järjestelmiä, geneettisillä algoritmeilla voi tehokkaasti optimoida neuroverkkojen rakennetta, ja Bayes-verkot antavat matemaattisesti perustellun ja siksi voimakkaan mallin samoihin päättelyongelmiin, joita ratkotaan neuroverkoilla tai sumeilla järjestelmillä. Toiseksi, ja ehkä vieläkin tärkeämpänä syynä, on menetelmien perusidea ja lähtökohta — enemmän tai vähemmän biologisista tai kognitiivisista prosesseista johdetaan toimintamalleja, jotka yksinkertaistettuina ja muunneltuina sovitetaan oppiviksi ja älykkäiksi ratkaisuiksi käytännön ongelmiin. Tärkeää on nimen omaan laskennallisuuden korostaminen: oppivat ja älykkäät järjestelmät, ainakin realistisissa isoissa ongelmissa, toimivat tehokkaidenkin tietokoneiden äärirajoilla, eikä niitä olisi voinut ajatellakaan käytännön menetelminä vielä kymmenenkään vuotta sitten. Tietokoneiden tehon edelleen kasvaessa tulemme varmasti näkemään uusia laskennallisesti älykkäitä menetelmiä, jotka aina käyttävät kaiken saatavilla olevan laskentatehon yhä parempien ja parempien päättelymallien rakentamiseksi.

Mitä soveltaja tästä alojen niputtamisesta sitten hyötyy? Suoranaisesti ei paljoakaan. Kuitenkin ne valmiit menetelmät, joita hän ongelmiinsa kokeilee tai käyttää, ovat syntyneet tutkimuksen tuloksena, jossa erityyppisten oppivien ja älykkäiden järjestelmien vuorovaikutus on antanut uusia ideoita ja johtanut parempiin tuloksiin.

Bayes-verkot ovat päättelyjärjestelmiä, jotka nojautuvat vanhaan ja koeteltuun todennäköisyyslaskentaan, etenkin ns. bayesiläiseen päättelyyn. Lyhyesti sanottuna tämä tarkoittaa, että kaikille mallituksen elementeille — itse malliarkkitehtuurille, sovitettaville parametreille ja sovitukseen käytettävälle datalle — oletetaan todennäköisyysjakaumat, ja valitsemme sen mallin ja ne parametrit, jotka ovat kaikkein todennäköisimmät ottaen huomioon saamamme mittausdatan. Ajatus on elegantti ja sallii asiantuntijan tietämyksen käyttämisen, koska hän voi kokemuksensa perusteella valita erilaisten mallien ja parametrien etukäteistodennäköisyydet (priorit). Vaikeutena toisaalta on, että todennäköisyysformalismi eleganttisuudessaan voi olla liiankin yleinen — suurta joukkoa tarvittavia todennäköisyysjakaumia on hankala estimoida ja niitä yhdistäviä hyvin moniulotteisia integraaleja on vaikea laskea. Tällöin joudutaan voimakkaisiin yksinkertaistuksiin, jotka murentavat mallien tyylikästä formalismia. Kuitenkin yksinkertaistetutkin mallit ovat monissa käytännön ongelmissa, esimerkiksi luokittelussa, osoittautuneet hyvin tehokkaiksi ja toimiviksi.

Erkki Oja

Tiivistelmä

Oppivien ja älykkäiden järjestelmien tutkimuksen tavoitteena on rakentaa sellaisia tietokoneohjelmana toteutettuja sovellusalueen yksinkertaistettuja malleja, joita voidaan käyttää hyväksi erilaisissa ongelmanratkaisutilanteissa. Bayesiläisessä mallinnuksessa käytetään kaikkien tällaisen järjestelmän rakentamisessa esiintyvien ongelmien ratkaisemisessa samaa teoreettista, todennäköisyyslaskentaan perustuvaa kehikkoa. Jotta tätä eleganttia, täysin yleistä lähestymistapaa voitaisiin soveltaa käytännössä, on käytettävien mallien ominaisuuksia rajoitettava jollakin joukolla ongelmakenttää koskevia perusoletuksia. Aikaisemmin on usein arvioitu tarvittavien oletusten tällöin yksinkertaistavan malleja niin, että ne tulevat hyödyttömiksi käytännössä. Bayes-verkkoina tunnetun malliperheen kehittäminen joitakin vuosia sitten, ja etenkin lukuisat Bayes-verkkotekniikoihin perustuvat viimeaikaiset menestykselliset sovellukset, ovat kuitenkin osoittaneet tällaiset arviot vääriksi. Bayes-verkkoteknologiaan perustuvien sovellusten kehittäminen onkin muodostunut yhdeksi oppivien ja älykkäiden järjestelmien tutkimuksen avainkysymyksistä.

Tässä raportissa kuvataan bayesiläisen mallinnuksen yleiset periaatteet, ja esitellään lyhyesti tärkeimmät Bayes-verkkomalleihin liittyvät käsitteet. Raportissa tarkastellaan myös bayesiläisen mallinnuksen suhdetta muihin oppivien ja älykkäiden järjestelmien rakentamisessa käytettyihin menetelmiin. Raportti sisältää katsauksen Bayes-verkkojen teolliseen soveltamiseen, lukuisia sovellusesimerkkejä, ja tietoa Bayes-verkkosovellusten kehittämiseen tarkoitetuista ohjelmistoista, alan kirjallisuudesta sekä tutkimusryhmistä.

Executive Summary

Research on adaptive and intelligent systems aims at building problem domain models, implemented as computer software, that can be used in various problem solving situations. In Bayesian modeling, all the related problems in building such models are solved within the same theoretical framework based on probability theory. In order to be able to apply this completely general, theoretically elegant approach in practice, the set of possible models has to be constrained by some basic assumptions on the problem domain. It has sometimes been argued that the required assumptions simplify models to the point where they become useless for practical applications. However, this argumentation has been disproved by recent theoretical developments leading to the use of what are known as Bayesian networks, and by several successful applications based on this technology. The development of fielded Bayesian network applications has become one of the key challenges in the research on adaptive and intelligent systems.

This report discusses briefly the general principles of Bayesian modeling, and introduces the most important concepts related to Bayesian network models. The relationships between the Bayesian approach and other methods used for building adaptive and intelligent systems are also explored. The report includes an overview of industrial applications of Bayesian networks, and a survey of Bayesian network software, literature and research activities.

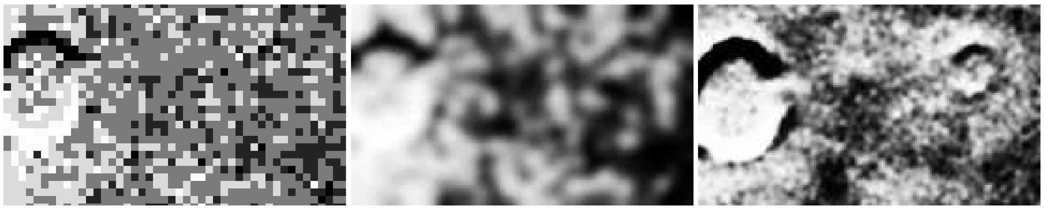
Sisältö

Luku 1

Johdanto

Motivointia ja sovellusesimerkkejä: kuvankäsittelystä älykkäisiin agentteihin. Raportin aihepiiri ja kohderyhmät. Raportin yleisrakenne.

Bayesiläinen mallintaminen on lahjakkaan amatöörimatemaatikon Thomas Bayesin (1702–1761) mukaan nimensä saanut todennäköisyyslaskentaan perustuva yleinen lähestymistapa monimutkaisissa järjestelmissä esiintyvän epätäsmällisen informaation hallitsemiseksi. Yksinkertaisimmillaan bayesiläisen mallinnuksen voi ymmärtää menetelmänä, joka mallintaa sitä kuinka käsityksemme jostakin asiasta (uskomuksemme asian todenperäisyydestä) muuttuu, kun saamme asiaan liittyvää uutta tietoa. Bayesiläinen mallintaminen on kunnioitettavasta iästään huolimatta yksi nopeimmin kehittyviä ja kasvavia mallinnusmenetelmiä, ja sen soveltamisella on kauaskantoisia ja merkittäviä vaikutuksia mm. lääketieteessä, teollisessa mallintamisessa, ekonometriassa ja yhteiskuntatieteissä [92]. Yksi syy bayesiläisen mallintamisen nopeasti kasvavaan suosioon on vasta joitakin vuosia sitten nykyisen muotonsa saanut *Bayes-verkkoina (Bayesian networks)* tunnettu malliperhe, jonka myötä poistui monia bayesiläiseen mallinnukseen aikaisemmin liittyneitä käytännön ongelmia. Vuoden 1997 IJCAI-konferenssissa (maailman johtava tekoälykonferenssi) Bayes-verkkoteknologioihin perustuvien sovellusten kehittäminen todettiin yhdeksi tekoälytutkimuksen avainkysymyksistä tulevaisuudessa [38]. Havainnollistaaksemme bayesiläisen mallintamisen monipuolisia sovellusmahdollisuuksia esittelemme seuraavassa lyhyesti joukon eri sovellusaloilla (kuvankäsittely, tiedon koodaus ja tiivistys, ohjelmistotekniikka) viime vuosina kehitettyjä merkittäviä teknologioita, jotka kaikki — erilaisista sovellusalueistaan huolimatta — perustuvat bayesiläiseen mallintamiseen. Lisää sovellusesimerkkejä löytyy luvusta 5.1.

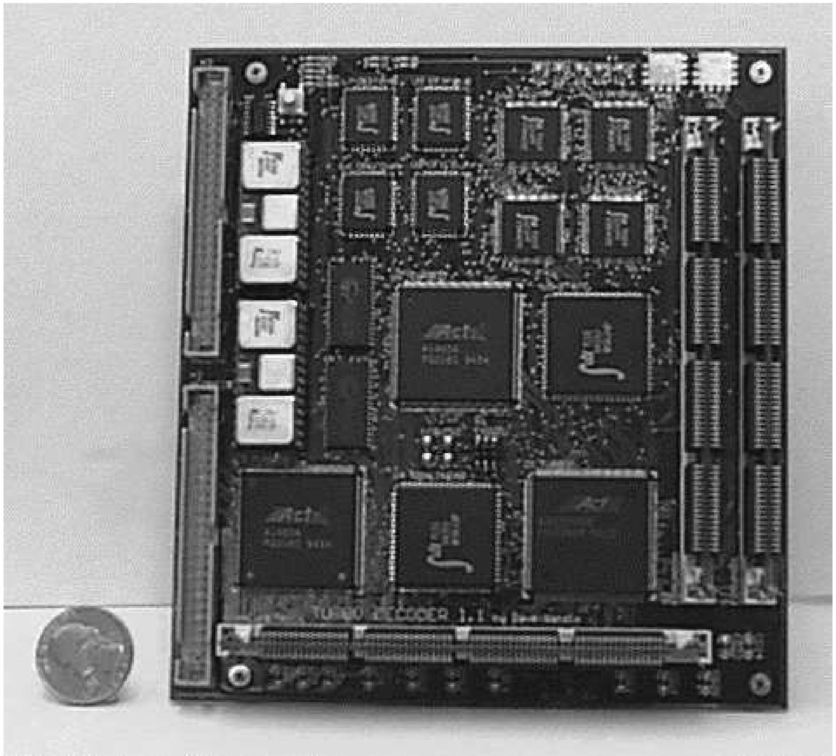


Kuva 1.1: Vasemmalla: yksi Viking-luotaimen Marsin pinnasta ottamista 24 alkuperäisestä kuvasta. Keskellä: kuva perinteisillä kohinansuodatustekniikoilla käsiteltynä. Oikealla: 24 kuvan aineistosta yhdistetty superresoluutiokuva.

Superresoluutio. Viking-avaruusluotain tuottaa useita kuvia Mars-planeetasta. Kuvat on otettu samankaltaisissa valaistusolosuhteissa ja samoista kohteista. Tällöin instrumenttien pienet orientaatio- ja kohdennuserot tuottavat kuitenkin hieman erilaisia kuvia samasta kohteesta. Tällaisesta kuvajoukosta voidaan yhdistelemällä tuottaa tietokoneella superresoluutiokuva: kuva, joka on tarkempi kuin millään kohinansuodatustekniikalla yksittäisestä kuvasta saatu kuva (katso kuva 1.1). National Aeronautics & Space Administration (NASA) käyttää tätä useaan kuvaan perustuvaa superresoluutiotekniikkaa avaruusluotainten tuottamien kohinaisten kuvien parantamiseen, ja tällaisella kuvankäsittelytekniikalla on merkittäviä sovelluksia mm. lääketieteessä ja biotieteissä.

Turbo-koodaus. Tietoliikenteen määrän jatkuvasti kasvaessa tulee tiedon tehokkaasta koodaamisesta ja tiivistämisestä yhä keskeisempi tietotekninen ongelma. Vuonna 1993 kehitettyä turbo-koodausta pidetään koodausteorian merkittävimpänä keksintönä vuosikymmeniin. Turbo-koodaus soveltuu tiedon välittämiseen kohinaista tiedonsiirtokanavaa käyttäen, ja sen tehokkuus tiedon välityksessä on moninverroin parempi aiempiin menetelmiin verrattuna. Turbo-koodit ovat erityisen merkittävä edistysaskel avaruusteknologiassa, jossa ne ylittävät mm. Voyager-luotaimen käyttämän ns. katenointikoodin (concatenated code) moninkertaisesti. Turbo-koodaus osoittaa myös kuinka nopeasti teoreettinen tutkimus siirtyy käytännön sovellukseksi: erilaisia turbo-koodeihin perustuvia piirejä valmistavat jo useat yhtiöt, kuten esimerkiksi Efficient Channel Coding Inc. (katso kuva 1.2).

Älykkäät agentit. Nykyiset tekstin ja kuvien käsittelyyn kehitetyt ohjelmistot ovat laajoja ohjelmakokonaisuuksia, joiden monilukuisten piirteiden tehokas käyttäminen on vaikeaa ja käytön oppiminen hidasta. Microsoft Office'97-ohjelmistossa käyttäjän apuna ovat *älykkäät agentit* (Office Assistants), jotka itsenäisinä moduuleina tarkkailevat käyttäjän toimintaa ja tarjoavat apuaan käyttäjälle mm. monimutkaisten muotoilujen toteuttamisessa

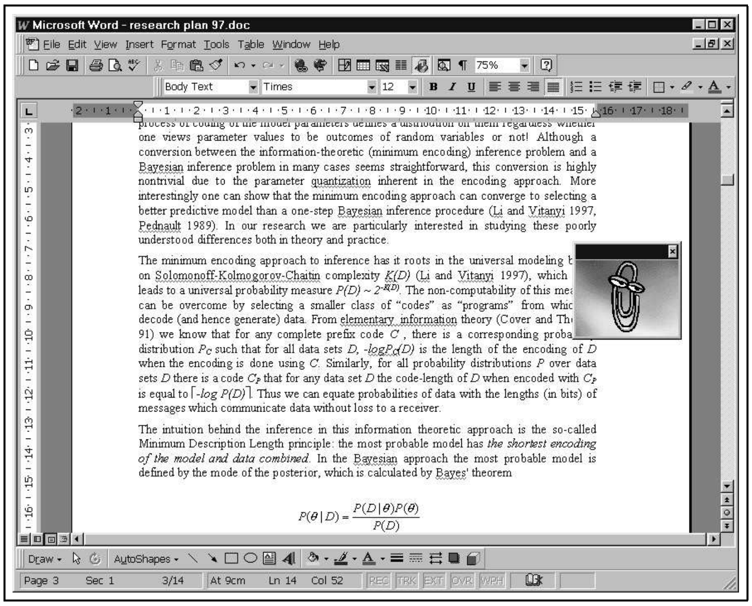


Kuva 1.2: Turbo-koodauksen toteuttava piiri.

(kuva 1.3). Tällainen teknologia edellyttää varsin monimutkaista ohjelman toimintojen ja käyttäjän tarpeiden yhteismallinnusta, jotta agentti pystyy tarjoamaan käyttäjälle triviaalien kommenttien sijaan älykkäitä ja hyödyllisiä ohjeita. Älykäs agenttiteknologia on muodostumassa erityisesti verkottumisen myötä yhdeksi tärkeimmistä uusista ohjelmistosuuntauksista, ja sen keskeisimpiä sovellusalueita ovat tiedon haussa avustaminen, liikkuva tietojenkäsittely sekä ohjelmistojen ja laitteiden konfiguroinnissa avustaminen.

Tässä raportissa käsitellään bayesiläistä mallintamista keskittyen erityisesti Bayes-verkkomalleihin, jotka ovat mm. edellä esitetyn turbo-koodauksen ja Office Assistant-teknologian perustana. Raportti on suunnattu suomalaisen teollisuuden edustajille, mutta alansa ensimmäisenä suomenkielisenä yleistajuisena esityksenä löytänee lukijoita myös oppilaitoksista ja julkishallinnosta. Tavoitteensa takia esityksen näkökulma on käytännöllinen: raportissa keskitytään Bayes-verkkojen teollisten sovellusmahdollisuuksien esittelemiseen, ei niinkään Bayes-verkkojen teorian yksityiskohtien selvittämiseen.

Raportin aluksi luvussa 2 tarkastellaan bayesiläistä lähestymistapaa yleisesti oppivien ja älykkäiden järjestelmien rakentamisen kannalta, ja sitä kuin-



Kuva 1.3: Microsoftin Office Assistant-agentti.

ka lähestymistapaa voidaan soveltaa mallien konstruointiin ja epätäydellisen tiedon varassa päättelyyyn. Vaikka raporttia muuten on mahdollista lukea valikoiden, tässä luvussa esiteltävät käsitteet ovat keskeisiä myöhempien lukujen kannalta.

Luvussa 3 arvioidaan bayesiläisen mallinnuksen suhdetta muihin oppivien ja älykkäiden järjestelmien yhteydessä esiintyviin menetelmiin kuten esimerkkeihin perustuvaan päättelyyn (Case-Based Reasoning, CBR), ja tarkastellaan lyhyesti bayesiläiseen mallintamiseen liittyvää informaatioteoreettista tulkintaa. Tämä luku on erityisen kiinnostava niille, jotka haluavat vastauksia yleisesti esitettyihin bayesiläistä mallintamista koskeviin kysymyksiin.

Luku 4 on raportin keskeisin osa, ja tarjoaa katsauksen erilaisiin Bayes-verkkomalleihin ja niihin liittyviin käsitteisiin. Luku sisältää runsaasti yksityiskohtaisia esimerkkejä siitä, miten erityyppisiä ongelmia voidaan mallintaa Bayes-verkoilla. Näiden perusesimerkkien lisäksi luvussa 5 tarkastellaan Bayes-verkkojen teollisia sovelluksia ja Bayes-verkkojen sovellusten kehittämiseen suunniteltuja kaupallisia ohjelmistoja. Tämä luku on tarkoitettu antamaan tietoa ja viitteitä soveltajille niistä sovellusalueista, joilla Bayes-verkkoja voidaan käyttää. Bayes-verkkoteknologian kokeilusta kiinnostuneille luku tarjoaa tietoa saatavilla olevista ohjelmistotyökaluista, joista useista on olemassa ilmainen demonstraatioversio.

On syytä korostaa, että tätä raporttia ei ole tarkoitettu käytettäväksi yksinään alan oppikirjana. Esityksessä on pyritty antamaan Bayes-verkkoihin liittyvästä teoriasta ja siihen liittyvistä käsitteistä ja menetelmistä intuitiivisesti mahdollisimman ymmärrettävä yleiskuva, paikoitellen matemaattisesta täsmällisyydestä joustaen. Bayes-verkkoteorian ja teollisuudessa sovellettavien algoritmien täsmälliset yksityiskohdat ovat löydettävissä tutustumalla luvussa 6 esiteltyihin alueen keskeisiin tietolähteisiin. Raportissa annetaan myös suuri joukko URL-osoitteita, joista löytyy lisämateriaalia, kirjallisuusluetteloita, sovellusesimerkkien kuvauksia, ilmaisohjelmia sekä yhtiöiden, järjestöjen ja tutkijoiden kotisivuja. Kyseiset Internet-osoitteet löytyvät myös kirjoittajien tutkimusryhmän kotisivulta[1].

[1] URL: http://www.cs.Helsinki.FI/research/cosco

Luku 2

Bayesiläinen mallinnus

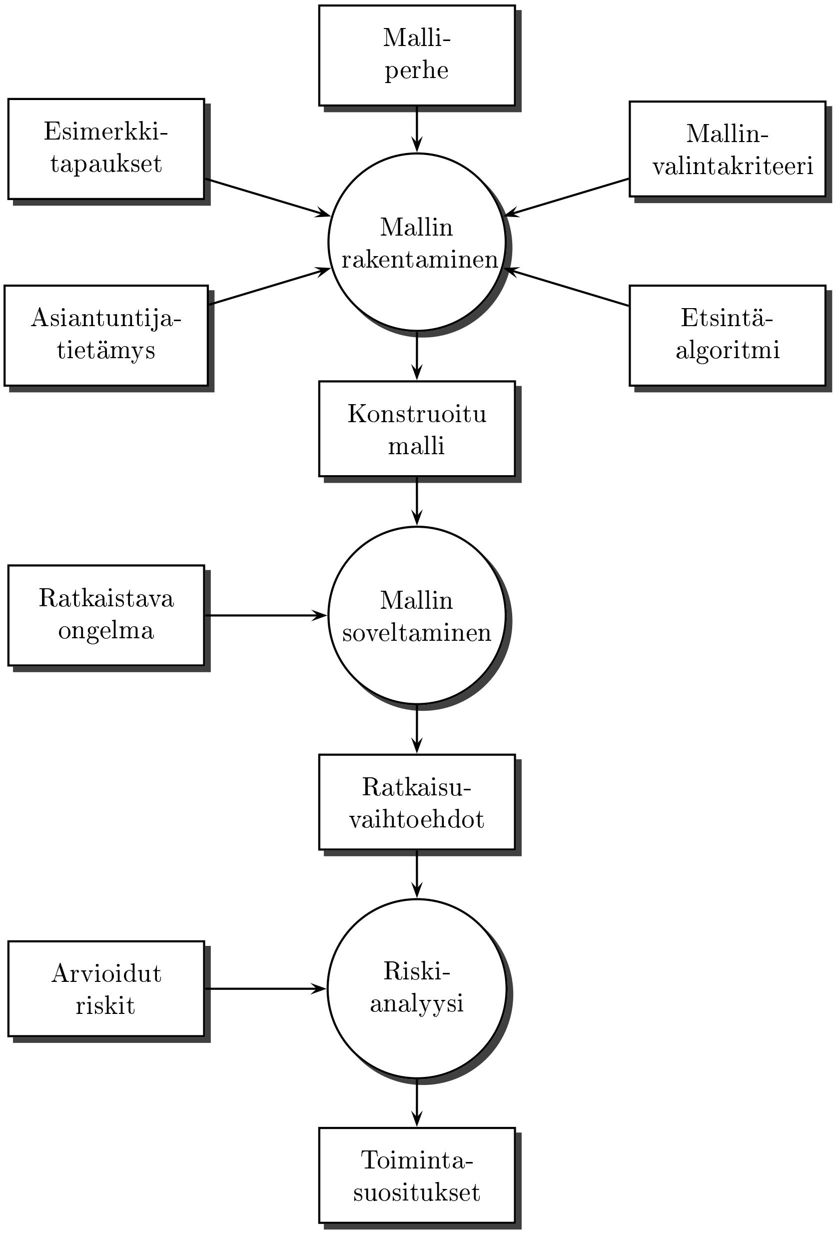
Oppivien ja älykkäiden järjestelmien yleinen rakenne — mallien oppiminen ja niiden soveltaminen. Mallien oppiminen bayesiläisittäin: mallien rakenteen ja parametrien oppimisen erottaminen. Ylioppiminen ja Occamin partaveitsi. Oppimisen ja taustatietämyksen yhdistäminen. Bayesiläisten mallien soveltaminen.

2.1 Bayesiläinen mallintaminen oppivissa ja älykkäissä järjestelmissä

Oppivissa ja älykkäissä järjestelmissä pyritään rakentamaan käsiteltävästä sovellusalueesta yksinkertaistettu matemaattinen malli siten, että mallin toteuttavaa tietokoneohjelmaa voidaan käyttää avuksi erilaisissa ongelmanratkaisutilanteissa. Kuvassa 2.1 on tällaisen järjestelmän rakenne jaettu kolmeen komponenttiin: mallien rakentaminen, mallien soveltaminen ja riskianalyysi. 'Mallien soveltaminen' on tässä yhteydessä intuitiivisesti määritelty käsite, joka kattaa mm. seuraavat ongelmakentät:

Tilastollinen päättely ja ennustaminen:

Rakennettu malli on ongelmakentän matemaattinen kuvaus, ja mallinsovellusmoduulille annettua syötettä voidaan tässä tapauksessa ajatella mallinnetussa maailmassa esiintyvän tilanteen osittaisena kuvauksena. Tilastollisessa päättelyssä ja ennustamisessa on päämääränä selvittää mitä annetuista tiedoista seuraa annetun mallin puitteissa, tai mitkä ovat annettuun tilanteeseen johtaneet syyt. Tähän ryhmään kuuluvat muun muassa luokittelu-, diagnosointi- ja aikasarjaongelmat.



Kuva 2.1: Oppivien ja älykkäiden järjestelmien yleisrakenne.

Ongelmakentästä kerätyn tiedon analysointi (data mining):
Ongelmakentästä kerätystä tiedosta muodostetaan malli, jonka avulla tutkitaan ongelmakentässä esiintyviä yleisiä säännönmukaisuuksia. Mallille annettu syöte (ratkaistava ongelma) määrittelee tässä tapauksessa esimerkiksi millaisista säännönmukaisuuksista käyttäjä on kiinnostunut.

Teollisten prosessien säätö ja ohjaus:
Rakennettu malli kuvaa jotakin todellisen maailman teollista prosessia. Mallille annettu syöte on tässä tapauksessa joukko malliin vaikuttavia parametreja, ja mallin soveltamisessa on tehtävänä päätellä kuinka määrätä loput parametrit optimaalisella tavalla annetun mallin puitteissa.

Tiedon tiivistäminen ja signaalinkäsittely:
Mallinnettu todellisen maailman prosessi käsittelee tässä tapauksessa signaalien tiivistämistä (compression), suodattamista (filtering), tai korjaamista (correction). Mallin sovellusvaiheessa on tehtävänä optimoida tätä prosessia ohjaavat parametrit.

Mikäli mallien sovellusmoduulin antamien tulosten perusteella tehdään toiminnallisia päätöksiä, on riskianalyysivaiheen tehtävänä arvioida eri vaihtoehdoista seuraavat hyödyt ja haitat.

Bayesiläinen mallintaminen on yksi kiinnostava lähestymistapa oppivien ja älykkäiden järjestelmien rakentamiseksi. Lähestymistapa tarjoaa yhtenäisen, todennäköisyyslaskentaan perustuvan formaalin kehikon, jonka avulla rakennettavan järjestelmän eri vaiheissa esiintyvät osaongelmat voidaan muotoilla ristiriidattomalla, teoreettisesti elegantilla tavalla. Termiä 'bayesiläinen' käytetään seuraavassa korostamaan kahta mallintamiselle keskeistä seikkaa. Ensinnäkin verrattuna moniin muihin lähestymistapoihin, bayesiläisen mallintamisen lähtökohtana on konstruoida probabilistinen malli ongelmakentän *yhteistodennäköisyysjakaumalle*, siis malli ongelmakentässä esiintyville tilanteille kokonaisuutena, eikä pelkästään tietylle ongelmakentän osalle (esimerkiksi yhdelle muuttujalle). Yhteistodennäköisyysjakaumaa voidaan toki soveltaa vain tietyn ongelmakentän osa-alueen jakauman estimointiin, jos näin halutaan. Toiseksi termi viittaa myös siihen, että lähestymistavassa käytetään todennäköisyyksien semanttisessa tulkinnassa modernia, *subjektivistista* näkökantaa, jonka mukaan todennäköisyys on subjektiivinen epävarmuuden mitta, eikä perinteisen, ns. *objektivistisen* lähestymistavan mukainen "toistokoefrekvenssi" (tätä seikkaa käsitellään tarkemmin luvussa 3.2). Todennäköisyyslaskennan käyttämisessä ei tietenkään sinänsä ole mitään subjektiivista, vaan teoria johtaa yhteen tiettyyn ristiriidattomaan eli konsistenttiin tapaan käsitellä epävarmaa tietoa. Subjektivistinen lähestymistapa hyväksyy kuitenkin sen tosiseikan, että mikäli päättelyn lähtökohtana käytetään ihmisten arvioimia todennäköisyyksiä, saattavat eri ihmisten antamat arviot tapahtumien todennäköisyyksistä vaihdella hyvinkin paljon, mikä puolestaan vaikuttaa laskennan lopputulokseen. Yksi käytettävän teoreettisen formalismin suurimmista ansioista on se, että kunkin osaongelman ratkaisemisessa käytettävät perusoletukset on luetteloitava eksplisiittisesti, jolloin eri ratkaisujen järkevyyttä voidaan arvioida paitsi empiirisesti, myös analyyttisesti. Seuraavassa käsitellään kutakin järjestelmän kolmesta komponentista erikseen.

2.2 Mallien rakentaminen

Mallien rakentamista voidaan ajatella yleisesti etsintäprosessina, jossa tavoitteena on löytää ongelmakenttää hyvin kuvaavia numeerisia malleja. Etsintäavaruuden muodostava mallien joukko, *malliperhe*, määritellään yleensä käyttäen ns. parametrisia malliperheitä[1]: yksittäistä mallia voidaan tässä lähestymistavassa ajatella parina (M, θ), missä M määrittelee mallin *struktuurin*, esimerkiksi neuroverkon rakenteen tai sumeiden sääntöjen määrän ja muodon, kun taas θ kiinnittää struktuuriin liittyvät parametrit, esimerkiksi neuroverkon kaarten painot tai sumeiden sääntöjen jäsenyysasteet. *Probabilistisella mallilla* tarkoitetaan seuraavassa mallia (M, θ), joka tuottaa todennäköisyysjakauman $P(\mathbf{d} \mid M, \theta)$ mallimaailmassa esiintyville tilanteille[2].

Käytännössä mallien rakentamisessa joudutaan aina rajoittumaan äärelliseen määrään mallistruktuureja, esimerkiksi joukkoon kolmikerroksisia suunnattuja backpropagation-neuroverkkoarkkitehtuureja, joissa piilosolmujen lukumäärä vaihtelee välillä 1–K. Tässä tapauksessa käytettävissä olevia mallistruktuureja on K kappaletta: $M_1, \ldots, M_K$. Malliperheen voidaan nyt ajatella muodostuvan siitä joukosta malleja (M_k, θ), jotka saadaan aikaan käyttämällä mallistruktuurijoukkoa $\{M_1, \ldots, M_K\}$. Yksittäisen neuroverkon kaarien painot on tässä esimerkissä koodattu parametrivektorina θ. Mallistruktuuria M_k vastaavaksi *malliluokaksi* kutsutaan sitä malliperheen osajoukkoa, joka sisältää kaikkia muotoa (M_k, θ) olevat mallit. Tyypillisiä esimerkkejä oppivissa ja älykkäissä järjestelmissä käytetyistä mallistruktuureista ovat erilaiset neuroverkkorakenteet, päätöspuut, sumeat sääntöjoukot ja Bayes-verkot, joita käsitellään tarkemmin luvussa 4.

[1] Ei-parametrisia malleja ja esimerkkeihin perustuvaa päättelyä (case-based reasoning) käsitellään luvussa 3.3.

[2] Tässä merkinnällä $P(\mathbf{d} \mid M, \theta)$ tarkoitetaan ns. ehdollista todennäköisyyttä, eli tilanteen $\mathbf{d}$ todennäköisyyttä olettaen että mallin (M, θ) kuvaamat rajoitteet ovat voimassa.

Mallistruktuuri M: Rakenne, joka määrää mallien määrittelemiseen tarvittavat osat (parametrit). Esimerkiksi neljä syötesolmua, kaksi piilosolmua, ja neljä tulossolmua sisältävä täydellisesti kytketty neuroverkkoarkkitehtuuri muodostaa yhden mallistruktuurin. Yksittäisen mallin määrittelevät parametrit ovat tässä tapauksessa verkon kaarien painot.

Malli (M, θ): Mallistruktuuri M + siihen liittyvien parametrien ilmentymät eli parametrien kiinnitetyt arvot θ (esimerkiksi $4\times2\times4$–neuroverkkostruktuuri + siihen liittyvien kaarten painot).

Malliperhe $\mathcal{M}$: Joukko mallistruktuureja (esimerkiksi kaikki muotoa $4\times \mathrm{k} \times 4$ olevat neuroverkot, joissa on neljä syötesolmua ja neljä tulossolmua, ja yksi k:n solmun piilokerros). Malliperhettä voidaan ajatella myös kaikkien kiinnitetyllä mallistruktuurijoukolla aikaansaatavien mallien muodostamana joukkona.

Opetusjoukko $\mathcal{D}$: Ongelmakentästä saatu otos esimerkkitapauksia.

Oppiminen: Mallin (siis mallistruktuurin ja malliparametrien arvojen) valinta kiinnitetystä malliperheestä annettua opetusjoukkoa hyväksikäyttäen.

Bayesiläisessä lähestymistavassa mallien rakentamisprosessissa erotetaan eksplisiittisesti *mallinvalintakriteeri*, jonka perusteella päätetään mitkä mallit ovat sopivia käytettäviksi eri ongelmakentissä, ja *etsintäalgoritmi*, jonka avulla etsintäavaruudesta yritetään löytää mallinvalintakriteerin mielessä hyviä malleja. Mikäli käytetty menetelmä mallien konstruoimiseksi ei tuota toivotun kaltaisia tuloksia, voidaan menetelmän evaluoinnissa tämän erottelun ansiosta keskittyä tarkastelemaan erikseen joko käytettyä valintakriteeriä tai etsintäalgoritmia. Esimerkiksi useissa koneoppimisen alueella käytetyistä päätöspuualgoritmeissa ei tämänlaatuista erottelua ole mahdollista tehdä, vaan etsintä ja valintakriteeri muodostavat yhtenäisen kokonaisuuden, jolloin algoritmien heikkouksien analysointi on vaikeaa.

Kuvassa 2.1 esitetyssä järjestelmässä mallien konstruoiminen tapahtuu käyttäen ongelmankentästä olevaa (asiantuntija)tietämystä, ja/tai saatavilla olevaa ns. *opetusjoukon* (training set) muodostavaa esimerkkiaineistoa. Oletetaan seuraavassa aluksi yksinkertaisuuden vuoksi että käytettävä malli halutaan konstruoida pelkästään opetusjoukon perusteella, jolloin mallien rakennusprosessia kutsutaan mallien *oppimiseksi*, ja käytetään kyseisestä otosjoukosta merkintää $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$. Asiantuntijatietämyksen yhdistämistä oppimisprosessiin käsitellään luvussa 2.2.4.

Bayesiläisessä mallintamisessa mallien oppiminen voidaan jakaa kahteen erilliseen vaiheeseen: ensimmäisessä vaiheessa määrätään mallistruktuuri M, ja toisessa vaiheessa määrätään kiinnitetyn mallistruktuurin M parametrit θ. Kuvassa 2.2 esitetään tarkennettu kaavio mallien oppimisprosessista.

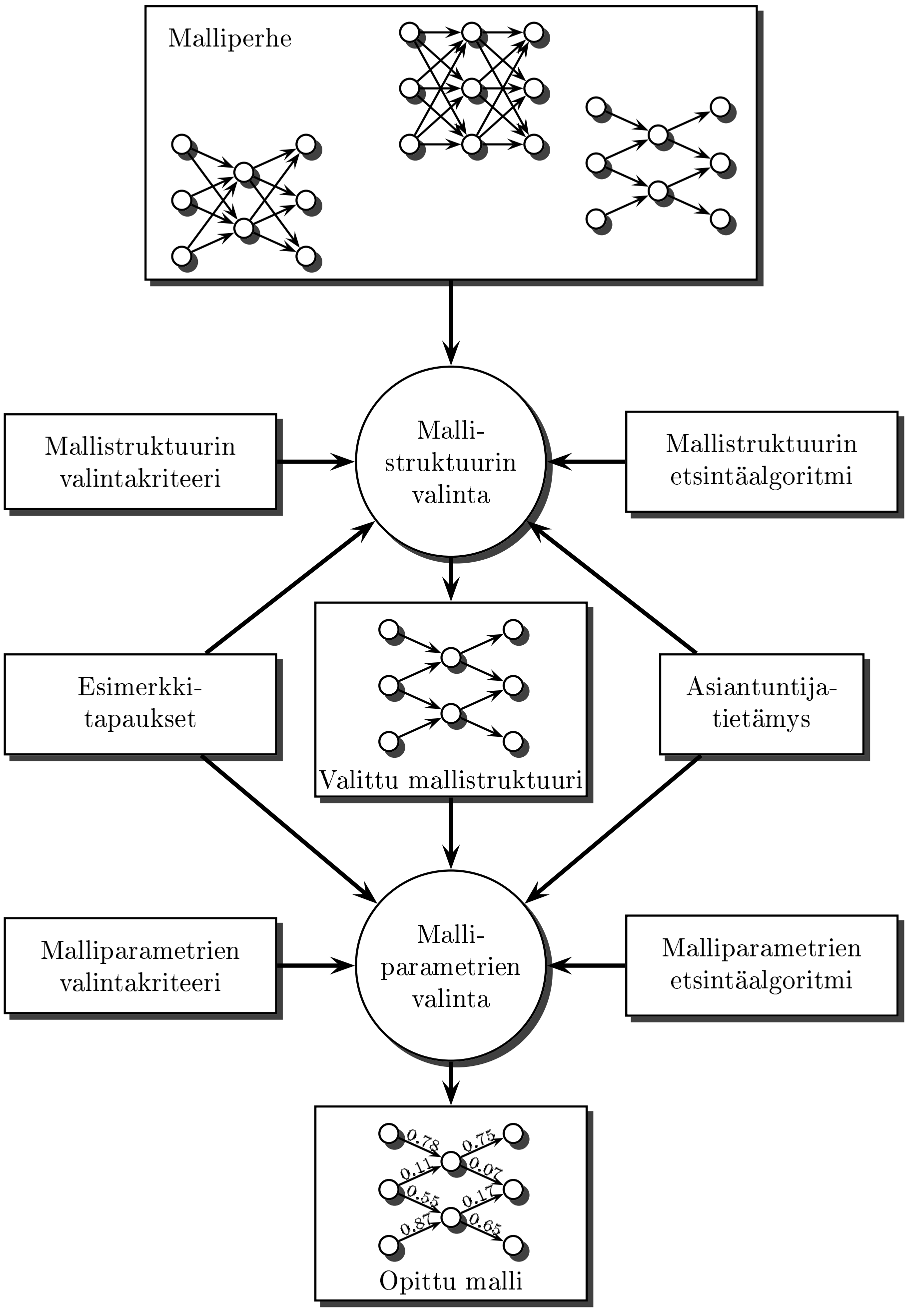
Keskeinen ongelma mallien konstruoimisessa on mallien ylisovittamisen välttäminen, jota käsitellään tarkemmin seuraavassa luvussa. Luvuissa 2.2.2 ja 2.2.3 kuvataan bayesiläinen kriteeri mallistruktuurin ja malliparametrien valitsemiseksi. *Tässä yhteydessä on huomattava että bayesiläinen mallintaminen ei ota kantaa oppimisessa käytettävään etsintäalgoritmiin*, vaan malliavaruutta voidaan läpikäydä mitä tahansa hakumenetelmää käyttäen. Yhden vaihtoehdon tällaiseksi algoritmiksi muodostavat *geneettiset algoritmit*, joita käsitellään tarkemmin erillisessä TEKESin teknologiaraportissa.

2.2.1 Mallien ylisovittaminen ja Occamin partaveitsi

Kuten Jorma Rissanen "Älykkäiden ja oppivien järjestelmien sovellukset"-teknologiaohjelman evaluointiraportissaan [118] toteaa, valitettavan suuri osa oppiviin ja älykkäisiin järjestelmiin liittyvästä tutkimuksesta on keskittynyt nimenomaan mallien parametrien estimointiongelmaan, vaikka mallin struktuurin valitseminen on käytännön kannalta paljon merkittävämpi ongelma: rakenteellisesti liian monimutkaisissa malleissa on parametrien lukumäärä niin suuri, että mallit voidaan *ylisovittaa* (overfit) annettuun opetusjoukkoon. Ylisovittamisen tuloksena syntyvät *ylioppineet* (overtrained) mallit adaptoituvat liian tarkasti opetusjoukkoon, jolloin mallien *yleistyskyky* (generalization capability) eli ennustustarkkuus opetusjoukon ulkopuolisissa tilanteissa on huono. Sanomattakin on selvää, että tällainen järjestelmä, joka toimii hyvin vain tietyissä, ennalta kiinnitetyissä tapauksissa, on käytännön sovellusten kannalta hyödytön.

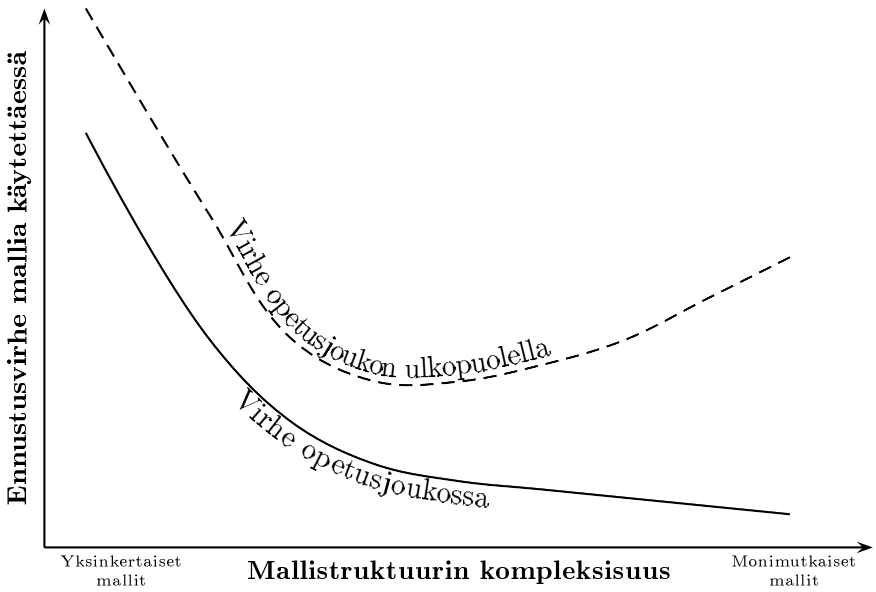
Ylioppiminen: Mallin sovittaminen opetusjoukkoon niin tarkasti, että mallin käyttökelpoisuus on hyvä vain opetusjoukkoon kuuluvissa tapauksissa, ei opetusjoukon ulkopuolella.

Yksi syy ylioppimisilmiöön on se, että koska opetusjoukko on vain suppea ja lisäksi yleensä lukuisia virheitä sisältävä otos ongelmakentästä, se ei voi edustaa ongelmakentän yhteistodennäköisyysjakaumaa täydellisen tarkasti. Siten opetusjoukkoa vastaavan yhteistodennäköisyysjakauman pikkutarkka



Kuva 2.2: Tarkennettu kuva mallien rakennusvaiheesta.

estimointi johtaa malliin, joka edustaa huonosti ongelmakentän todellista todennäköisyysjakaumaa. Kuvassa 2.3 on esitetty tyypillinen esimerkki ylioppimisesta: opetusaineistosta malliparametrien sovittamisen tuloksena opitun mallin ennustustarkkuus opetusjoukossa (ennustustarkkuus testattuna käyttäen vain opetusjoukossa annettuja esimerkkitapauksia) on sitä parempi, mitä monimutkaisempi käytetty mallistruktuuri on (mitä enemmän mallissa on sovitettavia parametreja), kun taas opitun mallin yleistyskyky eli ennustustarkkuus opetusjoukon ulkopuolisissa tapauksissa alkaa huonontua jossakin vaiheessa mallin monimutkaisuuden lisääntyessä. Koska mallien oppimisessa on tavoitteena nimenomaan hyvän yleistyskyvyn saavuttaminen, on ylioppimisen välttäminen koneoppimisen keskeisimpiä ongelmia.



Kuva 2.3: Esimerkki ylioppimisesta.

Ylioppimista voidaan yrittää välttää noudattamalla *Occamin partaveitseksi* kutsuttua periaatetta, jonka mukaan paras mahdollinen malli on sellainen, joka sopii ongelmakentästä saatuihin havaintoihin mahdollisimman tarkasti, mutta jonka struktuuri on toisaalta mahdollisimman yksinkertainen. Mallistruktuurin valinnassa on siis löydettävä sopiva tasapaino mallin rakenteen monimutkaisuuden ja sen esitysvoimakkuuden välillä: monimutkaiset mallit saadaan sovitettua havaintoihin hyvin tarkasti, mutta niiden struktuurit ovat kompleksisisia, yksinkertaisia malleja ei puolestaan voida useinkaan sovittaa tehtyihin havaintoihin kovin tarkasti. Tilastollisessa oppimisessa tätä tasapainotusongelmaa kutsutaan nimellä *vinouma-varianssi-*

ongelma (bias-variance dilemma) (katso esim. [15, 36]).

Vinouma-varianssi-ongelman ratkaisemiseksi on esitetty useita menetelmiä, joista useimmat perustuvat empiirisin argumentein perusteltuihin heuristisiin algoritmeihin (katso esim. [45, 46]). Ongelman ratkaisemiseksi on myös esitetty monimutkaisia teoreettisia kehikoita mm. ristiintestauksen (cross-validation) [4] ja VC-ulottuvuuden [142] käsitteisiin perustuen. Esimerkiksi monet neuroverkoissa käytetyt regularisointitekniikat (regularization) [137, 110] pyrkivät ratkaisemaan ylioppimisongelman lisäämällä mallinvalintakriteeriin "rankaisutermin", jonka arvo kasvaa neuroverkon rakenteen monimutkaistuessa. Seuraavassa luvussa näemme kuinka bayesiläinen lähestymistapa sisältää "sisäänrakennetun" *automaattisen Occamin partaveitsi-periaatteen*, minkä ansiosta ylioppimisongelmaa ei tarvitse käsitellä bayesiläisessä lähestymistavassa erikseen, vaan se ratkeaa luonnollisena osana mallien konstruointiprosessia!

2.2.2 Mallistruktuurin valinta

Edellä esitetyssä kaksivaiheisessa oppimisprosessissa tarvitsemme oppimiskriteerin sekä mallistruktuurien että mallien parametrien arvojen evaluoimiseksi. Bayesiläisessä oppimisessa käytettävä kriteeri perustuu todennäköisyyteen: mallistruktuuri M_1 on "parempi" kuin mallistruktuuri M_2, jos se on näistä vaihtoehdoista *todennäköisempi*. Tämä periaate pätee sekä mallistruktuureille että mallien parametreille: mallistruktuurit ja malliparametrit evaluoidaan niiden *posterioritodennäköisyyden* mukaan. Mallistruktuurin tapauksessa sana 'posteriori' viittaa siihen, että struktuurin M todennäköisyys $P(M \mid \mathcal{D})$ lasketaan sen jälkeen kun opetusaineisto on nähty, siis *ehdollisena todennäköisyytenä* annettuna opetusjoukko $\mathcal{D}$. Malliparametrien tapauksessa posterioritodennäköisyydet saadaan vastaavasti ehdollisesta jakaumasta $P(\theta \mid \mathcal{D}, M)$.

Todennäköisyyslaskennan perusaksioomia käyttäen on helppo osoittaa että

$$P(M \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid M)P(M)}{P(\mathcal{D})}. \tag{2.1}$$

Tämä kaava tunnetaan *Bayesin teoreemana*, jonka esittäjän, Thomas Bayesin (1701–1761) mukaan koko tutkimusalue on saanut nimensä. Koska todennäköisyyttä $P(\mathcal{D})$ voidaan pitää tässä yhteydessä vakiona, Bayesin teoreemasta seuraa, että mallistruktuurien bayesiläisenä oppimiskriteerinä voidaan käyttää tuloa, jonka tekijöinä ovat oppimisjoukon *kokonaisuskottavuus* eli *evidenssi (marginal likelihood, evidence)* $P(\mathcal{D} \mid M)$, ja mallistruktuurin *prioritodennäköisyys* $P(M)$. Tässä yhteydessä on huomattava että kokonaisuskottavuus eroaa tilastotieteestä tutusta uskottavuustermistä $P_{\mathcal{D}}$,

$$P_{\mathcal{D}} = P(\mathcal{D} \mid M, \theta),$$

joka antaa opetusjoukon todennäköisyyden annettuna *täydellinen malli*, siis mallistruktuurin lisäksi myös sitä vastaavien parametrien arvot. Kokonais- eli *marginaaliuskottavuus* saadaan "tavallisesta" uskottavuudesta nimensä mukaisesti *marginalisoimalla* eli *integroimalla* uskottavuustermi yli kaikkien mallistruktuuria M vastaavien mallien, painotettuna parametriasetusten θ prioritodennäköisyydellä:

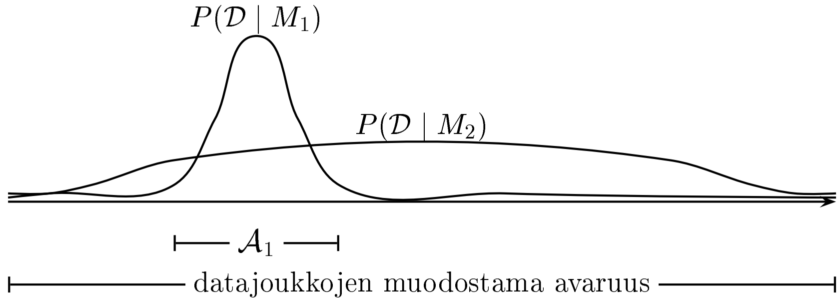
$$P(\mathcal{D} \mid M) = \int P(\mathcal{D} \mid M, \theta) P(\theta \mid M) d\theta. \tag{2.2}$$

Näemme että kokonaisuskottavuutta voidaan pitää normaalin uskottavuustermin odotusarvona niiden mallien joukossa, joiden struktuuri on M.

Mikäli yllä esiintyvä integraali voidaan laskea, antavat kaavat (2.1) ja (2.2) bayesiläisen kriteerin eri mallistruktuurien vertailemiseksi. Luvussa 4.4 näemme, kuinka tämä kriteeri voidaan laskea hyvin tehokkaasti Bayes-verkkojen muodostaman probabilistisen malliperheen tapauksessa. Bayesiläistä lähestymistapaa voidaan soveltaa myös tiettyjen neuroverkkomallien tapauksessa [90, 99, 12, 14, 117], joskaan ei niin suoraviivaisesti kuin probabilististen malliperheiden kanssa.

Yksi tapa soveltaa edellisessä luvussa kuvattua Occamin partaveitsi-periaatetta bayesiläisessä lähestymistavassa saadaan määrittelemällä priorijakauma $P(M)$ siten, että mallistruktuurien prioritodennäköisyys pienenee struktuurin monimutkaisuuden kasvaessa. Mikäli mitään muuta syytä — esimerkiksi ongelmakentän asiantuntijan esittämää arvioita mallinnettavan maailman luonteesta — tällaisen priorijakauman käyttämiseen ei kuitenkaan ole, voidaan ei-tasaisen priorijakauman käytöstä luopua, koska *kokonaisuskottavuuskriteeri sisältää jo itsessään automaattisen Occamin partaveitsi-periaatteen!* Tätä aluksi hieman yllättävältä tuntuvaa seikkaa on käsitelty laajemmin mm. teoksissa [115, 90, 99] — asian ymmärtämiseksi intuitiivisella tasolla tarkastelkaamme seuraavaa esimerkkiä.

Olkoon M_1 ja M_2 kaksi mallistruktuuria siten, että M_1 on hyvin yksinkertainen, ja M_2 hyvin monimutkainen mallistruktuuri. Koska siis mallistruktuurin M_2 tapauksessa löytyy aina suuri määrä parametriasetuksia θ, joiden avulla mallistruktuuri voidaan sovittaa tehtyihin havaintoihin $\mathcal{D}$, voimme olettaa että kokonaisuskottavuus $P(\mathcal{D} \mid M_2)$ on kohtalaisen suuri hyvin monelle datajoukolle $\mathcal{D}$. Vastaavasti mallistruktuurin M_1 tapauksessa kokonaisuskottavuus on suuri vain hyvin rajoitetulle määrälle datajoukkoja;



Kuva 2.4: Intuitiivinen perustelu sille miksi kokonaisuskottavuus noudattaa automaattisesti Occamin partaveitsi-periaatetta: molempien käyrien rajaaman alueen pinta-alan on oltava yhtä suuri (yksi), jolloin yksinkertainen mallistruktuuri M_1 antaa jollakin alueella (tässä $\mathcal{A}_1$) suuremman kokonaistodennäköisyyden arvon kuin monimutkaisempi mallistruktuuri M_2.

olkoon tämä alue $\mathcal{A}_1$. Koska kokonaisuskottavuus määrää jakauman kaikille mahdollisille datajoukoille $\mathcal{D}$, on selvää että $P(\mathcal{D} \mid M_2)$ ei voi olla suurempi kuin $P(\mathcal{D} \mid M_1)$ kaikilla $\mathcal{D}$. Kuva 2.4 havainnollistaa tätä seikkaa tarkemmin: monimutkaisen mallistruktuurin M_2 tapauksessa on kokonaisuskottavuus levittynyt tasaisesti laajalle, monien datajoukkojen $\mathcal{D}$ muodostamalle alueelle, kun taas mallistruktuurin M_1 tapauksessa kokonaisuskottavuuden merkitsevät arvot saadaan alueella $\mathcal{A}_1$, ja sen ulkopuolella kokonaisuskottavuus on lähellä nollaa. Yksinkertaisempaa mallistruktuuria M_1 vastaava kokonaisuskottavuus on siis suurempi, mikäli käytettävä datajoukko on alueelta $\mathcal{A}_1$!

Occamin partaveitsi-periaatteen toteutumisen näemme myös soveltamalla kokonaisuskottavuuden laskemisessa ns. BIC-approksimaatiota [125, 114, 67], jonka mukaan

$$P(\mathcal{D} \mid M) \approx \frac{P(\mathcal{D} \mid M, \tilde{\theta})}{N^{\frac{d(M)}{2}}},$$

missä merkintä $\tilde{\theta}$ tarkoittaa uskottavuustermin $P(\mathcal{D}|M, \theta)$ maksimoivia *suurimman uskottavuuden parametreja (maximum likelihood parameters)*, $\mathrm{d}(M)$ mallistruktuuria M vastaavien malliparametrien lukumäärää, ja N datajoukon $\mathcal{D}$ kokoa. Näemme siis, että kokonaisuskottavuus tasapainoilee Occamin partaveitsi-periaatteen mukaisesti mallin esityskyvyn ja monimutkaisuuden välillä: mallin struktuurin monimutkaisuuden kasvaessa saa osoittajassa esiintyvä uskottavuustermi yhä suurempia arvoja, mutta samanaikaisesti kasvaa myös nimittäjässä esiintyvä mallin monimutkaisuudesta rankaiseva termi.

Jos esitämme BIC-approksimaation hieman toisessa muodossa (ottamalla logaritmit kyseisistä termeistä), saamme

$$-\log P(\mathcal{D} \mid M) \approx -\log P(\mathcal{D} \mid \tilde{\theta}(\mathcal{D})) + \frac{d(M)}{2} \log N,$$

mistä näemme että approksimaatio on sama kuin Rissasen teoksessa [114] esittämä *Minimum Description Length (MDL)*-periaatteeseen perustuva kaksiosainen informaatioteoreettinen kriteeri mallistruktuurin valitsemiseksi. Bayesiläisen ja informaatioteoreettisten lähestymistapojen yhteyksiä käsitellään myöhemmin tarkemmin luvussa 3.4.

2.2.3 Malliparametrien valinta

Kun kriteerin (2.1) mielessä optimaalinen mallistruktuuri M on löydetty, on seuraava tehtävä oppimisprosessissa määrätä mallistruktuuriin liittyvät parametrit θ. Monissa ei-probabilistisissa malliperheissä, kuten esimerkiksi neuroverkkomalleissa, käytetty mallinvalintakriteeri on tässä tapauksessa yleensä estimointivirhe opetusjoukossa. Kuten edellä nähtiin, tämä johtaa kuitenkin helposti ylioppimiseen mikäli mallistruktuuria ei ole valittu oikein. Lisäksi opetusvirhettä minimoivat etsintäalgoritmit (esimerkiksi backpropagation) ovat yleensä yksinkertaisia iteratiivisia gradienttimenetelmän variaatioita, joiden konvergenssi saattaa olla hyvin hidasta. On myös huomattava, että gradienttimenetelmät ovat paikallisia (lokaaleja) menetelmiä, jotka parhaassakin tapauksessa konvergoivat vain lokaaliin optimiin, eivätkä välttämättä löydä niitä globaalisti optimaalisia parametriarvoja jotka maksimoivat käytetyn parametrien evaluointikriteerin.

Bayesiläisessä mallintamisessa sitä vastoin parametrien arvojen valinnassa käytettävä oppimiskriteeri on sama kuin mallistruktuurien valinnassa: parametriarvot θ evaluoidaan niiden posterioritodennäköisyyden $P(\theta \mid M, \mathcal{D})$ mukaan. Todennäköisyyslaskennan perusaksioomista saamme että

$$P(\theta \mid M, \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta, M) P(\theta \mid M)}{P(\mathcal{D} \mid M)}. \tag{2.3}$$

Yksittäisen mallin θ posterioritodennäköisyys saadaan siis kertomalla opetusjoukon uskottavuus $P(\mathcal{D} \mid \theta, M)$ mallin prioritodennäköisyydellä $P(\theta \mid M)$, ja normalisoimalla syntyvä tulo kokonaisuskottavuudella (2.2), jonka arvo on jo määrätty mallistruktuurin valinnassa.

Tietyissä probabilistisissa malliperheissä voidaan posteriorijakauman (2.3) maksimikohta laskea suljetussa muodossa yksinkertaisella matemaattisella kaavalla, jolloin *mallin parametrien säätämiseen ei tarvita lainkaan iteratiivista aikaavievää oppimisalgoritmia*! Esimerkin tällaisesta malliperheestä muodostavat tässä raportissa käsiteltävät Bayes-verkot, joita tarkastellaan tarkemmin luvussa 4. Luvussa 4.4.2 esitetään yksinkertainen laskukaava, jonka avulla annetun Bayes-verkon parametrit voidaan määrätä suoraan ilman iteratiivista oppimisprosessia.

2.2.4 Oppimisen ja asiantuntijatietämyksen yhdistäminen

Edellä oletettiin mallistruktuurin valinnassa käytettävän pelkästään tilastollista havaintoaineistoa $\mathcal{D}$. Usein saatavilla on kuitenkin esimerkkitapauksien lisäksi myös runsaasti "ihmistietämystä": ongelmakentän asiantuntijoiden kokemusperäistä tietoa, tai "maalaisjärkeen" (common sense) perustuvaa tietoa ongelmakentän luonteesta. Monien malliperheiden, kuten esimerkiksi neuroverkkojen, tapauksessa merkittäväksi ongelmaksi on muodostunut se, kuinka nämä kaksi tietolähdettä saadaan yhdistettyä. Bayesiläisessä mallintamisessa yhdistäminen on suoraviivaista: kaavoissa (2.1) ja (2.2) esiintyy kaksi prioritodennäköisyysjakaumaa, $P(M)$ ja $P(\theta \mid M)$. Sana 'priori' viittaa tässä yhteydessä siihen, että nämä jakaumat on määrättävä "ennen" opetusjoukon $\mathcal{D}$ tarkastelua, siinä mielessä että ne eivät saa riippua opetusjoukon muodostavasta tilastollisesta aineistosta. Bayesiläinen formalismi antaa nyt mahdollisuuden koodata ongelmakentästä saatavilla olevaa yleistä tietämystä näiden prioritodennäköisyyksien avulla.

Luvussa 4.4 näemme, että Bayes-verkkojen tapauksessa mallistruktuurien prioritodennäköisyyksien arvioiminen on erityisen luontevaa, koska kukin Bayes-verkkostruktuuri edustaa joukkoa muuttujien välisiä riippuvuuksia, ja tällaisten riippuvuuksien arviointi on ongelmakentän asiantuntijoille monessa tapauksessa sangen helppoa. Lisäksi priorijakauma $P(\theta \mid M)$ voidaan formalisoida Bayes-verkkojen tapauksessa tavalla, joka antaa asiantuntijoille mahdollisuuden paitsi koodata prioritietämyksensä, myös arvioida prioritiedon merkityksellisyyttä suhteessa saatavilla olevaan tilastolliseen aineistoon. Toisaalta on myös korostettava, että koska Bayes-verkkojen tapauksessa on sekä mallistruktuurilla että siihen liittyvillä parametreilla selvä semanttinen merkitys, voidaan Bayes-verkkoja konstruoida aivan hyvin myös ilman tilastollista aineistoa, yksinomaan asiantuntijoilta saatua tietämystä käyttäen.

2.3 Mallien soveltaminen

Kuten edellä näimme, bayesiläisessä mallintamisessa muodostetaan ongelmakentästä malli (M, θ) käyttäen apuna opetusesimerkkejä $\mathcal{D}$ ja/tai ongelmakentän asiantuntijoiden prioritietämystä. Tässä luvussa käsittelemme sitä, kuinka opittuja (tai rakennettuja) malleja voidaan soveltaa erilaisissa ongelmanratkaisutilateissa.

Oletetamme jatkossa, että ongelmakentän kuvauksessa käytetään n attribuuttia (satunnaismuuttujaa) $X_1, \ldots, X_n$. Merkinnällä $X_i = x_i$ tarkoitetaan seuraavassa sitä, että attribuutin X_i arvo on x_i, missä x_i voi olla joko reaalilukuku, jolloin kyseessä on *jatkuva-arvoinen* muuttuja, tai sitten x_i kuuluu johonkin äärelliseen joukkoon mahdollisia arvoja, missä tapauksessa attribuuttia X_i kutsutaan *diskreetiksi.* Jatkossa keskitymme yksinkertaisuuden vuoksi hyvin paljon diskreetteihin muuttujiin — jatkuvien muuttujien käsittelystä puhutaan enemmän luvussa 4.5.

2.3.1 Probabilistinen päättely

Seuraavassa oletamme, että käytetty malliperhe on probabilistinen, missä tapauksessa käytettävä malli (M, θ) määrittelee yhteistodennäköisyysjakauman $P(X_1 = x_1, \ldots, X_n = x_n \mid M, \theta)$ muuttujien $X_1, \ldots, X_n$ eri arvokombinaatioille. Tätä yhteistodennäköisyysjakaumaa käyttäen on mahdollista ratkaista erilaisia mallien soveltamisessa esiintyviä estimointitehtäviä. Tätä havainnollistaaksemme olettakaamme että attribuutit on jaettu kolmeen erilliseen osajoukkoon siten, että ensimmäisen osajoukon $\mathbf{S}_1$ muodostavat ne attribuutit joiden arvoa halutaan estimoida, toisen osajoukon $\mathbf{S}_2$ ne attribuutit joiden arvo on annettu, ja kolmannen osajoukon $\mathbf{S}_3$ muut attribuutit. *Probabilistisessa päättelyssä* tavoitteena on laskea ehdollinen todennäköisyysjakauma kiinnostuksen kohteena oleville attribuuteille, annettuna tunnettujen attribuuttien arvot. Tuloksena on siis jakauma $P(\mathbf{S}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$, missä merkintä $\mathbf{S}_2 = \mathbf{s}_2$ tarkoittaa että kukin attribuutti joukossa $\mathbf{S}_2$ on asetettu tunnettuun arvoonsa.

Tässä yhteydessä on korostettava, että attribuuttien jako yllämainittuihin kolmeen osajoukkoon ei ole millään tavalla kiinteä, vaan voi vaihdella tilanteen mukaan. Niinpä probabilistisessa päättelyssä ei tarvitse etukäteen päättää mitä halutaan estimoida, tai mitä tietoa estimointitehtävässä on saatavilla. Tämä poikkeaa monista vaihtoehtoisista lähestymistavoista, kuten esimerkiksi päätöspuista ja monista neuroverkkomalleista, joissa estimointitehtävä (esimerkiksi luokittelutehtävä) on kiinnitettävä etukäteen jo ennen mallin rakentamista. Probabilistisessa päättelyssä sitä vastoin riittää, jos yhteistodennäköisyysjakauma $P(X_1, \ldots, X_n \mid M, \theta)$ on määrätty: toden-

Muuttujajoukko $\mathbf{S}_1$:
Ne muuttujat, joiden arvo ei ole tiedossa, mutta joiden arvojen tuntemisen katsotaan olevan edellytyksenä käsillä olevan ongelmanratkaisutilanteen ratkaisemiksi.

Muuttujajoukko $\mathbf{S}_2$:
Muuttujat joiden arvo on tunnettu.

Muuttujajoukko $\mathbf{S}_3$:
Muuttujat, joiden arvo ei ole tiedossa, eikä arvojen tuntemisen katsota olevan välttämätön käsillä olevan ongelmanratkaisutilanteen kannalta.

Probabilistinen päättely:
Todennäköisyysjakauman $P(\mathbf{S}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$ estimointi.

näköisyyslaskennan perusaksioomista seuraa, että ehdolliset todennäköisyydet voidaan esittää muodossa

$$\begin{aligned} P(\mathbf{S}_1 = \mathbf{s}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta) &= \frac{P(\mathbf{S}_1 = \mathbf{s}_1, \mathbf{S}_2 = \mathbf{s}_2 \mid M, \theta)}{P(\mathbf{S}_2 = \mathbf{s}_2 \mid M, \theta)} \\ &= \frac{\sum_{\mathbf{S}_3} P(\mathbf{S}_1 = \mathbf{s}_1, \mathbf{S}_2 = \mathbf{s}_2, \mathbf{S}_3 \mid M, \theta)}{\sum_{\mathbf{S}_1, \mathbf{S}_3} P(\mathbf{S}_1, \mathbf{S}_2 = \mathbf{s}_2, \mathbf{S}_3 \mid M, \theta)}, \end{aligned} \quad (2.4)$$

missä merkintä $\sum_{\mathbf{S}_i}$ tarkoittaa sitä, että summassa käydään läpi kaikki joukossa $\mathbf{S}_i$ esiintyvien muuttujien arvokombinaatiot. Mikä tahansa ehdollinen todennäköisyys saadaan siis laskettua summaamalla eli *marginalisoimalla* yli tietyn yhteistodennäköisyyksien joukon. Jotta tätä lähestymistapaa olisi mahdollista soveltaa käytännössä, on ratkaistava seuraavat kaksi ongelmaa:

1. Kuinka tallettaa yhteistodennäköisyydet $P(X_1, \ldots, X_n \mid M, \theta)$ siten, että niiden talletus ei vaadi eksponentiaalista muistitilaa ja että ne ovat tehokkaasti käytettävissä?

2. Kuinka laskea kaavassa (2.4) esiintyviä marginaalisummia tehokkaasti?

Luvussa 4 osoitamme kuinka yllä esitetyt ongelmat voidaan ratkaista kun yhteistodennäköisyysjakaumat määritellään käyttäen Bayes-verkkojen malliperhettä: luvussa 4.1 käsitellään yhteistodennäköisyyksien tallettamisen ongelmaa, ja luvussa 4.3 marginalisointiongelmaa Bayes-verkkojen tapauksessa.

Probabilistisessa päättelyssä estimoidaan siis muotoa (2.4) olevaa ehdollista jakaumaa. Luvussa 2.4 käsitellään sitä, miten tätä jakaumaa voidaan käyttää hyväksi päätöksenteossa arvioimalla eri toimenpiteistä seuraavien hyötyjen ja haittojen odotusarvoja. Ennen sitä esittelemme kuitenkin ensin joitakin tyypillisiä probabilistisen päättelyn muotoja.

2.3.2 Esimerkkejä

Yksinkertaisimmassa probabilistisen päättelyn muodossa joukko $\mathbf{S}_1$ käsittää vain yhden muuttujan X, ja tehtävänä on määrätä muuttujan X arvojen jakauma lähtötietojen $\mathbf{S}_2 = \mathbf{s}_2$ pohjalta. Tyypillisen esimerkin tällaisesta ongelmasta muodostavat *luokitteluongelmat*, joissa tehtävänä on määrittää diskreetin luokkamuuttujan X arvojen todennäköisyydet (esimerkki 1).

Esimerkki 1 Yhden diskreetin muuttujan jakauman estimointi (luokitteluongelma).

Ongelma1:	Sairastaako potilas tautia X havaittujen oireiden ja tehtyjen kokeiden ($\mathbf{s}_2$) perusteella?
Ongelma2:	Mikä on asiakkaan luottokelpoisuusluokka (X) havaitun osittaisen asiakasprofiilin ($\mathbf{s}_2$) perusteella?
Ratkaisu1:	Laske $P(X = 1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$.
Ratkaisu2:	Laske $P(X = x \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$ kaikille luottokelpoisuusluokille x.

Esimerkissä 1 käytettyjen diskreettien muuttujien jakauma voidaan määrätä laskemalla todennäköisyys kullekin muuttujan mahdolliselle arvolle. Mikäli X on jatkuva-arvoinen, jakauma voidaan tietyin jakaumaoletuksin kuvata antamalla jakauman tunnusluvut, esimerkiksi odotusarvo ja varianssi (katso esimerkki 2).

Esimerkki 2 Jatkuvan muuttujan jakauman estimointi.

Ongelma1:	Mikä on tuotteen valmistuskustannusten (X) oletettava arvo, jos täytämme asiakkaan tuotteelle asettamat vaatimukset ($\mathbf{s}_2$)?
Ongelma2:	Mikä on ulostuloventtiilin X virtauksen oletettava voimakkuus, jos säätimet $\mathbf{S}_2$ asetetaan arvoihin $\mathbf{s}_2$?
Ratkaisu:	Laske jakauman $P(X \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$ maksimikohta.

Esimerkissä 1 esitetyissä diskreeteissä luokitteluongelmissa probabilistista päättelyä käytettiin diagnostiseen päättelyyn, eli syiden (esim. sairaus) todennäköisyyksien arviointiin, annettuna havaitut seuraukset (oireet). Toisaalta, kuten edellä mainittiin, muuttujien jakoa joukkohin $\mathbf{S}_1$, $\mathbf{S}_2$ ja $\mathbf{S}_3$ voidaan muuttaa dynaamisesti mielivaltaisella tavalla. Niinpä siirtämällä muuttujia joukosta toiseen voimme soveltaa probabilistista päättelyä myös päinvastaiseen suuntaan, seurausten ennustamiseen syiden pohjalta. Tämänkaltaista päättelyä voidaan käyttää esimerkiksi ongelmakentän luonteen analysointiin (data mining) esimerkissä 3 esitetyillä tavoilla.

Esimerkki 3 Ongelmakentän analysointi.

Ongelma1:	Kuinka suuri osa tautia X sairastavista potilaista kärsii äkillisistä tajuttomuuskohtauksista ($Y = 1$)?
Ongelma2:	Mikä on parhaaseen luottokelpoisuusluokkaan (AAA) kuuluvien asiakkaiden yleisin tilimuoto (Y)?
Ratkaisu1:	Laske $P(Y = 1 \mid X = 1, M, \theta)$.
Ratkaisu2:	Laske $P(Y = y \mid X = AAA, M, \theta)$ kaikille tilimuodoille y.

Esitettyjä kahta päättelyn muotoa (seurauksien arviointi syistä, syiden päättely seurauksista) voidaan tietenkin edelleen sekoittaa mielivaltaisella tavalla ratkaisemaan mm. esimerkissä 4 esitettyjen tapausten kaltaisia ongelmia.

Esimerkki 4 Eri päättelytapojen yhdistäminen.

Ongelma1:	Onko oletettavaa, että potilas voi kärsiä tajuttomuuskohtauksista ($Y = 1$), jos hänellä on tauti X, ja hänen ruumiinlämpötilansa on z?
Ongelma2:	Onko tehtyjen havaintojen ($\mathbf{s}_2$) perusteella oletettavaa, että potilas sairastaa tautia X, ja että hän voi saada vakavia tajuttomuuskohtauksia?
Ratkaisu1:	Laske $P(Y = 1 \mid X = 1, Z = z, M, \theta)$.
Ratkaisu2:	Laske $P(X = 1, Y = 1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$.

Mikäli joukko $\mathbf{S}_1$ sisältää useita muuttujia, tulee jakauman $P(\mathbf{S}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$ täydellisestä määräämisestä vaikeaa, koska se vaatii diskreetissäkin tapauksessa kaikkien joukossa $\mathbf{S}_1$ olevien muuttujien arvokombinaatioiden läpikäynnin, ja näitä arvokombinaatioita on tietenkin eksponentiaalinen määrä. Niinpä tällaisessa tilanteessa tyydytäänkin usein etsimään se muuttujien $\mathbf{S}_1$ arvokombinaatio $\mathbf{s}_1$ (tai n tässä mielessä parasta arvokombinaatiota), jolla on korkein todennäköisyys, ts. arvokombinaatio, joka maksimoi todennäköisyyden $P(\mathbf{S}_1 = \mathbf{s}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$. Koska

$$\max_{\mathbf{s}_1} P(\mathbf{S}_1 = \mathbf{s}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta) = \max_{\mathbf{s}_1} P(\mathbf{S}_1 = \mathbf{s}_1, \mathbf{S}_2 = \mathbf{s}_2 \mid M, \theta), \tag{2.5}$$

näemme että tässä mielessä optimaalinen arvokombinaatio täydentää vapaille muuttujille $\mathbf{S}_1$ arvot siten, että syntyvän kokonaisarvokombinaation todennäköisyys on maksimaalinen. Tyypillisiä esimerkkejä tämänkaltaisen probabilistisen päättelyn sovellusalueista ovat konfiguraatio-ongelmat, joissa täydennetään osakonfiguraatio siten, että täydennetyn konfiguraation todennäköisyys on mahdollisimman suuri annetun mallin muodostamassa todennäköisyysjakaumassa. Jos malli (M, θ) muodostetaan esimerkiksi jotakin tuotetta valmistavan yrityksen valmistettujen tuotteiden tietokannasta, voidaan tällä tavalla täydentää asiakkaan tuotteelle esittämät toivomukset siten, että uusi tuote on mahdollisimman samankaltainen jo valmistettujen tuotteiden kanssa, ja siten otaksuttavasti edullinen valmistaa (katso esimerkki 5).

Esimerkki 5 Optimaalisen konfiguraation etsiminen.

Ongelma1:	Täydennä valmistettavan tuotteen kuvaus ($\mathbf{S}_1$) siten, että tuote täyttää asiakkaan sille esittämät vaatimukset ($\mathbf{s}_2$), ja sopii yrityksen tuoteprofiiliin (eli siitä tehtyyn malliin) mahdollisimman hyvin.
Ongelma2:	Konstruoi asiakkaan täydellinen asiakasprofiili ($\mathbf{S}_1$) annettujen osatietojen $\mathbf{s}_2$ pohjalta.
Ratkaisu:	Etsi $\mathbf{s}_1$ siten, että todennäköisyyden $P(\mathbf{S}_1 = \mathbf{s}_1, \mathbf{S}_2 = \mathbf{s}_2 \mid M, \theta)$ arvo on maksimaalinen.

Ehdollisia todennäköisyyksiä voidaan käyttää myös etäisyysmittana, jonka avulla on mahdollista etsiä esimerkiksi valmistettujen tuotteiden joukosta se tapaus, joka muistuttaa eniten asiakkaan antamaa osittaista tuotteen kuvausta. Tällaista probabilistista etäisyysmetriikkaa käsitellään lähteessä [71].

Konfiguraatio-ongelmia läheisesti muistuttavan sovellusalueen muodostavat asiakasprofilointitehtävät (katso esimerkki 5, ongelma 2). Tavoitteena on täydentää asiakkaan profiili asiakkaasta kerättyjen tietojen ja kaikkien asiakkaiden muodostamasta tietokannasta muodostetun mallin perusteella. Syntyvää täydennettyä asiakasprofiilia voidaan käyttää hyväksi mm. suoramarkkinoinnin suuntaamisessa: probabilistista päättelyä voidaan käyttää esimerkiksi ennustamaan olisiko asiakas kiinnostunut tietyntyyppisestä luottokortista, annettuna hänestä kerätyt tiedot.

Koska probabilistista päättelyä käyttäen voidaan muodostaa täydellinen mallimaailman tilanteen kuvaus, voidaan yhteistodennäköisyyden maksimoivia konfiguraatioita käyttää myös probabilistisen päättelyn selitysmekanismina: muodostetusta täydennetystä konfiguraatiosta voidaan esimerkiksi päätellä mitä muita tauteja tai oireita jo diagnosoidulla potilaalla mahdollisesti on (katso esimerkki 6). Ennustettujen lisäoireiden havaitseminen lisää luonnollisesti jo tehdyn diagnoosin luotettavuutta.

Esimerkki 6 Päättelyä tukevat selitykset.

Ongelma:	Oletetaan, että potilas sairastaa tautia X ja että hänellä on oireet $\mathbf{S}_2 = \mathbf{s}_2$. Mitä muita oireita potilaalla luultavasti on? Sairastaako potilas mahdollisesti myös jotain muuta tautia?
Ratkaisu:	Etsi $\mathbf{s}_3$ siten, että todennäköisyyden $P(X = 1, \mathbf{S}_2 = \mathbf{s}_2, \mathbf{S}_3 = \mathbf{s}_3 \mid M, \theta)$ arvo on maksimaalinen.

Kuten edellä esitetyistä esimerkeistä huomaamme, puuttuva tieto käsitellään probabilistisessa päättelyssä erityisen elegantilla tavalla: jos esimerkiksi alkuperäinen suunnitelmamme oli estimoida jakaumaa $P(X_1 \mid X_2 = x_2, \ldots, X_5 = x_5)$, mutta muuttujan X_3 arvoa ei olekaan saatavilla, käytämme päättelyssä yksinkertaisesti jakaumaa $P(X_1 \mid X_2 = x_2, X_4 = x_4, X_5 = x_5)$. Kullakin hetkellä käytetään siis jakaumaa, joka saadaan ottamalla huomioon kaikki saatavilla oleva tieto. Monien muiden malliperheiden tapauksessa on puuttuvan tiedon käsittely huomattavan hankalaa: esimerkiksi päätöspuiden toimivuuden perusedellytyksenä on se, että kussakin päätössolmussa käytettävän muuttujan arvo on aina mahdollista saada selville.

Probabilistista päättelyä voidaan käyttää myös puuttuvan tiedon estimointiin: sen sijaan että estimoisimme jakaumaa $P(\mathbf{S}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$, asetamme estimoitavat muuttujat johonkin arvokombinaatioon $\mathbf{s}_1$, ja kysymme kuinka muuttujat $\mathbf{S}_3$ on asetettava kun halutaan maksimoida halutun arvokombinaation todennäköisyys $P(\mathbf{S}_1 = \mathbf{s}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, \mathbf{S}_3, M, \theta)$. Esimerkissä 7 esitetyt tapaukset kuvaavat tämäntyyppisiä sovellusalueita.

Esimerkki 7 Puuttuvan tiedon estimointi.

Ongelma1:	Mitä lääkettä ($\mathbf{S}_3$) potilaalle on annettava, jotta toipumisen todennäköisyys olisi maksimaalinen?
Ongelma2:	Kuinka asettaa säätimet $\mathbf{S}_3$ siten, että läpivirtaus putkessa X olisi optimaalinen tilanteessa, jossa säätimet $\mathbf{S}_2$ halutaan pitää nykyisissä asetuksissa $\mathbf{s}_2$?
Ongelma3:	Muodosta valmistettavan tuotteen kuvaus ($\mathbf{S}_3$) siten, että se on konsistentti asiakkaan vaatimusten ($\mathbf{s}_2$) kanssa, ja voitto on maksimaalinen.
Ratkaisut1&2:	Etsi $\mathbf{s}_3$ siten, että todennäköisyyden $P(X = 1 \mid \mathbf{S}_2 = \mathbf{s}_2, \mathbf{S}_3 = \mathbf{s}_3, M, \theta)$ arvo on maksimaalinen.
Ratkaisu3:	Etsi $\mathbf{s}_3$ siten, että jakauman $P(X \mid \mathbf{S}_2 = \mathbf{s}_2, \mathbf{S}_3 = \mathbf{s}_3, M, \theta)$ odotusarvo on maksimaalinen.

Tässä yhteydessä on huomautettava, että esimerkissä 7 esiintyvä päättelyn muoto, puuttuvan tiedon estimointi, on monissa tilanteissa laskennallisesti raskaampi tehtävä kuin jakauman $P(\mathbf{S}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$ estimointi, päättely puuttuvan tiedon vallitessa.

Paitsi probabilistisessa päättelyssä esiintyvissä syötearvoissa, puuttuvaa tietoa voi esiintyä myös opetusaineistossa $\mathcal{D}$. Lähteessä [73] tarkastellaan bayesiläisen lähestymistavan käyttämistä tämän ongelman ratkaisemisessa.

2.4 Riskianalyysi

Edellisessä luvussa kuvattiin, kuinka probabilistisessa päättelyssä on tavoitteena muodostaa ehdollinen todennäköisyysjakauma estimoitaville muuttujille $\mathbf{S}_1$, annettuna tunnettujen muuttujien $\mathbf{S}_2$ arvot. Oletetaan seuraavassa yksinkertaisuuden vuoksi, että estimoimme jakaumaa $P(X \mid \mathbf{S}_2 = \mathbf{s}_2, \theta, M)$, missä X on diskreetti muuttuja, jonka arvo meidän on kiinnitettävä annettujen lähtötietojen $\mathbf{s}_2$ pohjalta. Intuitiivisesti luonnollisimmalta tuntuvin vaihtoehto on tietenkin valita se muuttujan X arvo x, joka maksimoi todennäköisyyden $P(X = x \mid \mathbf{S}_2 = \mathbf{s}_2, \theta, M)$. Käytännön sovelluksissa saatetaan kuitenkin tehdä muuttujan arvon valinnan perusteella päätöksiä, joiden seuraukset voivat olla hyvin kauaskantoisia, ja siksi päätöksenteon huolellinen analyysi on paikallaan. Todennäköisyyslaskenta tarjoaa *päätösteoriana (decision theory)* tunnetun teoreettisen kehikon eri päätöstilanteissa esiintyvien ongelmien ratkaisemiseksi. Päätösteorian mukaan meidän on valittava se vaihtoehto, joka minimoi odotettavissa olevan haitan (maksimoi odotettavissa olevan hyödyn). Esimerkki 8 havainnollistaa tätä periaatetta.

Jos ehdolliset todennäköisyydet $P(X = x \mid \mathbf{S}_2 = \mathbf{s}_2, \theta, M)$ on arvioitu oikein, bayesiläisen päätösteorian mukaan toimivan agentin toiminta on optimaalista, tulkittuna siten, että tietyllä ajanjaksolla saavutettavan hyödyn odotusarvo on tätä politiikkaa käyttäen maksimaalinen. Toisaalta, vaikka päätöksissä käytettävä todennäköisyysjakauma olisikin arvioitu väärin, bayesiläiseen päätösteoriaan perustuvan toiminnan voidaan osoittaa silti olevan *ristiriidattomasti rationaalista* siinä mielessä, että ei ole mahdollista konstruoida sellaista päätöksentekotilanteiden jonoa, jonka tuloksena bayesiläinen päätöksentekijä häviäisi varmasti (ns. "*Dutch book*"-argumentti, katso esim. [10, 9, 32]). Lisäksi voidaan osoittaa, että bayesiläinen päätösteoria on *ainoa* tässä mielessä rationaalinen päätöksentekopolitiikka [105].

Päätösteoriaa on kritisoitu siitä, että eri tilanteista seuraavien haittojen arviointi on usein kohtuuttoman vaikeaa tai jopa mahdotonta. Tämä on usein totta: esimerkissä 8 haittojen arvottaminen oli yksinkertaista, koska niitä arvioitiin pelkästään yhtiölle muodostuvien kustannusten kautta, mutta jos mukaan olisi otettu työntekijälle koituvien kärsimysten arviointi, tilanne olisi ollut paljon monimutkaisempi. Päätösteoriassa toteutuu kuitenkin

Esimerkki 8 Päätösteoreettinen riskianalyysi.

Työpaikkalääkäri saa eteensä työntekijän, jolle tehdyt alustavat testit osoittavat henkilön potevan erästä tautia 60% todennäköisyydellä. Tauti on sinänsä vaaraton, mutta jos se pääse hoitamattomana taudin myöhempään vaiheeseen, joutuu henkilö kymmenen päivän sairaslomalle. Tauti voidaan diagnosoida 100% luotettavuudella erään nopean, mutta melko kalliin lisätestin avulla. Testin hinta vastaa työntekijän yhden päivän työpanosta. Jos henkilöllä tämän testin avulla havaitaan kyseinen tauti ennen kuin se kehittyy myöhempään vaiheeseensa, voidaan tarvittavien sairaspäivien määrä vähentää puoleen.

Lääkäri pohtii, tulisiko hänen määrätä työntekijälle kallis lisätesti, vai luokitella henkilö terveeksi ja lähettää hänet takaisin töihin. Eri vaihtoehdoissa kustannukset muodostuvat siis seuraavasti, kun yksikkönä käytetään menetettyjä työpäiviä:

Aiheutuvat kustannukset:	Potilas on sairas	Potilas on terve
Tehdään lisätesti	6	1
Ei tehdä lisätestiä	10	0

Jos potilaalle tehdään lisätesti, on aiheutuvien kustannusten odotusarvo $0.6 \cdot 6 + 0.4 \cdot 1 = 4$, kun taas toisessa tapauksessa aiheutuvien kustannusten odotusarvo on $0.6 \cdot 10 + 0.4 \cdot 0 = 6$. Päätösteorian mukaan eri toimintavaihtoehdoista on valittava se, josta seuraava odotettavissa oleva kustannus on minimaalinen (tai vastaavasti odotettavissa oleva hyöty maksimaalinen), joten lääkärin tulisi määrätä potilas lisätesteihin. Jos lääkärille tulee 1000 samanlaista tapausta vuodessa, toimimalla päätösteorian mukaisesti hän säästää yhtiölleen vuosittain (keskimäärin) 2000 työpäivää vastaavan määrän rahaa.

bayesiläisen mallintamisen pääperiaate, jonka mukaan meidän on kirjattava täsmällisesti kaikki ongelmanratkaisussa käyttämämme oletukset. Jos käyttämämme järjestelmä ei toimi toivotulla tavalla, tiedämme että syy on joko siinä, että olemme arvioineet haittavaikutukset väärin, tai siinä, että emme ole pystyneet estimoimaan käytettyjä ehdollisia todennäköisyyksiä riittävän tarkasti. Päätösteoria antaa meille kuitenkin teoreettisesti optimaalisen päämäärän, jota voimme approksimoida parhaan kykymme mukaan, ja jos järjestelmämme ei toimi halutulla tavalla, voimme analysoida mikä tekemistämme approksimaatioista ei pidä paikkaansa.

Tässä yhteydessä on korostettava, että ihmisen rooli bayesiläisessä mallintamisessa on ensiarvoisen tärkeä: valittavissa olevista vaihtoehdoista seuraavien haittojen (tai hyötyjen) arviointi on tehtävä, joka jää viime kädessä ihmisen tehtäväksi. Päätösteoria voi antaa toimintasuosituksen annettujen haittavaikutusten kannalta, mutta ihmisen tehtäväksi jää arvioida tehtyjen oletusten paikkansapitävyys ja siten toimintasuositusten järkevyys. Tästä syystä epätäsmällistä päättelyä suorittavia järjestelmiä onkin ryhdytty kutsumaan yhä useammin *päätöksentukijärjestelmiksi (decision support systems)* entisen termin *asiantuntijajärjestelmä (expert system)* sijaan.

Luku 3

Bayesiläisen lähestymistavan arviointia

Bayesiläisen mallintamisen edut: ristiriidaton päättely, ylioppimisen välttäminen, tilastollisen aineiston ja asiantuntijatietämyksen yhdistäminen, puuttuvan tiedon luonteva käsittely, teoreettinen kehikko hybridimalleille. Vastauksia bayesiläisen lähestymistavan kritiikkiin. Vertailua esimerkkeihin perustuvaan päättelyyn ja ei-parametrisiin malleihin. Yhteys informaatioteoreettisiin lähestymistapoihin: MDL, MML ja stokastinen kompleksisuus.

Luvussa 2 esitettyä bayesiläistä mallinnuskehikkoa kohtaan voidaan esittää kritiikkiä kohdistamalla se joko parametriseen mallintamislähestymistapaan sinänsä, tai nimenomaan bayesiläiseen mallintamiseen. Luvussa 3.3 käsittelemme lyhyesti esimerkkeihin perustuvaa päättelyä ja ei-parametrisia malliperheitä, jotka eivät kuitenkaan osoittaudu kovin erilaisiksi lähestymistavoiksi parametriseen mallintamiseen verrattuna. Sitä ennen kertaamme luvussa 3.1 tärkeimmät bayesiläisen lähestymistavan tarjoamat edut, ja arvioimme bayesiläisyyttä kohtaan esitettyä kritiikkiä luvussa 3.2. Luvussa 3.4 käsitellään informaatioteoreettista lähestymistapaa mallintamiseen, ja osoitetaan sen läheinen yhteys bayesiläisyyteen.

3.1 Bayesiläisen mallintamisen etuja

Bayesiläinen mallinnus tarjoaa teoreettisesti elegantin, yhtenäisen lähestymistavan kaikkiin oppivien ja älykkäiden järjestelmien rakentamisessa esiintyviin ongelmiin. Mikäli käytettävä malliperhe on lisäksi probabilistinen, saavutetaan lähestymistavalla monia käytännön sovellusten kannalta merkittäviä etuja:

Ristiriidaton kalkyyli epätäsmällisen tiedon käsittelemiseksi.
Todennäköisyyslaskenta tarjoaa epätäsmällisen tiedon käsittelemiseksi matemaattisen kalkyylin, missä uskomusastetta tiettyyn väitteeseen kuvataan reaaliluvulla väliltä [0,1]. Syntyvä kalkyyli on *ristiriidaton* eli *konsistentti*, mikä tarkoittaa sitä että arvioidessamme jonkin suureen todenperäisyyden astetta, on todennäköisyyslaskennan tarjoama vastaus aina yksikäsitteinen. Mielivaltaisen lukuja väliltä [0,1] manipuloivan numeerisen järjestelmän ristiriidattomuus ei suinkaan ole itsestäänselvyys. Itse asiassa voidaan osoittaa, että tiettyjen luontevien oletusten vallitessa todennäköisyyslaskenta on ainoa olemassaoleva konsistentti epätäsmällistä tietoa käsittelevä kalkyyli. Vaikka kaikkia ko. todistuksen perustana olevia oletuksia ei hyväksyttäisikään, tämä teoreettinen tulos osoittaa että konsistentin kalkyylin luominen on erittäin vaikeaa (enemmän aiheesta löytyy mm. lähteestä [138]).

Todennäköisyyslaskennan käyttäminen poistaa yhden epätäsmällistä päättelyä suorittavien järjestelmien perusongelmista: jos esimerkiksi rakennamme sumean sääntökannan, ja käytämme jotakin mielivaltaisesti valittua kalkyyliä (esimerkiksi jotakin min-max-yhdistelysääntöjen variaatiota) sääntökannan soveltamisessa, ja järjestelmä ei toimi toivotulla tavalla, emme voi tietää onko meidän syytä vaihtaa käytettyä kalkyyliä, järjestelmän stuktuuria (sääntökannan muotoa) vai järjestelmän parametreja (jäsenyysasteita). Bayesiläisessä lähestymistavassa käytettävä kalkyyli on yksikäsitteinen, joten järjestelmän puutteiden korjaamiseksi riittää tarkastella käytetyn mallin struktuuria ja siihen liittyviä parametreja.

Ylioppimista välttävä mallinvalintakriteeri.
Kuten luvussa 2.2.1 todettiin, bayesiläinen mallinvalintakriteeri sisältää "automaattisen Occamin partaveitsen", minkä ansiosta mallin struktuurin monimutkaisuus voidaan valita siten, että mallin tarkkuus myös opetusjoukon ulkopuolella (mallin yleistyskyky) on mahdollisimman hyvä. Bayesiläinen oppiminen pyrkii siis automaattisesti tasapainoon mallien monimutkaisuuden ja niiden kuvausvoiman välillä. Vaikka bayesiläistä mallinvalintakriteeriä on luonnollisinta käyttää aidosti probabilististen malliperheiden, kuten esimerkiksi Bayes-verkkojen tapauksessa (katso luku 4.4), voidaan kriteeriä soveltaa myös muiden malliperheiden, kuten esimerkiksi neuroverkkojen [90, 99, 12, 14, 117] ja päätöspuiden [115, 103] oppimisessa.

Tilastollisen aineiston ja asiantuntijatietämyksen yhdistäminen luonnollisella, teoreettisesti tyydyttävällä tavalla.
Luvussa 2.2.4 hahmoteltiin, kuinka asiantuntijatietämys voidaan yhdistää tilastolliseen aineistoon priorijakaumia käyttäen. Toisaalta monissa probabilis-

tisissa malliperheissä (kuten esimerkiksi luvussa 4 esitettyjen Bayes-verkkojen tapauksessa) voidaan sekä mallistruktuureille että malliparametreille antaa selkeä semanttinen tulkinta, jolloin malleja voidaan konstruoida myös suoraan asiantuntijatietämystä käyttäen, ilman tilastollista oppimista (toisin kuin neuroverkkojen ja muiden ns. "black box"-mallien tapauksessa). Luvussa 4.4 sovellamme näitä periaatteita käytännössä, ja esitämme kuinka Bayes-verkkomalleja voidaan konstruoida käyttäen pelkästään tilastollista aineistoa, pelkästään asiantuntijatietämystä, tai yhdistämällä nämä kaksi tietolähdettä.

Parametrien oppimisen nopeus.
Kuten luvussa 2.3 todettiin, bayesiläisessä mallintamisessa voidaan optimaaliset malliparametrit löytää monessa tapauksessa suoraan ilman aikaavievää iteratiivista oppimisprosessia (vrt. neuroverkkojen parametrien oppiminen backpropagation-algoritmilla). Luvussa 4.4.2 annamme yksinkertaisen laskukaavan, jonka avulla Bayes-verkkomallin parametrit voidaan laskea suoraan annetusta opetusjoukosta $\mathcal{D}$.

Tarvittavan tilastollisen aineiston vähyys.
Bayesiläinen mallintaminen mahdollistaa asiantuntijatietämyksen hyväksikäytön mallien oppimisprosessissa. Jos asiantuntijatietämystä on saatavilla, on tarvittavan tilastollisen aineiston määrä hyvin pieni — itse asiassa mallit voidaan konstruoida suoraan asiantuntijatietämyksen perusteella, ilman tilastollista oppimista! Toisaalta, vaikka asiantuntijatietämystä ei olisikaan saatavilla, todellisilla aineistoilla suoritetut empiiriset kokeet [76, 77] ovat osoittaneet, että bayesiläisessä oppimisessa jo hyvin pieni esimerkkiotos (parhaissa tapauksissa vain muutamista esimerkeistä muodostuva joukko) riittää hyvän tuloksen saavuttamiseksi parametrien oppimisessa.

Monipuoliset sovellusmahdollisuudet yhteisjakauman mallintamisen ansiosta.
Kuten luvussa 2.3 näimme, bayesiläinen mallintaminen mahdollistaa hyvin monenlaisten ongelmien kuvaamisen yhdessä ristiriidattomassa teoreettisessa kehikossa. Koska probabilistiset mallit kuvaavat mallinnetussa maailmassa esiintyvien tilanteiden yhteisjakaumaa, voidaan samaa probabilistista mallia käyttää joustavasti hyvin erityyppisissä sovelluksissa.

Mahdollisuus sekoittaa jatkuvia ja diskreettejä muuttujia.
Probabilistisissa malleissa voidaan käyttää joko moniarvoisia diskreettejä

muuttujia, jatkuva-arvoisia muuttujia, tai molempia yhtäaikaa. Tähän kysymykseen palataan luvussa 4.5.

Puuttuvan tiedon käsittelymekanismit.
Kuten luvussa 2.3 näimme, probabilistinen lähestymistapa tarjoaa selkeän teoreettisen kehikon puuttuvan tiedon käsittelemiseksi: jos puuttuva tieto ei sinänsä ole kiinnostuksen kohteena, probabilistinen päättely voidaan suorittaa olemassaolevan tiedon perusteella marginalisoimalla (integroimalla) puuttuvan tiedon vaikutus pois. Puuttuva tieto voidaan toisaalta myös täydentää optimaalisella tavalla vertailemalla eri täydennysten todennäköisyyksiä.

Toimintavaihtoehtojen päätösteoreettinen analyysi.
Toisin kuin vaihtoehtoiset lähestymistavat, bayesiläinen päätösteoria tarjoaa mahdollisuuden arvioida eri toimintavaihtoehtoihin liittyviä odotettavia riskejä ja hyötyjä. Kuten luvussa 2.4 näimme, päätösteorian avulla voidaan etsiä odotusarvoisesti pienimmän riskin (tai suurimman hyödyn) tuottava toimintavaihtoehto, tai analysoida jo tehtyjen päätösten seurauksia. Päätösteoriaa voidaan luonnollisesti soveltaa myös ei-probabilististen mallien antamien tulosten käsittelyssä, jos mallien antamat tulokset esitetään todennäköisyysjakaumana. On kuitenkin huomattava, että pelkkä numeroiden suora normeeraus siten että niiden summaksi tulee yksi, ei välttämättä tuota todennäköisyysjakaumaa, joka vastaisi ongelmakentän todellista jakaumaa hyvin, eikä päätösteorian soveltaminen tällaiselle jakaumalle tuota hyviä tuloksia.

Teoreettinen kehikko hybridimallien konstruoimiseksi.
Luvussa 2.2 esitetyssä lähestymistavassa valittiin annetun opetusdatan $\mathcal{D}$ perusteella yksi mallistruktuuri M ja yksi parametriarvoilmentymä $\hat{\theta}$, joiden avulla konstruoitiin ongelmakenttää mallintava yhteistodennäköisyysjakauma $P(X_1 = x_1, \ldots, X_n = x_n \mid M, \hat{\theta}, \mathcal{D})$. On kuitenkin huomattava, että vaikka esitetty bayesiläinen kriteeri valitsee kaikkein todennäköisimmän parametrivektorin $\hat{\theta}$, saattaa malliavaruudesta löytyä myös muita parametriarvovektoreita, jotka ovat lähes yhtä todennäköisiä kuin $\hat{\theta}$. Jos päämääränä on tehdä tarkkoja ennustuksia saatavilla olevan aineiston $\mathcal{D}$ perustella, on helppo nähdä että ennustusten tarkkuutta voidaan parantaa käyttämällä yhden mallin sijasta joukkoa todennäköisiä parametrivektoreita, painottaen niiden antamia tuloksia parametrivektorien todennäköisyydellä. Itse asiassa bayesiläinen lähestymistapa ei pääty edes vielä tähän, vaan tarkasti ottaen bayesiläisessä mallintamisessa edellytetään, että optimaalisen ennustustarkkuuden saavuttamiseksi on käytettävä jakaumaa, joka saadaan laskemalla integraali

("painotettu summa") yli *kaikkien* (äärettömän monen) parametrivektorin θ:

$$P(X_1 = x_1, \ldots, X_n = x_n \mid M, \mathcal{D}) \\ = \int P(X_1 = x_1, \ldots, X_n = x_n \mid M, \theta) P(\theta \mid M, \mathcal{D}) d\theta.$$

Tämä lähestymistapa johtaa Bayes-verkkojen tapauksessa odotusarvoparametrien käyttöön, joista puhutaan luvussa 4.4.2. Toisaalta vastaavaa ajatuskulkua voidaan soveltaa myös mallistruktuurin valinnassa, jolloin päädytään käyttämään jakaumaa

$$P(X_1 = x_1, \ldots, X_n = x_n \mid \mathcal{D}) \\ = \sum_M (X_1 = x_1, \ldots, X_n = x_n \mid M, \mathcal{D}) P(M \mid \mathcal{D}).$$

Bayesiläisen filosofian mukaan siis opetusdatan perusteella saadaan tarkin ongelmakentän jakauman esitys muodostamalla painotettu summa saatavilla olevista mallistruktuureista, kun painokertoimina käytetään mallistruktuurien posterioritodennäköisyyksiä $P(M \mid \mathcal{D})$. Näin voidaan yhdistää esimerkiksi monen eri Bayes-verkon antamat tulokset, tai muodostaa vaikkapa Bayes-verkon, neuroverkon, ja päätöspuun muodostama hybridimalli. Jälkimmäisessä tapauksessa ongelmaksi saattaa kuitenkin muodostua se, kuinka konstruoida jakaumamalli, joka antaa eri malliperheistä tuleville mallistruktuureille yhteismitalliset todennäköisyydet. Lisäksi on huomattava, että hybridimalleille ei voida antaa semanttisesti yhtä yksinkertaista tulkintaa kuin yksittäisille probabilistisille malleille, joten lähestymistapa ei sovellu hyvin ongelmakentän analysointiin tähtääviin sovelluksiin (data mining).

3.2 Bayesiläisen lähestymistavan kritiikkiä

Edellisessä luvussa esitetyistä bayesiläisen mallintamisen eduista huolimatta on bayesiläinen mallinnus toistaiseksi jäänyt ehkä vähemmälle huomiolle kuin vaihtoehtoiset laskennalliseen älykkyyteen tähtäävät lähestysmistavat. Yksi syy tähän on varmaankin se, että bayesiläiseen mallinnukseen saattaa (malliperheestä riippuen) liittyä sen formaalin luonteen vuoksi joitakin teknisiä ongelmia, vaikka erityisesti luvussa 4 kuvatun Bayes-verkkomalliperheen kehittäminen onkin poistanut monet näistä ongelmista. Saavutetut tulokset ovat kuitenkin verrattain uusia (Bayes-verkkoteoriaa on tutkittu laajemmin vasta noin 10 vuoden ajan), eivätkä siksi vielä kovin laajalti tunnettuja. Maailmalla liikkuukin bayesiläisestä mallintamisesta monia harhakäsityksiä, jotka eivät pidä paikkaansa. Seuraavassa kokoelma yleisimpiä bayesiläiseen lähestymistapaan kohdistuvia kriittisiä kommentteja.

Bayesiläinen mallinnus edellyttää epärealistisen suuren parametrijoukon arvojen määräämistä.
Bayesiläisen mallinnuksen myönnetään usein olevan teoreettiselta kannalta “se oikea” lähestymistapa, mutta probabilististen mallien oletetaan edellyttävän niin suuren parametrijoukon arvojen tarkkaa määrittämistä, että lähestymistapa ei ole käytännössä mahdollinen. Luvussa 4 esitetyt Bayes-verkkomallit ovat vasta joitakin vuosia sitten kehitetty elegantti ratkaisu tähän ongelmaan: Bayes-verkkoja käyttäen tarvittavien parametrien määrä saadaan vähennettyä käytännön sovellukset mahdollistavalle tasolle. Toisaalta, kuten edellä jo todettiin, bayesiläinen mallintaminen tarjoaa mahdollisuuden oppia malleja suoraan tilastollisesta aineistosta, suoraan asiantuntijatietämystä soveltaen, tai yhdistämällä molemmat tietolähteet. Tämä mahdollisuus siirtää parametrien määräämistehtävän ihmiseltä koneelle suoritettavaksi. Lisäksi on vielä todettava, että bayesiläisten mallien on empiirisesti havaittu kestävän hyvin parametrien arvojen epätarkkuutta, jolloin parametrien arvojen tarkka määrittäminen ei ole tärkeää — käytännön sovelluksia ajatellen on usein riittävää, jos parametrit ovat oikeaa suuruusluokkaa [59].

Priorijakaumien määrääminen on vaikeaa.
Kuten edellä näimme, bayesiläinen mallintaminen edellyttää että järjestelmälle annetaan kaksi priorijakaumaa: mallistruktuurien priorijakauma $P(M)$ ja malliluokan parametrien priorijakauma $P(\theta \mid M)$. Tämä seikka mainitaan usein nurinkurisesti bayesiläisen mallintamisen heikkoutena, vaikka tosiasiassa se on yksi lähestymistavan suurimpia etuja — nimenomaan priorijakaumienhan käytön ansiosta voidaan tilastollista tietoa ja asiantuntijatietämystä yhdistää luonnollisella tavalla. Mikäli prioritietämystä ei ole saatavilla, hyvä vaihtoehto käytännössä on käyttää ns. ei-informatiivista, tasaista priorijakaumaa. Toinen käyttökelpoinen vaihtoehto on käyttää informaatioteoreettisia lähestymistapoja, joita tarkastellaan luvussa 3.4.

Bayesiläinen päättely on epäintuitiivista.
Bayesiläistä lähestymistapaa on kritisoitu siitä, että sen tarjoama epätäsmällisen päättely olisi jollakin tavoin “epäintuitiivista”, erilaista kuin ihmisten suorittama epätäsmällinen päättely. Suurin syy tähän uskomukseen on vanha *objektivistinen* todennäköisyyskäsitteen tulkinta, jonka mukaan todennäköisyydet määritellään toistokokeissa esiintyvinä frekvensseinä. Uudemman, *subjektivistisen* todennäköisyystulkinnan mukaan todennäköisyydet määritellään uskomusastetta kuvaavina lukuina väliltä $[0, 1]$. Tämän ansiosta voidaan subjektivistisessa lähestymistavassa puhua esimerkiksi todennäköisyydestä sille, että maailmanloppu tulee vuonna 2000, kun taas vanhassa fre-

kventistisessä mallissa kyseisen todennäköisyyden käyttäminen oli ongelmallista. Luvussa 4.3 annamme probabilistisesta päättelystä esimerkkejä, jotka tukevat intuitiivisesti hyvin luonnolliselta tuntuvia päättelyn muotoja (esimerkiksi ns. syiden poisselittämisilmiötä).

Bayesiläisyyden semantiikan tulkintaa haittaavat myös monet päätösteoriaan liittyvät paradokseina esitellyt esimerkkitapaukset. Tyypillisessä esimerkissä koehenkilölle annetaan kaksi vaihtoehtoa: vaihtoehdossa A hän saa kymmenen miljoona markkaa todennäköisyydellä 1 (siis varmasti), kun taas vaihtoehdossa B hän saa kymmenen miljoona markkaa todennäköisyydellä 0.25, neljäkymmentä miljoonaa markkaa todennäköisyydellä 0.25, ja ei mitään todennäköisyydellä 0.5. Vaihtoehto A on varmaankin itse kunkin mielestä paljon houkuttelevampi, vaikka päätösteoria suosittelee vaihtoehdon B valitsemista, koska tässä tapauksessa saavutettavan voiton odotusarvo on yli kymmenen miljoona markkaa! Esimerkki paljastaa kuinka tärkeää päätösteoriassa on odotettavissa olevan hyödyn mittaaminen oikein: siinä oletettavassa lähtötilanteessa, jossa koehenkilö ei ole miljonääri, ei neljänkymmenen miljoonan markan voittaminen ole hänelle neljä kertaa arvokkaampaa kuin kymmenen miljoonan voittaminen: jo kymmenen miljoonaa markkaa riittää siihen, että henkilö voi elää loppuelämänsä mukavissa olosuhteissa ilman tarvetta palkkatyöhön, joten voittosumman kasvattaminen nelinkertaiseksi ei tuo merkittävää lisäystä voiton aikaansaamaan elämänmuutokseen. Jos eri markkamääriä vastaavat todelliset arvostukset otetaan huomioon, päätösteoria toimii inhimillisesti katsoen "järkevällä" tavalla.

Lopuksi on vielä todettava, että vaikka bayesiläinen päättely olisikin jossakin mielessä epäintuitiivista, ehkäpä se olisi vain hyvä asia: ihmiset ovat monesti osoittautuneet huonoiksi päätöksentekijöiksi monimutkaisia suureita käsittelevissä tilanteissa. Kuten edellä kuvattu "paradoksi" osoittaa, ihmiselle jää oppivien ja älykkäiden järjestelmien käyttämisessä tärkeä rooli: on vaikeaa kuvitella tietokoneohjelmaa, joka pystyisi tyydyttävästi määräämään ongelmakentästä tärkeimmät olemassaolevat toimintavaihtoehdot, ja arvioimaan realistisesti niihin liittyviä hyötyjä ja haittoja.

Bayesiläinen mallintaminen käsittelee vain tiedon epävarmuutta, ei tiedon epätäsmällisyyttä.

Todennäköisyyslaskentaa on kritisoitu siitä, että se käsittelee "perinteisiä", täsmällisiä objekteja, eikä epätäsmällisyyttä kuvaavia ns. *sumeita käsitteitä* (katso esim. [60]). Koska todennäköisyysjakaumia voidaan määritellä mielivaltaisen moniarvoisille tai jatkuva-arvoisille muuttujille, voidaan bayesiläisessä mallintamisessa kuitenkin käsitellä samoja kysymyksiä kuin sumeassa logiikassakin [21, 20].

Bayesiläisen mallintamisen soveltaminen on vaikeaa.
Bayesiläisen mallintamisen edellyttämä kaikkien oletusten eksplisiittisen formalisoinnin täydellinen ymmärtäminen edellyttää tiettyjä matemaattisia valmiuksia. Tällaisen mallintamisen soveltaminen käytännössä ei kuitenkaan edellytä näiden teknisten yksityiskohtien syvempää ymmärrystä, vaan bayesiläisen mallintamisen teoriaa voidaan käyttää tuottamaan yksinkertaisia laskenta-algoritmeja, kuten lukuisat saatavilla olevat ohjelmistot todistavat (ks. luku 5.2). Toisaalta monet heuristisesti kehitetyt "ad hoc"-menetelmät voidaan tulkita todennäköisyyslaskennan tarjoamassa kehikossa. Tämän avulla on pystytty esimerkiksi selittämään ne epärealistiset riippumattomuusoletukset, jotka joudutaan (implisiittisesti) tekemään käytettäessä ns. varmuuskertoimiin perustuvia sääntökantoja [47].

3.3 Esimerkkeihin perustuva päättely ja ei-parametriset mallit

Ei-parametrisissa (non-parametric) lähestymistavoissa luovutaan mallien eksplisiittisestä, mallistruktuurin kiinnittämästä parametroidusta esitysmuodosta. *Esimerkkeihin perustuvassa päättelyssä (case-based reasoning, CBR)*[1] ongelmakentässä esiintyvät päättelyongelmat yritetään ratkaista suoraan esimerkkitapauksia soveltaen, kaksivaiheista laskentaprosessia käyttäen. Ensimmäisessä, *esimerkkien vertailuvaiheessa (case matching)* käsillä olevaa tilannetta verrataan muistissa oleviin aikaisempiin esimerkkitapauksiin (cases), ja tapauksille annetaan *yhteensopivuusaste (matching score)*, jonka arvo on sitä suurempi mitä enemmän muistissa oleva tapaus muistuttaa nykyistä tilannetta. Päättelyn toisessa, *esimerkkien sovellusvaiheessa (case adaptation)* muodostetaan ratkaisuehdotus muokkaamalla suurimman yhteensopivuusasteen saaneiden vanhojen tapausten ratkaisemisessa käytetyistä menetelmistä ehdotus nykyisen tilanteen ratkaisemiseksi.

Yksinkertaisin esimerkki CBR-tyyppisestä lähestymistavasta on ns. *lähimmän naapurin menetelmä (nearest neighbor method)*, jossa esimerkkien yhteensopivuusasteena käytetään niiden (euklidista) etäisyyttä, ja tarvittavat tiedot kopioidaan suoraan korkeimman yhteensopivuusasteen saaneesta esimerkkitapauksesta. Monimutkaisemmissa CBR-malleissa tarvittavat tiedot muodostetaan käyttämällä opetusjoukon tapauksien painotettua summaa, käyttäen painokertoimina tapausten yhteensopivuusasteita. Esimerk-

[1] Lisätietoa esimerkkeihin perustuvasta päättelystä löytyy esimerkiksi lähteistä [70, 144]. Lähestymistavasta käytetään myös termejä *memory-based reasoning* [136], *instance-based learning* [2], *lazy learning* [1] ja *transductive inference* [142].

keihin perustuvan päättelyn toimintaperiaate on hyvin samanlainen kuin tilastotieteessa *ydinestimaattoreiksi (kernel estimators)* kutsutuilla menetelmillä (katso esim. [126]). Ydinestimaattoreissa muodostetaan annetun opetusjoukon kunkin tapauksen ("ytimen") ympärille jakauma, jonka keskikohta on ytimessä, ja kokonaisjakauma saadaan tällaisten ydinjakaumien summana. Neuroverkkotutkimuksen alueella vastaavia menetelmiä kutsutaan *ydinkantafunktioiksi (radial basis functions)* [95, 110] tai *probabilistisiksi neuroverkoiksi (probabilistic neural networks)* [133].

On tärkeää huomata, että nimestään huolimatta ei-parametriset mallit eivät suinkaan ole parametrittomia: esimerkiksi vaikka ydinestimaattoreissa on ydinjakaumien määrä ja niiden keskikohta kiinnitetty, on jakaumien tarkka muoto määrättävä joukolla parametreja (esim. ydinjakaumakohtaiset varianssit), jotka on arvioitava esimerkkiaineistosta. Esimerkkeihin perustuvassa päättelyssä puolestaan on tapausten yhteensopivuusasteen laskemisessa käytettävä metriikka käyttäjän vapaasti määrättävissä, ja eri metriikat voidaan ajatella jonkin (mahdollisesti hyvin monimutkaisen) malliperheen parametrisoituina malleina. Se, halutaanko puhua eri metriikoista ja mallien soveltamisessa käytettävissä laskentamenetelmistä, vai parametrisista malleista, näyttää olevan pitkälti makuasia.

Bayesiläisiin parametrisiin malleihin perustuvilla lähestymistavoilla on se etu, että menetelmä pakottaa määrittelemään täsmällisesti kaikki oletukset, joiden varassa rakennettavan järjestelmän toiminta on. Ilman tällaista lähestymistapaa päädytään ad hoc-tyyppisiin menetelmiin, joiden toimivuudesta ei ole teoreettisia takeita. Voidaan esimerkiksi kysyä, mihin perustuu se useimmissa CBR-menetelmissä implisiittisesti tehtävä oletus, että euklidisen etäisyysmitan mielessä toisiaan muistuttavien tapausten ratkaisemisessa voidaan soveltaa samoja menetelmiä? Tämä oletus saattaa tietenkin pitää paikkansa ongelmasta, ja sen tapausten koodaamisessa käytetystä menetelmästä riippuen, mutta menetelmien käyttäminen ilman tehtyjen implisiittisten oletusten tiedostamista saattaa johtaa vakaviin virhepäätelmiin.

Suurin ero ei-parametristen lähestymistapojen ja parametrisen mallintamisen välillä näyttää olevan siis lähinnä siinä, että ei-parametrisissa malleissa parametrien lukumäärä ei ole kiinteä, vaan se riippuu käytettävissä olevan opetusjoukon koosta. Kuten edellä kuitenkin näimme, bayesiläisessä mallintamisessa on mallistruktuurin valinta kiinteä osa oppimisprosessia, ja valittava mallistruktuuri (ja sitä kautta malliparametrien lukumäärä) riippuu käytettävästä opetusjoukosta[2]! Parametristen ja erilaisten ei-parametristen lähestymistapojen ero ei siis näytä olevan periaattellinen, vaan lähinnä ter-

[2]Tästä syystä tähän periaatteeseen perustuvia malleja kutsutaan joskus *semiparametrisiksi (semi-parametric)* malleiksi.

minologinen. Itse asiassa sekä esimerkkeihin perustuva päättely että ydinestimaattorit (ja muut vastaavat mallit) voidaan ajatella erikoistapauksena *äärellisistä sekajakaumamalleista (finite mixture models)* [35, 141], joita käsitellään enemmän luvussa 4.5. Äärellisten sekajakaumamallien ja esimerkkeihin perustuvan päättelyn samankaltaisuus osoitetaan lähteissä [140, 139, 98]. Yleisempää todennäköisyyslaskentaan perustuvaa formalismia esimerkkeihin perustuvalle päättelylle on ehdotettu lähteessä [75]. On huomattava, että ehdotetut bayesiläiset formalismit eivät välttämättä noudata esimerkkeihin perustuvan päättelyn normaalia kaksivaiheista päättelyprosessia, vaan päättely voi tapahtua suoraan ilman esimerkkien yhteensopivuusasteen laskemista. Bayesiläinen menetelmä tapausten yhteensopivuusasteen laskemiseksi esitetään lähteessä [71].

3.4 Informaatioteoreettiset lähestymistavat: MDL ja MML

3.4.1 MDL-periaate ja stokastinen kompleksisuus

Mallistruktuurin valinnassa voidaan käyttää myös informaatioteoreettista lähestymistapaa, joka on hyvin läheisessä suhteessa bayesiläiseen mallintamiseen. Rissasen *Minimum Description Length (MDL)*-periaatteen [113, 114, 115, 116] mukaan mallistruktuureista on valittava se, jonka avulla opetusjoukko $\mathcal{D}$ voidaan koodata siten, että syntyvä koodi on pituudeltaan mahdollisimman lyhyt. Rissanen kutsuu lyhintä tällaista koodinpituutta opetusjoukon $\mathcal{D}$ *stokastiseksi kompleksisuudeksi (stochastic complexity)*, annettuna mallistruktuuri M. Tällä koodausteoreettiselle lähestymistavalle voidaan antaa intuitiivisesti selkeä tulkinta: lyhimmän koodin tuottava mallistruktuuri on ongelmakentän paras kuvaus siksi, että lyhyen koodin aikaansaamiseksi on mallistruktuuriin koodattava kaikki mahdolliset aineistossa $\mathcal{D}$ esiintyvät säännönmukaisuudet.

Kuinka stokastinen kompleksisuus sitten määritellään? Ensiksikin on korostettava, että on mahdotonta konstruoida määritelmää joka tuottaisi lyhimmän koodin kaikille mahdollisille opetusjoukoille $\mathcal{D}$: Koodausteorista tiedämme että jos $L(\mathcal{D} \mid M)$ kuvaa opetusjoukkoa $\mathcal{D}$ vastaavan koodin pituutta mallistruktuurin M tapauksessa, määrittelee

$$P(\mathcal{D} \mid M) = 2^{-\log L(\mathcal{D} \mid M)}$$

todennäköisyysjakauman yli kaikkien mahdollisten joukkojen $\mathcal{D}$. Koska

$$\sum_{\mathcal{D}} P(\mathcal{D}|M) = 1,$$

ei ole mahdollista että olisi olemassa sellainen yksittäinen todennäköisyysjakauma, joka antaisi kaikille datajoukoille $\mathcal{D}$ suuremman todennäköisyyden kuin muut todennäköisyysjakaumat. Niinpä ei myöskään voi olla sellaista koodia L, joka tuottaisi lyhimmän mahdollisen kuvauksen kaikille joukoille $\mathcal{D}$. Stokastinen kompleksisuus täytyykin määritellä siten, että se tuottaa lyhyimmän kuvauksen "tyypillisimmille" joukoille $\mathcal{D}$. Se kuinka "tyypillisyys" määritellään, ratkaisee viime kädessä eri määritelmien hyvyyden.

Varhaisessa tuotannossaan [113, 114] Rissanen määritteli stokastisen kompleksisuuden seuraavasti:

$$L(\mathcal{D} \mid M) = -\log P(\mathcal{D} \mid M, \tilde{\theta}) + \frac{d(M)}{2} \log N, \tag{3.1}$$

missä $(M, \tilde{\theta})$ on malliluokan M *suurimman uskottavuuden malli* (maximum likelihood model), ts. malli, joka maksimoi uskottavuustermin $P(\mathcal{D} \mid M, \theta)$, $d(M)$ on mallin parametrien lukumäärä, ja N opetusjoukon $\mathcal{D}$ koko. Näemme, että tämä informaatioteoreettinen stokastisen kompleksisuuden määritelmä antaa saman tuloksen kuin luvussa 2.2.2 esitetty kokonaisuskottavuuden BIC-approksimaatio.

Koodin (3.1) jälkimmäinen termi voidaan tulkita koodinpituudeksi, joka tarvitaan mallin $\tilde{\theta}$ koodaamiseksi, ja ensimmäinen termi koodinpituudeksi, joka tarvitaan opetusjoukon $\mathcal{D}$ koodaamiseksi, annettuna malli $(M, \tilde{\theta})$. Tälla tavalla muodostettua koodia kutsutaan *kaksiosaiseksi koodiksi* (two-part code). Myöhemmin (katso esim. [115]) Rissanen osoitti, että määritelmä

$$L(\mathcal{D} \mid M) = -\log P(\mathcal{D} \mid M) = -\log \int P(\mathcal{D} \mid M, \theta) P(\theta \mid M) d\theta \tag{3.2}$$

tuottaa koodeja, joiden voidaan osoittaa olevan monessa suhteessa parempia kuin määritelmän (3.1) tuottamat kaksiosaiset koodit. Määritelmästä (2.1) näemme että tämä kriteeri vastaa bayesiläistä lähestymistapaa, jossa mallistruktuurien prioriksi on oletettu tasainen jakauma, ts. $P(M_1) = P(M_2) = \ldots = P(M_K)$, jolloin mallistruktuurin valintakriteerinä käytetään siis yksinomaan kokonaisuskottavuutta $P(\mathcal{D} \mid M)$.

Äskettäin [116] Rissanen esitteli uuden, tietyssä mielessä vielä tehokkaamman koodin, jossa stokastinen kompleksisuus määritellään seuraavasti:

$$L(\mathcal{D} \mid M) = -\log \frac{P(\mathcal{D} \mid \tilde{\theta}(\mathcal{D}), M)}{\sum_{\mathcal{D}'} P(\mathcal{D}' \mid \tilde{\theta}(\mathcal{D}'), M)}, \tag{3.3}$$

missä $\tilde{\theta}(\mathcal{D})$ on opetusjoukkoa $\mathcal{D}$ vastaava suurimman uskottavuuden malli, ja summa nimittäjässä käy yli kaikkien mahdollisten opetusjoukkojen. Voidaan osoittaa, että tämä uusi stokastisen kompleksisuuden määritelmä antaa asymptoottisesti saman tuloksen kuin määritelmä (3.2), jos käytämme integraalissa (3.2) parametrien priorijakaumana erityistä ns. Jeffreyn prioria. Tätä kysymystä käsitellään tarkemmin lähteissä [79, 77, 78].

3.4.2 MML-periaate

Wallacen kehittämä *MML (minimum message length)*-periaate on läheistä sukua Rissasen MDL-periaatteelle, mutta MML-lähestymistavassa keskitytään malliparametrien valitsemisen ongelmaan, kun taas MDL-periaatetta sovelletaan yleensä mallistruktuurin valintaongelmaan. MML-lähestymistavassa hylätään edellä esitetty mallien kaksivaiheinen oppiminen, ja sulautetaan kaikki malliluokat M yhdeksi suureksi kaikkien mallien muodostamaksi malliperheeksi. Koska malliluokan käsitettä ei tässä lähestymistavassa käytetä, malliperhe muodostuu erikokoisista malleista eli eripituisista parametrivektoreista θ.

MML-periaatteen mukaan *paras malli $\bar{\theta}$ on se malli, joka tuottaa odotusarvoisesti lyhimmän koodin silloin, kun ensin koodataan käytetty malli $\bar{\theta}$, ja sitten opetusaineisto $\mathcal{D}$ käyttäen mallia $\bar{\theta}$*:

$$\bar{\theta} = \arg\min_{\theta} L(\theta) + L(\mathcal{D} \mid \theta).$$

MML-lähestymistavassa keskitytään siis aina kaksiosaisiin koodeihin, toisin kuin stokastisen kompleksisuuden uudemmissa laskennallisissa muodoissa (3.2) ja (3.3).

Jos mallien parametrit oletetaan reaaliluvuiksi, ei MML-optimaalista koodia voida konstruoida, koska puhtaiden reaalilukujen esittäminen edellyttää kykyä esittää lukuja äärettömän tarkasti, mikä taas vaatii äärettömän pitkiä koodeja. Siksi MML-lähestymistavassa oletetaan, että mallien parametrit pystytään esittämään vain äärellisellä tarkkuudella (mikä tietysti on realistinen oletus käytännön sovelluksia ajatellen), ja mallit θ diskretisoidaan äärellisen koodin saamiseksi. Diskretisointi on tehtävä niin, että syntyvän koodin pituus on mahdollisimman lyhyt.

Baysiläisessä oppimisessa toisaalta yksittäisen mallin θ soveltuvuus mitattiin sen posterioritodennäköisyydellä

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}. \tag{3.4}$$

Jos unohdamme nimittäjässä olevan vakion $P(\mathcal{D})$, saamme

$$-\log P(\theta \mid \mathcal{D}) \propto -\log P(\mathcal{D} \mid \theta) - \log P(\theta).$$

Kun merkitsemme $L(\theta) = -\log P(\theta)$ ja $L(\mathcal{D} \mid \theta) = -\log P(\mathcal{D} \mid \theta)$, näemme että MML-periaate on bayesiläinen siinä mielessä, että se valitsee mallin, joka maksimoi posterioritodennäköisyyden (3.4), kun mallit on ensin diskretisoitu optimaalisella tavalla. MML-periaatteesta löytyy lisämateriaalia lähteistä [79, 102, 101, 6], ja Monash-yliopiston tutkimusryhmän ylläpitämältä WWW-sivulta[3].

[3] URL: http://www.cs.monash.edu.au/~lloyd/tildeMML/

Luku 4

Bayes-verkot

Mitä ovat Bayes-verkot? Kausaalinen mallintaminen Bayes-verkoilla. Päättely epätäydellisillä tiedoilla Bayes-verkkoja käyttäen. Bayes-verkkojen muodostaminen automaattisesti: verkon rakenteen oppiminen ja parametrien oppiminen. Bayes-verkkojen muunnelmia: päätösverkot, jatkuva-arvoiset Bayes-verkot, sekajakaumamallit, Naiivi Bayes-malli, kvalitatiiviset Bayes-verkot.

Kuten kuvasta 2.1 näemme, ensimmäinen askel bayesiläisessä mallinnusprosessissa on malliperheen valinta. Vaikka bayesiläistä mallinnusta voi periaatteessa soveltaa myös ei-probabilistisiin malliperheisiin, kuten esimerkiksi neuroverkkoihin (ks. [90, 99, 12, 14, 117]), saavutetaan bayesiläisen lähestymistavan edut helpoimmin käyttäen probabilistisia malleja. *Bayes-verkot* (Bayesian networks, Bayesian belief networks) muodostavat tällaisen malliperheen, joka on saavuttanut viime aikoina suurta huomiota. Intuitiivisesti ottaen Bayes-verkkoja voidaan ajatella malleina, joissa mallin struktuuri esitetään solmuista ja niitä yhdistävistä kaarista muodostuvana verkkorakenteena. Verkon solmut vastaavat ongelmakentän määrittelemisessä käytettäviä muuttujia, ja verkon kaaret kuvaavat muuttujien välisiä riippuvuuksia. Mallin parametrit muodostuvat joukosta ehdollisia todennäköisyyksiä, jotka kuvaavat muuttujien välisten riippuvuuksien voimakkuuksia.

Pääsyy Bayes-verkkoihin kohdistuvaan kiinnostukseen johtuu 80-luvulla kehitetystä teoreettisesta kehikosta, jonka avulla mallien luomiseen tarvittavien parametrien määrää voidaan huomattavasti pienentää, ja johon perustuen on kehitetty tehokkaita probabilistisia päättelyalgoritmeja. Luvussa 4.1 selvitetään lyhyesti Bayes-verkkojen pääperiatteet, puuttumatta tässä yhteydessä tarkemmin teknisiin yksityiskohtiin, jotka selviävät esimerkiksi lähteistä [108, 100, 62, 18]. Luvussa 4.3 näemme, kuinka luvussa 2.3 esitettyjä probabilistisen päättelyn eri muotoja on mahdollista toteuttaa Bayes-verkkomallien avulla. Bayes-verkkojen rakentamista käsitellään luvussa 4.4.

4.1 Bayes-verkkomalliperhe

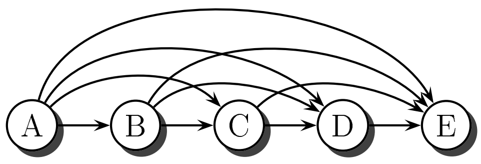
Esitämme seuraavassa Bayes-verkkomalliperheen perusominaisuudet yksinkertaistettujen esimerkkien avulla. Tässä yhteydessä on korostettava, että vaikka käytetyissä esimerkeissä esiintyy yleensä vain binäärisiä diskreettejä muuttujia, voidaan lähestymistavassa käyttää myös sekä moniarvoisia diskreettejä muuttujia että jatkuvia muuttujia. Tähän kysymykseen palaamme luvussa 4.5.

Oletetaan seuraavassa esimerkissä, että ongelmakentän kuvauksessa käytetään viittä binääristä attribuuttia (satunnaismuuttujaa) A, B, C, D, E, ja käytetään binäärimuuttujien mahdollisista arvoista merkintöjä 0 ja 1. Voimme ajatella, että arvoasetus 'A=1' vastaa tilannetta, jossa attribuutin A mallintaman ongelmakentän piirteen on havaittu olevan 'tosi'. Tässä matemaattisessa mallissa maailman kaikki mahdolliset tilat koostuvat muuttujien A, B, C, D, E arvojen muodostamista binäärivektoreista, joita on tässä yksinkertaisessa viiden binäärimuuttujan tapauksessa kaikkiaan $2^5 = 32$ kappaletta. Bayesiläisessä mallinnuksessa on päämääränä muodostaa matemaattinen malli, jonka avulla voidaan määrätä todennäköisyys mille tahansa mahdolliselle maailman tilalle (5-komponenttiselle binäärivektorille). Tällaista *yhteistodennäköisyysjakaumaa* hyväksikäyttäen voidaan suorittaa luvussa 2.3 esitettyjä probabilistisen päättelyn eri muotoja, kuten luvussa 4.3 tulemme näkemään.

Bayesiläisen mallintamisen päämääränä oleva yhteistodennäköisyysjakauma voidaan luonnollisesti määrätä luettelemalla mallimaailman kaikki mahdolliset tilat ja niitä vastaavat todennäköisyydet, mutta koska tilojen lukumäärä on eksponentiaalinen suhteessa muuttujien lukumäärään, ei tämä lähestymistapa ole käytännössä esiintyvissä tilanteissa mahdollinen. Toisaalta todennäköisyyslaskennan perusaksioomia käyttäen on helppo osoittaa, että

$$P(A, B, C, D, E) = \\ P(A)P(B|A)P(C|A, B)P(D|A, B, C)P(E|A, B, C, D). \quad (4.1)$$

Muuttujien A, B, C, D, E yhteistodennäköisyysjakauman määräämiseksi siis riittää, että tunnemme yhtälössä (4.1) oikealla puolella esiintyvät ehdolliset todennäköisyydet. Tässä yhteydessä on huomattava, että tarvittavien ehdollisten todennäköisyyksien muoto riippuu käytetystä muuttujien järjestyksestä. Jos yhtälössä (4.1) käytettyä ketjusääntöä sovelletaan esimerkiksi muuttujien järjestykseen (E, D, C, B, A), saadaan yhtälön oikealle puolelle ehdolliset todennäköisyydet $P(E)$, $P(D \mid E)$, $P(C \mid E, D)$, $P(B \mid E, D, C)$ ja $P(A \mid E, D, C, B)$. Käytettyä permutaatiota (ja siis sitä vastaavaa ehdollisten todennäköisyyksien joukkoa) voidaan havainnollistaa graafisesti: kuvas-



Kuva 4.1: Hajotelman (4.1) esitys suunnattuna verkkona.

sa 4.1 on esitetty syklitön suunnattu verkko, jossa muuttujasta Y on muuttujaan X kaari ainostaan silloin, kun jokin hajotelmassa (4.1) käytettävistä ehdollisista todennäköisyyksistä on muotoa $P(X \mid *, Y, *)$.

Kuvassa 4.1 esitetty suunnattu verkko määrää siis minkä tyyppisiä ehdollisia todennäköisyyksiä on käytettävä, jos halutaan määrätä muuttujien yhteisjakauma $P(A, B, C, D, E)$ käyttäen ketjusääntöä (4.1). Kiinnittämällä verkkoesitystä vastaavien ehdollisten todennäköisyyksien arvot on yhteisjakauma yksikäsitteisesti määrätty. Toisaalta, kuten luvussa 2.2 esitettiin, bayesiläisessä mallintamisessa malli määritellään parina (M, θ), missä M eli mallin struktuuri määrää mitä parametreja mallin identifioimiseksi on määrättävä, ja θ on joukko parametriarvoja, jotka kiinnittävät yhden mallin. Kuvassa 4.1 esitettyä graafista verkkoesitystä voidaan siis ajatella mallin strukturiina, ja verkkoa vastaavia ehdollisia todennäköisyyksiä voidaan pitää mallin parametreina.

Koska $P(X = 1|*) = 1 - P(X = 0|*)$, voimme laskea, että kuvassa 4.1 esitettyä mallistruktuuria käytettäessä tarvitsemme $1 + 2 + 4 + 8 + 16 = 31$ parametria (ehdollisen todennäköisyyden arvoa) yhteistodennäköisyysjakauman määräämiseksi. Koko avaruuden tilojen lukumäärään (32) nähden emme ole siis säästäneet paljoa. Keskeinen tekijä Bayes-verkkoteoriassa onkin ehdollisen riippumattomuuden käsite, jota käyttäen tarvittaen parametrien määrää voidaan radikaalisti pienentää:

Määritelmä 1 (Ehdollinen riippumattomuus)
Olkoot $\mathbf{X}$, $\mathbf{Y}$, *ja* $\mathbf{Z}$ *erillisiä muuttujajoukkoja. Joukon* $\mathbf{X}$ *sanotaan olevan ehdollisesti riippumaton muuttujajoukosta* $\mathbf{Y}$, *annettuna joukko* $\mathbf{Z}$, *jos*

$$P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}).$$

Selventääksemme ehdollisen riippumattomuuden käsitettä olettakaamme esimerkiksi, että ongelmakenttämme liittyy lääketieteelliseen diagnosointiin, ja esimerkkimuuttujillamme A, B, C, D, E on taulukossa 4.1 esitetyt semanttiset tulkinnat[1].

[1]Esimerkki on teoksesta [108, sivu 196].

Merkintä	*Tulkinta*
$A = 1$	Tutkittavalla henkilöllä on aivosyöpä.
$B = 1$	Tutkittavan henkilön veren kalsiumpitoisuus on noussut epänormaalille tasolle.
$C = 1$	Tutkittavalla henkilöllä on aivokasvain.
$D = 1$	Tutkittava henkilö saa tajuttomuuskohtauksia.
$E = 1$	Tutkittava henkilö kärsii vakavista pääkivuista.

Taulukko 4.1: Lääketieteellisessä esimerkissä käytetyt merkinnät.

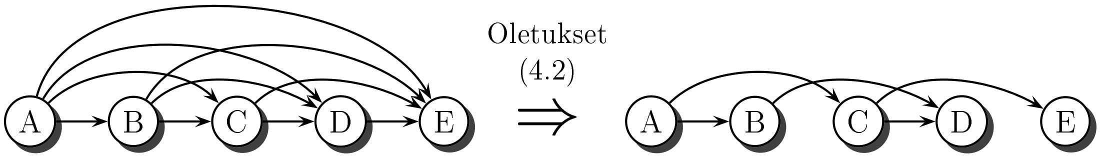
Ongelmakentän hyvin tunteva asiantuntija saattaisi kertoa muuttujien suhteesta esimerkiksi seuraavaa:

> Vakavien päänsärkykohtausten todennäköisyys $P(E = 1)$ riippuu ainoastaan siitä, onko henkilöllä aivokasvain vain ei (C). Toisaalta, jos tiedetään henkilön veren kalsiumpitoisuus (B) ja se, onko henkilöllä kasvain vai ei (C), voidaan määrätä tajuttomuuskohtauksien todennäköisyys $P(D = 1)$, joka ei tällöin riipu päänsäryn esiintymisestä (E), tai (suoraan) siitä, onko henkilöllä aivosyöpä (A) vai ei. Aivokasvaimen (C) esiintymisen todennäköisyys riippuu vastaavasti suoraan ainoastaan siitä, onko henkilöllä aivosyöpä (A), ei muista tekijöistä.

Yllä mainitut tiedot voidaan toisaalta myös päätellä tilastollisesta aineistosta ilman asiantuntijan apua, mikäli saatavilla on riittävä määrä esimerkkitapauksia. Ehdollisen riippumattomuuden käsitettä soveltaen hajotelmassa (4.1) esiintyneet ehdolliset todennäköisyydet voidaan saadun informaation valossa laskea nyt seuraavasti:

$$
\begin{aligned}
P(E \mid A, B, C, D) &= P(E \mid C) \\
P(D \mid A, B, C) &= P(D \mid B, C) \\
P(C \mid A, B) &= P(C \mid A)
\end{aligned}
\tag{4.2}
$$

Kun otamme nämä ehdolliset riippumattomuudet huomioon, kuvassa 4.1 esitetty verkko muuttuu kuvassa 4.2 esitettyyn muotoon. Kun edelleen piirrämme verkon hieman eri tavalla, ja merkitsemme kuvaan selvyyden vuoksi kutakin solmua vastaavan muuttujan semanttisen tulkinnan, saamme kuvassa 4.3 esitetyn Bayes-verkkostruktuurin.

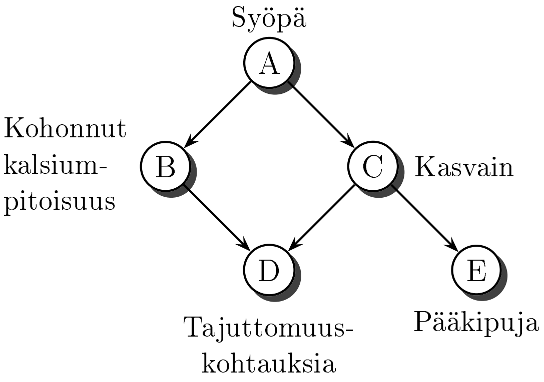


Kuva 4.2: Ehdollisen riippumattomuuden vaikutus verkkoesitykseen.

Bayes-verkot määritellään formaalisti seuraavasti:

Määritelmä 2 (Bayes-verkkomalli)
Bayes-verkkomalli on pari (M, θ)*, missä* M *on syklitön suunnattu verkko, jonka solmut vastaavat ongelmakentän muuttujia, ja jonka topologia täyttää seuraavan ehdon: kukin muuttuja on ehdollisesti riippumaton kaikista muuttujista, jotka eivät ole muuttujan jälkeläisiä verkossa, annettuna muuttujan edeltäjät, ja muuttujan edeltäjien muodostama joukko on suppein muuttujajoukko, joka täyttää tämän ehdon. Parametrijoukon* θ *muodostavat muotoa* $P(X \mid F(X))$ *olevat ehdolliset todennäköisyydet, missä* $F(X)$ *on muuttujan* X *edeltäjien joukko verkossa* M*.*

Tässä yhteydessä on korostettava, että vaikka sekä neuroverkko että Bayes-verkko ovat molemmat verkkoesityksiä, esitysten abstraktiotasossa on huomattava ero. Neuroverkothan ovat lähinnä *funktionaalisen* tason kuvauksia, siinä mielessä että neuroverkkoalgoritmien voidaan ajatella toimivan neuroverkkostruktuurin kaltaisessa arkkitehtuurissa, jossa verkon solmut ovat yksinkertaisia laskentayksiköitä, jotka lähettävät toisilleen yksittäisistä reaaliluvuista muodostuvia sanomia. Bayes-verkko puolestaan on *käsitteellisen (symbolisen)* tason esitys: Bayes-verkon solmut vastaavat ongelmakentän kuvaamisessa käytettyjä attribuutteja, ja kaaret kuvaavat attribuuttien välisiä riippuvuuksia. Bayes-verkon rakenne ei siis välttämättä suoraan kuvaa Bayes-verkkoalgoritmien toimintaa samalla tavoin kuin neuroverkon rakenne kuvaa neuroverkkoalgoritmien toimintaa (tähän kysymykseen palaamme luvussa 4.3). Erityisesti on huomattava, että toisin kuin neuroverkoissa, joissa kuhunkin kaareen liittyy yksi parametri (kaaren paino), Bayes-verkoissa parametrit eivät liity suoranaisesti verkon kaariin, vaan kutakin verkon *solmua* kohden tarvitaan joukko parametreja, joiden lukumäärä on eksponentiaalinen solmuun saapuvien kaarien (solmun edeltäjien) lukumäärän suhteen. Tarvittavien parametrien lukumäärän pienentämiseksi voidaan tätä parametrijoukkoa approksimoida lineaarisella määrällä parittaisia, muotoa $P(X = x \mid Y = y)$ olevia parametreja, joiden voidaan ajatella liittyvän verkon kaariin; yhtä tällaista approksimaatioita (ns. Noisy-Or-mallia) käsi-



$P(A=1)$	$=$	0.20
$P(B=1 \mid A=1)$	$=$	0.80
$P(B=1 \mid A=0)$	$=$	0.20
$P(C=1 \mid A=1)$	$=$	0.20
$P(C=1 \mid A=0)$	$=$	0.05
$P(D=1 \mid B=1, C=1)$	$=$	0.80
$P(D=1 \mid B=0, C=1)$	$=$	0.80
$P(D=1 \mid B=1, C=0)$	$=$	0.80
$P(D=1 \mid B=0, C=0)$	$=$	0.05
$P(E=1 \mid C=1)$	$=$	0.80
$P(E=1 \mid C=0)$	$=$	0.60

Kuva 4.3: Esimerkki Bayes-verkkomallista.

tellään tarkemmin luvussa 4.5.

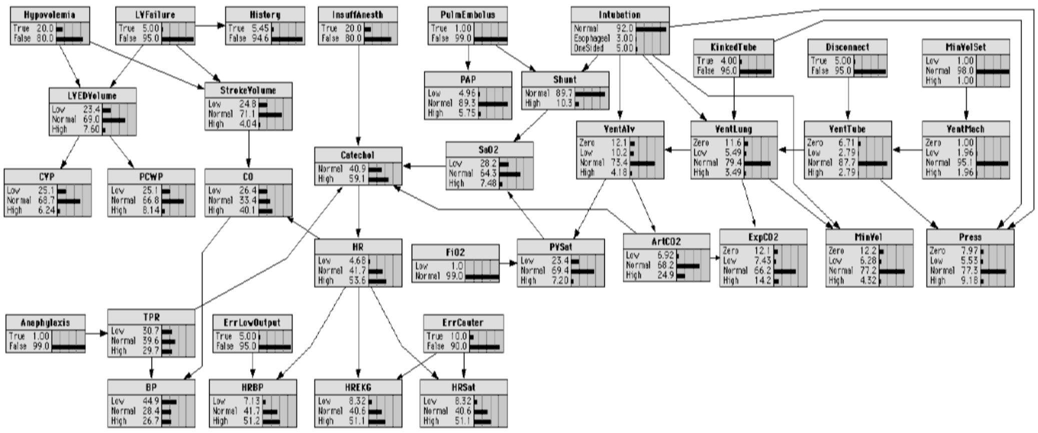
Kuvassa 4.3 on annettu esimerkki Bayes-verkkomallista kiinnittämällä verkon parametrit lähteessä [108, sivu 197] annettuihin arvoihin (on muistettava että kyseessä on yksinkertaistettu, rajoitettu esimerkki, joten annettuihin parametriarvoihin ei pidä suhtautua lääketieteelliseltä kannalta liian vakavasti). Kaikkien mahdollisten Bayes-verkkomallien (kaikki mahdolliset Bayes-verkkostruktuurit + kaikki mahdolliset struktuuria vastaavien parametrien asetukset) muodostamaa joukkoa kutsutaan seuraavassa *Bayes-verkkomalliperheeksi*. Bayes-verkkojen oppimisella tarkoitetaan yhden Bayes-verkkomallin etsimistä tästä joukosta. Luvussa 4.4 paneudumme tähän kysymykseen tarkemmin.

Bayes-verkkojen määritelmästä ja ehdollisen riippumattomuuden käsitteestä seuraa tärkeä Bayes-verkkomallien avulla koodattujen todennäköisyysjakaumien ominaisuus: jos ongelmakentän kuvauksessa käytetyt muuttujat ovat $X_1, \ldots, X_n$, niin minkä tahansa muuttujien arvoasetuskombinaation yhteistodennäköisyys saadaan laskettua kaavalla

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid F(X_i) = f(X_i)), \qquad (4.3)$$

missä kuten edellä, $F(X_i)$ on muuttujan X_i edeltäjät käytetyssä Bayes-verkossa, ja $f(X_i)$ joukkoa $F(X_i)$ vastaava arvovektori. Näemme, että Bayes-verkkomalli (M, θ) on probabilistinen, sillä annetun verkkostruktuurin parametrit kiinnittävät yhteistodennäköisyysjakauman muuttujille $X_1, \ldots, X_n$. Luvussa 4.3 esitämme, kuinka hajotelmaa (4.3) voidaan käyttää hyväksi toteutettaessa luvussa 2.3 esitettyjä probabilistisen päättelyn muotoja.

Kuvassa 4.3 esitetyssä Bayes-verkkomallissa on tarvittavien parametrien lukumäärä 11, mikä on edelleen yli 30% koko avaruuden koosta. On kuitenkin huomattava, että tarvittavat parametrit, kuten esimerkiksi $P(B|A)$,



Kuva 4.4: Sairaalapotilaiden valvonnassa käytettävä ALARM-verkko.

ovat muodoltaan sangen yksinkertaisia, ja siten myös helposti asiantuntijoiden määrättävissä, toisin kuin käsitteellisesti hankalat yhteistodennäköisyydet $P(A, B, C, D, E)$. Lisäksi on korostettava, että tarvittavien parametrien suhde avaruuden kokoon pienenee dramaattisesti muuttujien lukumäärän kasvaessa: kuvassa 4.4 on sairaalapotilaiden monitoroinnissa käytetty 37-solmuinen ALARM-verkkona tunnettu Bayes-verkkostruktuuri (katso [8]), jossa tarvittavien parametrien (ei näytetty kuvassa) lukumäärä on 509, mikä on ainoastaan $2,94 \cdot 10^{-12}\%$ koko avaruuden koosta ($2^{13} \cdot 3^{17} \cdot 4^{7} = 1,73 \cdot 10^{16}$). Kuvassa esitetyt todennäköisyydet ovat eri muuttujien arvojen prioritodennäköisyyksiä, jotka saadaan laskettua ALARM-verkon parametreista. Kuva on tuotettu Norsys-yhtiön Netica-ohjelmistolla, joka on Bayes-verkkojen konstruoimiseen ja niiden käyttämiseen tarkoitettu sovelluskehitin (katso luku 5.2).

Käytännön sovellusten kannalta on kuvassa 4.4 esitetty Bayes-verkkokin edelleen melko pieni — Bayes-verkko-ohjelmistoja valmistava HUGIN-yhtiö ilmoittaa teollisessa käytössä olevan yli tuhat solmua sisältäviä Bayes-verkkoja. Tässä yhteydessä on muistettava, että toisin kuin neuroverkoissa, ei Bayes-verkoissa solmujen lukumäärä ole vapaa parametri, vaan se määräytyy ongelmakentän kuvaamisessa käytettävien muuttujien lukumäärästä: jo sadallakin solmulla (muuttujalla) voidaan kuvata sangen monimutkaisia ongelmakenttiä. Toisaalta on korostettava, että vaikka solmujen lukumäärä on kiinnitetty, ei Bayes-verkkojen konstruoiminen ole kuitenkaan välttämättä yksinkertaista, koska mahdollisten verkkotopologioiden lukumäärä kasvaa erittäin nopeasti solmujen lukumäärän suhteen (katso luku 4.4).

4.2 Bayes-verkot ja kausaalisuus

Kuvassa 4.3 esitetyssä esimerkissä voidaan Bayes-verkkostruktuurin kaarille antaa selkeä kausaalinen tulkinta: esimerkiksi muuttujan A, joka on muuttujien B ja C edeltäjä verkossa, voidaan ajatella olevan muuttujien B ja C kausaalinen *syy*. Mielenkiintoinen kysymys onkin, kuinka Bayes-verkkostruktuurin kaaret suhtautuvat kausaalisen syy-seuraussuhteen käsitteeseen — onko kaikkien Bayes-verkkojen kaarilla sama kausaalinen tulkinta?

Yleisesti ottaen vastaus tähän kysymykseen on kielteinen: Bayes-verkon kaarien ei välttämättä tarvitse kuvastaa kausaalisia syy-seuraussuhteita. Tätä seikkaa havainnollistaa seuraava yksinkertainen esimerkki: olkoon ongelmakentän kuvauksessa käytettävät muuttujat A ja B, joilla on seuraavat semanttiset tulkinnat:

A: Tutkittavalla henkilöllä joko on flunssa (A=1) tai ei (A=0).
B: Tutkittavan henkilön nenä vuotaa (B=1) tai ei (B=0).

Vaihtoehtoisia Bayes-verkkostruktuureja on kaksi: $A \to B$ ja $A \leftarrow B$. Tarkastellaan ensiksi ensimmäistä vaihtoehtoa, ja oletetaan että verkkostruktuuria vastaavat parametrit ovat

$$\begin{aligned} P(B=1 \mid A=1) &= 0.1, \\ P(B=1 \mid A=0) &= 0.2, \\ P(A=1) &= 0.4. \end{aligned}$$

Nämä määrittelevät seuraavan yhteisjakauman:

$$\begin{aligned} P(A=0,B=0) = P(A=0)P(B=0|A=0) = 0.6 \cdot 0.8 = 0.48, \\ P(A=0,B=1) = P(A=0)P(B=1|A=0) = 0.6 \cdot 0.2 = 0.12, \\ P(A=1,B=0) = P(A=1)P(B=0|A=1) = 0.4 \cdot 0.9 = 0.36, \\ P(A=1,B=1) = P(A=1)P(B=1|A=1) = 0.4 \cdot 0.1 = 0.04. \end{aligned}$$

Toisaalta, jos otamme verkkostruktuurivaihtoehdon $A \leftarrow B$, ja asetamme struktuuria vastaavat parametrit seuraavasti:

$$\begin{aligned} P(B=1) &= 0.16, \\ P(A=1 \mid B=1) &= 0.25, \\ P(A=1 \mid B=0) &= 0.36/0.84 \approx 0.43, \end{aligned}$$

voidaan yksinkertaisella laskutoimituksella osoittaa, että syntyvä yhteistodennäköisyysjakauma on täsmälleen sama kuin ensimmäisessä tapauksessa. Esitettyjen verkkostruktuurien sanotaan olevan *ekvivalentteja*, mikä tarkoittaa sitä, että niillä voidaan esittää täsmälleen sama joukko todennäköisyysjakaumia.

Kumpi kahdesta yllä annetusta ekvivalentista verkkostruktuurista on sitten "oikea", kumpaa tulisi käyttää käytännön sovelluksissa? Puhtaasti pragmaattiselta kannalta katsottuna tällä kysymyksella ei ole merkitystä: koska syntyvät mallit esittävät täsmälleen samoja todennäköisyysjakaumia, ne käyttäytyvät täsmälleen samalla tavalla. Bayes-verkkojen soveltamisen kannalta on siis tässä mielessä yhdentekevää, käytetäänkö esimerkissämme verkkostruktuuria $A \rightarrow B$, joka vastannee useimpien käsitystä muuttujien kausaalisesta syy-seuraus-riippuvuussuhteesta, vai verkkostruktuuria $A \leftarrow B$. Bayes-verkkojen kaaret eivät siis välttämättä vastaa kausaalisia syy-seuraussuhteita.

Toisaalta on kuitenkin todettava, että vaikka kaarien kausaalisella tulkinnalla ei ole merkitystä Bayes-verkkojen soveltamisen kannalta, saattaa kausaalisuuden käsitteellä olla suuri käytännön merkitys Bayes-verkkojen rakentamisprosessissa: on helppo nähdä, että muotoa $P(\text{seuraus} \mid \text{syy})$ olevien todennäköisyyksien arviointi on huomattavasti helpompaa kuin muotoa $P(\text{syy} \mid \text{seuraus})$ olevien todennäköisyyksien. Esimerkkitapauksessamme on varmaankin huomattavasti helpompi arvioida mikä on nenän valumisen todennäköisyys, annettuna että henkilöllä on flunssa, kuin arvioida todennäköisyyttä sille, että henkilöllä on flunssa, annettuna että hänen nenänsä valuu. Samoin flunssan esiintymisen prioritodennäköisyyden arvioiminen on helpompaa kuin nenän valumisen prioritodennäköisyyden arvioiminen, koska jälkimmäisessä tapauksessa on arvioitava kaikki mahdolliset tilanteet, joissa nenän valumista voi esiintyä. Jos malli siis konstruoidaan asiantuntijatietämystä käyttäen, on käytännössä yleensä järkevää pyrkiä rakentamaan verkko siten, että solmujen väliset kaaret vastaavat muuttujien kausaalisia riippuvuussuhteita.

Vaikka kysymyksellä Bayes-verkkojen kaarien kausaalisesta tulkinnasta ei ole käytännön merkitystä silloin, kun malleja sovelletaan probabilistiseen päättelyyn, on kysymys sitäkin tärkeämpi ongelmakentän ominaisuuksien tilastollisen analyysin (data mining) kannalta. Bayes-verkkoja, joissa kaikki verkon kaaret kuvaavat kausaalisia riippuvuussuhteita, kutsutaan *kausaaliverkoiksi (causal networks)*. Bayes-verkkotutkimus onkin avannut täysin uusia lähtökohtia kausaalisuuden tutkimukselle sinänsä (katso esim. [109, 135, 48, 53]).

4.3 Probabilistinen päättely Bayes-verkoissa

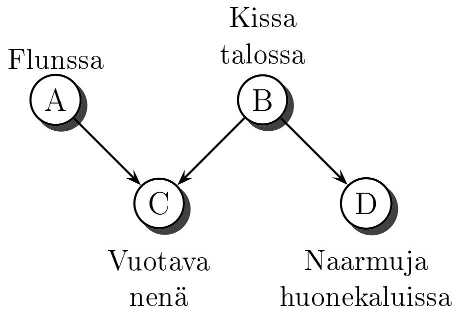
4.3.1 Esimerkkejä

Kuten luvussa 2.3 näimme, on probabilistisessa päättelyssä tavoitteena estimoida muotoa $P(\mathbf{S}_1 \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$ olevia todennäköisyyksiä joko niin, että tuloksena on yhteisjakauma joukon $\mathbf{S}_1$ muodostaville muuttujille, tai niin, että tuloksena saadaan n todennäköisintä arvokombinaatiota $\mathbf{s}_1$. Tarkastelemme seuraavassa esimerkkien valossa sitä, kuinka nämä probabilistisen päättelyn muodot toteutetaan Bayes-verkkojen tapauksessa. Yksinkertaisuuden vuoksi rajoitumme aluksi esimerkissä 9 kuvattuun tapaukseen, jossa joukon $\mathbf{S}_1$ koko on yksi, eli tarkastelemme muotoa $P(X = x \mid \mathbf{S}_2 = \mathbf{s}_2, M, \theta)$ olevia todennäköisyyksiä. Yleistys monimutkaisempiin tapauksiin on suoraviivainen.

Esimerkki 9 Kuvassa 4.5 esitettyyn Bayes-verkkoon liittyvän ongelmakentän kuvaus.

> Olet matkalle vierailulle, kun muistat äkkiä että et muistanut ottaa nenäliinoja mukaan, vaikka kaupungissa on paha influenssaepidemia (arvioit tartunnan todennäköisyyden suuruudeksi 0.05). Olet myös allerginen kissoille, etkä muista varmasti onko isäntäperheellä kissa vai ei. Arvioit todennäköisyydeksi sille, että talossa on kissa, 0.05. Olet huolestunut, koska sekä flunssa että kissaallergia saa nenäsi yleensä vuotamaan, jolloin nenäliinoille olisi ollut tarvetta. Kokemuksesta tiedät, että jos sinulla on joko flunssa, tai olet kosketuksissa kissojen kanssa, nenäsi alkaa vuotaa todennäköisyydellä 0.95, ja arvelet että jos molemmat pitäisivät yhtäaikaa paikkansa, nenän vuotamisen todennäköisyys nousisi arvoon 0.99. Toisaalta, ilman edellä mainittuja syitä on hyvin epätodennäköistä (todennäköisyysarviosi on 0.001), että nenäsi vuotaisi.
>
> Päätät perille tultuasi ensitöikseksi tarkkailla, näkyykö huonekaluissa kissan tekemiä naarmuja, joita kokemuksesi syntyy todennäköisyydellä 0.99, mikäli talossa asuu kissa. On tietysti myös mahdollista, että huonekaluista löytyy kissan tekemiä naarmuja, vaikka talossa ei kissaa olisikaan (huonekaluthan voivat olla vaikkapa ostettu käytettyinä), mutta arvioit tämän mahdollisuuden todennäköisyydeksi ainoastaan 0.1.

Esimerkissä 9 kuvattua ongelmakenttää voidaan mallintaa kuvassa 4.5 esitetyllä Bayes-verkolla, kun formalisoimme esimerkin kuvauksessa annetut



$$
\begin{aligned}
&P(A=1)=0.05\\
&P(B=1)=0.05\\
&P(C=1\mid A=0,B=0)=0.001\\
&P(C=1\mid A=1,B=0)=0.95\\
&P(C=1\mid A=0,B=1)=0.95\\
&P(C=1\mid A=1,B=1)=0.99\\
&P(D=1\mid B=1)=0.99\\
&P(D=1\mid B=0)=0.1
\end{aligned}
$$

Kuva 4.5: Esimerkin 9 Bayes-verkkokuvaus.

oletukset seuraavasti: nenän vuotamiseen (muuttuja C) vaikuttavat muuttujat ovat A (onko flunssa vai ei) ja B (onko talossa kissa vai ei). Huonekaluista löytyvien raapimisjälkiin (muuttuja D) vaikuttaa se, onko talossa kissa, eivät muut mallissa käytetyt muuttujat.

Mallinnetussa maailmassa on kaikkiaan 16 mahdollista tilaa, joita vastaavat yhteistodennäköisyydet saadaan laskettua esitetyn Bayes-verkkomallin (kuva 4.5) avulla soveltamalla kaavaa (4.3), josta näemme että

$$P(A,B,C,D) = P(A)P(B)P(C \mid A,B)P(D \mid B).$$

Vastaavat 16 yhteistodennäköisyyttä on esitetty taulukossa 4.2.

Taulukossa 4.2 esitettyjä yhteistodennäköisyyksiä hyväksikäyttäen voidaan laskea "oletusarvoinen" (priori)todennäköisyys sille, että nenä alkaa vuotaa esimerkissä 9 esitetyssä mallissa:

$$
\begin{aligned}
P(C=1) &= \sum_{a\in\{0,1\}}\sum_{b\in\{0,1\}}\sum_{d\in\{0,1\}} P(C=1,A=a,B=b,D=d)\\
&= 2.450\text{E-}03 + 2.475\text{E-}05 + 4.513\text{E-}03 + 4.061\text{E-}02\\
&\quad +4.467\text{E-}02 + 4.513\text{E-}04 + 9.025\text{E-}05 + 8.123\text{E-}04 = 9.36\%.
\end{aligned}
$$

Vastaavasti saamme naarmujen esiintymisen (priori)todennäköisyyden $P(D=1)$ arvoksi 14.45%. Flunssan (priori)todennäköisyys $P(A=1)$, samoin kuin kissan esiintymisen (priori)todennäköisyys $P(B=1)$ on 5%, mikä nähdään suoraan käytetystä Bayes-verkkomallista.

Sana "priori" viittaa edellä todennäköisyyteen, joka saadaan Bayes-verkkomallista joko suoraan tai yksinkertaisen laskutoimituksen kautta, ilman että tehdyt havainnot vaikuttavat tilanteeseen. Probabilististen mallien käyttökelpoisuus perustuu suoraviivaiseen kalkyyliin, jonka avulla eri satunnaismuuttujien todennäköisyydet muuttuvat dynaamisesti sitä mukaa, kun saamme ongelmakentästä uusia havaintoja. Kun esimerkiksi havaitset, että nenäsi alkaa vuotaa ($C=1$), voit bayesiläisen kalkyylin avulla *päivittää* uskomuksesi

A	B	C	D	P(A)	P(B)	P(C|A,B)	P(D|B)	P(A,B,C,D)
1	1	1	1	0.05	0.05	0.99	0.99	2.450E-03
1	1	1	0	0.05	0.05	0.99	0.01	2.475E-05
1	1	0	1	0.05	0.05	0.01	0.99	2.475E-05
1	1	0	0	0.05	0.05	0.01	0.01	2.500E-07
1	0	1	1	0.05	0.95	0.95	0.1	4.513E-03
1	0	1	0	0.05	0.95	0.95	0.9	4.061E-02
1	0	0	1	0.05	0.95	0.05	0.1	2.375E-04
1	0	0	0	0.05	0.95	0.05	0.9	2.138E-03
0	1	1	1	0.95	0.05	0.95	0.99	4.467E-02
0	1	1	0	0.95	0.05	0.95	0.01	4.513E-04
0	1	0	1	0.95	0.05	0.05	0.99	2.351E-03
0	1	0	0	0.95	0.05	0.05	0.01	2.375E-05
0	0	1	1	0.95	0.95	0.001	0.1	9.025E-05
0	0	1	0	0.95	0.95	0.001	0.9	8.123E-04
0	0	0	1	0.95	0.95	0.999	0.1	9.016E-02
0	0	0	0	0.95	0.95	0.999	0.9	8.114E-01
								1.000E+00

Taulukko 4.2: Kuvassa 4.5 esitetyn Bayes-verkon määrittelemä yhteistodennäköisyysjakauma.

siihen, että sinulla on flunssa, laskemalla flunssan ehdollinen todennäköisyys *annettuna tehdyt havainnot*:

$$\begin{aligned} P(A=1 \mid C=1) &= \frac{P(A=1, C=1)}{P(C=1)} \\ &= \frac{\sum_{b\in\{0,1\}} \sum_{d\in\{0,1\}} P(A=1, B=b, C=1, D=d)}{P(C=1)} \\ &= \frac{2.450\text{E-}03 + 2.475\text{E-}05 + 4.513\text{E-}03 + 4.061\text{E-}02}{9.36\text{E-}02} \\ &= 50.84\%. \end{aligned}$$

Flunssan todennäköisyys siis yli kymmenkertaistui tehdyn havainnon seurauksena. Vastaavasti saamme todennäköisyydeksi sille, että talossa on kissa, annettuna havainto nenän vuotamisesta,

$$P(B=1 \mid C=1) = 50.84\%,$$

ja todennäköisyydeksi sille, että huonekaluissa on naarmuja

$$P(D=1 \mid C=1) = 55.25\%.$$

Selitysten poissulkeminen (explaining away)
Annettuna nenän vuotamisesta tehty havainto, saamme sekä flunssan että kissan esiintymisen todennäköisyydeksi 50.84%, siis vain hieman yli 0.5. Ensi näkemältä voi tuntua siltä, että tehty havainto ei ole merkittävästi lisännyt tietoamme ongelmakentästä, mutta on muistettava että sekä flunssan että kissan esiintymisen prioritodennäköisyys oli mallissamme hyvin pieni, vain 5%, joten näiden muuttujien todennäköisyys on yli kymmenkertaistunut. Toisaalta näemme, että tehty havainto ei tässä vaiheessa auta meitä päättämään, kumpi seikka, flunssa vai kissa-allergia, on nenän valumisen syy. Jos havaitsemme edelleen, että talon huonekaluissa on kissan tekemiä naarmuja, satunnaismuuttujien posterioritodennäköisyydet päivittyvät jälleen:

$$\begin{aligned} P(B=1 \mid C=1, D=1) &\approx 91\%, \\ P(A=1 \mid C=1, D=1) &\approx 13\%. \end{aligned}$$

Tässä uudessa tilanteessa voimme olettaa, että nenän vuotamisen aiheuttaja on melko varmasti kissa-allergia. Tämä esimerkki havainnollistaa todennäköisyysmalleissa esiintyvää havaintojen selityksiin liittyvää *"poisselittämis" (explaining away)*-ilmiötä: yhden vaihtoehtoisen selityksen todennäköisyyden kasvaminen vähentää automaattisesti vaihtoehtoisten selitysten uskottavuutta, vaikka selitykset eivät sinänsä olisikaan poissulkevia. Tämän intuitiivisesti luontevalta tuntuvan periaatteen on havaittu toteutuvan myös inhimillisessä päättelyssä [57, 107, 89].

Edellä näimme, että annettuna havainto 'nenä vuotaa', saamme sekä flunssan että kissan esiintymisen todennäköisyydeksi 0.5084, ja naarmujen esiintymisen todennäköisyydeksi 0.5525. Jos kysymme nyt, mikä on todennäköisin mallimaailman tila, annettuna havainto $C = 1$, intuitiivisesti houkuttelevalta vaihtoehdolta tuntuisi asettaa kukin binäärimuuttuja itsenäisesti siihen arvoon, jonka todennäköisyys on suurempi, jolloin vastaus olisi

$$P(A = 1, B = 1, C = 1, D = 1) = 2.450\text{E-}03.$$

Taulukosta 4.2 on kuitenkin helppo nähdä, että muuttujien konfiguraatioavaruudesta löytyy kolme seikan $C = 1$ kanssa ristiriidatonta tilaa, joiden todennäköisyys on suurempi kuin yllä olevalla konfiguraatiolla:

$$\begin{aligned} P(A=0,B=1,C=1,D=1) &= 4.467\text{E-}02, \\ P(A=1,B=0,C=1,D=0) &= 4.061\text{E-}02, \text{ ja} \\ P(A=1,B=0,C=1,D=1) &= 4.513\text{E-}03. \end{aligned}$$

Kokonaistodennäköisuuden maksimoivan tilan etsiminen on siis aivan eri ongelma kuin kunkin yksittäisen satunnaismuuttujan ehdollisen todennäköisyysjakauman laskeminen.

4.3.2 Päättelyalgoritmit

Kuten edellä esitetystä esimerkistä näimme, mikä tahansa Bayes-verkkoesitystä vastaavan yhteistodennäköisyysjakauman arvo saadaan laskettua kertomalla asiaankuuluvat Bayes-verkon parametrit keskenään. Näin ollen probabilistisen päättelyn suorittamiseksi ei tarvitse tallettaa taulukon 4.2 kaltaista todennäköisyystaulukkoa, vaan riittää tallettaa Bayes-verkon parametrit, joita on yleensä pieni määrä koko malliavaruuden kokoon verrattuna, kuten edellisessä luvussa näimme. Toisaalta probabilistisessa päättelyssä käytettävien ehdollisten todennäköisyyksien laskeminen edellyttää yllä esitetyssä muodossaan monia tekijöitä sisältävien marginalisointisummien laskemista, mikä saattaa suurten Bayes-verkkojen tapauksessa olla käytännössä mahdotonta. Bayes-verkkojen määritelmään liittyvät ehdolliset riippumattomuusoletukset tarjoavat kuitenkin ratkaisun myös tähän ongelmaan, minkä ansiosta Bayes-verkoille voidaan kehittää tehokkaita päättelyalgoritmeja. Tämän seikan havainnollistamiseksi tarkastelkaamme yllä esitettyä esimerkkiä vielä kerran.

Olettakaamme, että tehtävänä on laskea kuvassa 4.5 esitettyyn Bayes-verkkomalliin liittyen ehdollinen todennäköisyys $P(B=1 \mid C=1, D=1)$. Todennäköisyyslaskennan perusaksioomia käyttäen näemme helposti, että

$$P(B=1 \mid C=1, D=1) = \frac{P(d^+ \mid b^+, c^+)P(c^+ \mid b^+)P(b^+)}{P(c^+, d^+)},$$

missä arvoasetuksesta $X=1$ on käytetty lyhennettä x^+ (asetuksesta $X=0$ käytetään jatkossa lyhennettä x^-). Bayes-verkkoihin liittyvän ehdollisen riippumattomuuden määritelmän perusteella muuttuja D on ehdollisesti riippumaton muuttujasta C, annettuna muuttujan B arvo, joten saamme

$$P(b^+ \mid c^+, d^+) = \frac{P(d^+ \mid b^+)P(c^+ \mid b^+)P(b^+)}{P(c^+, d^+)}.$$

Toisaalta, koska

$$P(c^+, d^+) = P(d^+ \mid c^+, b^+)P(c^+ \mid b^+)P(b^+) + P(d^+ \mid c^+, b^-)P(c^+ \mid b^-)P(b^-),$$

saamme, että

$$P(b^+ \mid c^+, d^+) = \frac{P(d^+ \mid b^+)P(c^+ \mid b^+)P(b^+)}{P(d^+ \mid b^+)P(c^+ \mid b^+)P(b^+) + P(d^+ \mid b^-)P(c^+ \mid b^-)P(b^-)}.$$

Haluamamme ehdollisen todennäköisyyden saamiseksi riittää siis laskea kaksi tuloa, $P(d^+|b^+)P(c^+|b^+)P(b^+)$ ja $P(d^+|b^-)P(c^+|b^-)P(b^-)$, jotka molemmat sisältävät vain Bayes-verkkomalliin talletettuja parametreja.

Bayes-verkkomallien avulla tapahtuvan probabilistisen päättelyn käytännön toteuttamista varten voimme ajatella kuvan 4.5 Bayes-verkkoa tietokoneohjelma-arkkitehtuurina, jossa kukin muuttuja X vastaa moduulia, jonka tehtävänä on laskea ehdollinen todennäköisyysjakauma $P(X \mid \mathbf{S}_2 = \mathbf{s}_2)$. Oletetaan lisäksi, että Bayes-verkon parametrit on talletettu siten, että kukin moduuli X näkee muotoa $P(X \mid F(X))$ olevat parametrit. Kuten edellä näimme, yllä esitetyssä tapauksessa moduuli B tarvitsee seuraavat luvut voidakseen määrätä ehdollisen todennäköisyysjakauman $P(b^+ \mid c^+, d^+)$:

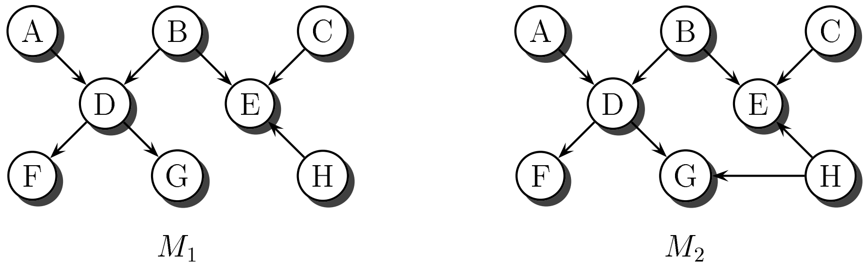
$$P(d^+ \mid b^+), P(c^+ \mid b^+), P(b^+), P(d^+ \mid b^-), P(c^+ \mid b^-), \text{ ja } P(b^-).$$

Luvut $P(b^+)$ ja $P(b^-)$ on talletettu paikallisesti moduuliin B Bayes-verkkomallin parametreina. Koska

$$P(c^+ \mid b^+) = P(c^+ \mid a^-, b^+)P(a^-) + P(c^+ \mid a^+, b^+)P(a^+),$$

moduuli C voi laskea tämän tarvittavan luvun, samoin kuin luvun $P(c^+|b^-)$, nopeasti ja tehokkaasti käyttäen paikallisesti talletettua informaatiota (moduuliin C talletettuja Bayes-verkon parametreja), sekä solmun välittömään edeltäjään A talletettuja parametreja $P(a^+)$ ja $P(a^-)$. Moduuli D voi vastaavasti laskea arvot $P(d^+|b^+)$ ja $P(d^+|b^-)$. Laskeakseen todennäköisyyden $P(b^+|c^+, d^+)$ moduulin B riittää siis lähettää tarvittavien lukujen laskupyyntö moduleille C ja D (jonka täyttämiseksi moduuli C lähettää parametrien välityspyynnön moduulille A), ja kertoa saadut vastaukset edellä esitetyllä tavalla halutun ehdollisen todennäköisyysjakauman tuottamiseksi.

Edellä kuvattua viestinvälitysarkkitehtuuriin perustuvaa menetelmää ehdollisten todennäköisyyksien laskemiseksi voidaan soveltaa kaikissa *yksipolkuisissa (singly connected)* Bayes-verkoissa: yksipolkuisissa verkoissa ei minkään kahden solmun välillä ole kuin yksi suuntaamaton polku, missä suuntaamaton polku tarkoittaa verkon kaaria seuraavaa reittiä, joka voi kulkea



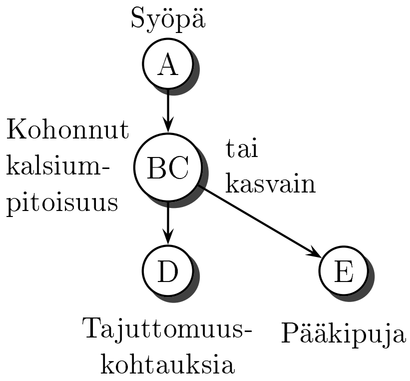
Kuva 4.6: Yksipolkuinen (M_1) ja monipolkuinen (M_2) Bayes-verkkostruktuuri.

suunnattuja kaaria kumpaan suuntaan tahansa. Kuvassa 4.6 yksipolkuinen Bayes-verkkostruktuuri M_1 muutetaan *monipolkuiseksi (multi-connected)* yhdistämällä solmut H ja G, jolloin syntyy kaksi (suuntaamatonta) polkua esimerkiksi solmujen B ja G välille.

Yksipolkuisten Bayes-verkkojen tapauksessa voidaan probabilistisen päättelyn päämääränä olevien ehdollisten todennäköisyyksien laskeminen suorittaa erittäin tehokkaasti, lineaarisessa ajassa Bayes-verkon parametrien lukumäärään nähden. Algoritmin yksityiskohtainen kuvaus löytyy esimerkiksi alan perusoppikirjoista [108, 100]. Maksimitodennäköisyyskonfiguraation etsiminen voidaan suorittaa samankaltaisella algoritmilla.

Viestinvälitystekniikkaa käyttävien tehokkaiden probabilististen päättelyalgoritmien toiminta perustuu siihen tosiseikkaan, että yksipolkuisissa verkoissa kukin solmu osittaa verkon kahteen erilliseen osaan, jotka voidaan käsitellä erikseen, ja sama pätee rekursiivisesti jokaiselle verkon solmulle. Monipolkuisissa verkoissa menetelmä ei valitettavasti toimi. Itse asiassa voidaan teoreettisesti osoittaa, että muotoa $P(X = 1 \mid \mathbf{S}_2 = \mathbf{s}_2)$ olevien ehdollisten todennäköisyyksien laskeminen [27], tai edes niiden approksimointi [29, 120], ja myös maksimitodennäköisyystilojen etsiminen [129], on erittäin vaikea ongelma monipolkuisten verkkojen tapauksessa: jos näille ongelmille olisi aina polynomisessa ajassa toimiva ratkaisumenetelmä, menetelmää voitaisiin käyttää ratkaisemaan kaikki ns. NP-täydelliset päätösongelmat polynomisessa ajassa, minkä ei yleisesti uskota olevan mahdollista. On kuitenkin syytä muistaa näiden teoreettisten tulosten käsittelevän pahimman tapauksen (worst-case) tilannetta: yhdenkin eksponentiaalisen ajan vaativan tapauksen löytäminen riittää teoreettisen tuloksen johtamiseksi, vaikka kaikki muut tapaukset voitaisiin käsitellä polynomisessa ajassa.

Yleisin lähestymistapa probabilistisen päättelyn toteuttamiseksi monipolkuisissa Bayes-verkoissa on yrittää muuttaa annettu monipolkuinen verkko yksipolkuiseksi, jolloin päättely voidaan toteuttaa tehokkaasti, kuten edel-



$P(A=1)$	$=$	0.20
$P(BC=00 \mid A=1)$	$=$	$0.20 \cdot 0.80 = 0.16$
$P(BC=01 \mid A=1)$	$=$	$0.20 \cdot 0.20 = 0.04$
$P(BC=10 \mid A=1)$	$=$	$0.80 \cdot 0.80 = 0.64$
$P(BC=11 \mid A=1)$	$=$	$0.80 \cdot 0.20 = 0.16$
$P(BC=00 \mid A=0)$	$=$	$0.80 \cdot 0.95 = 0.76$
$P(BC=01 \mid A=0)$	$=$	$0.80 \cdot 0.05 = 0.40$
$P(BC=10 \mid A=0)$	$=$	$0.20 \cdot 0.95 = 0.19$
$P(BC=11 \mid A=0)$	$=$	$0.20 \cdot 0.05 = 0.01$
$P(D=1 \mid BC=11)$	$=$	0.80
$P(D=1 \mid BC=01)$	$=$	0.80
$P(D=1 \mid BC=10)$	$=$	0.80
$P(D=1 \mid BC=00)$	$=$	0.05
$P(E=1 \mid BC=01)$	$=$	0.80
$P(E=1 \mid BC=11)$	$=$	0.80
$P(E=1 \mid BC=00)$	$=$	0.60
$P(E=1 \mid BC=10)$	$=$	0.60

Kuva 4.7: Kuvan 4.3 Bayes-verkko muutettuna yksipolkuiseksi verkoksi.

lä näimme. Yleisimmin käytetty transformointitekniikka on muuttujien *ryhmittely (clustering)*. Tarkastellaan esimerkiksi kuvassa 4.3 esitettyä Bayes-verkkoa. Ryhmittelemällä muuttujat uudelleen siten, että ne muodostavat neljä ryhmää $\{A\}$, $\{B, C\}$, $\{D\}$ ja $\{E\}$ näemme, että jos sulautamme muuttujat B ja C uudeksi muuttujaksi BC, syntyvä Bayes-verkko on yksipolkuinen (katso kuva 4.7). Koska muuttujilla B ja C oli kummallakin 2 mahdollista arvoa, on yhdistetyllä muuttujalla BC 4 mahdollista arvoa; merkitään näitä arvoja symboleilla 00, 01, 10 ja 11. Syntyneeseen uuteen verkkoon liittyvät todennäköisyydet on helppo laskea alkuperäisen verkon parametriarvoista: muotoa $P(D \mid BC)$ olevat todennäköisyydet vastaavat suoraan muotoa $P(D \mid B, C)$ olevia todennäköisyyksiä, ja ehdollisen riippumattomuuden määritelmän mukaan näemme, että $P(E \mid BC) = P(E \mid C)$ ja $P(BC \mid A) = P(B \mid A)P(C \mid A)$.

Edellä kuvatussa yksinkertaisessa tapauksessa oli helppo nähdä, kuinka annettu monipolkuinen verkko voidaan muuttaa yksipolkuiseksi kaksi muuttujaa yhdistämällä, mutta monimutkaisemmissa verkoissa tehtävä on huomattavasti vaikeampi, eivätkä ratkaisut ole aina yhtä yksinkertaisia. Käytännön kannalta tärkeä kysymys on, voidaanko tämä transformaatioprosessi automatisoida — onko mahdollista kehittää algoritmi, joka tuottaa mistä tahansa annetusta monipolkuisesta Bayes-verkosta yksipolkuisen Bayes-verkon polynomisessa ajassa? Vastaus on hieman yllättäen kyllä: tällaisia algoritmeja on kehitetty itse asiassa lukuisia (katso esim. [108, 100, 62, 18]). Useimmat esitetyistä menetelmistä perustuvat kaksivaiheiseen algoritmiin, jossa ensimmäisessä vaiheessa annettu Bayes-verkko *moralisoidaan*[2] yhdistämällä sol-

[2]Moralisointi-termi (moralization) johtuu siitä että prosessissa yhdistetään solmut (vanhemmat) joilla on yhteinen seuraaja (lapsi).

mut, joilla on yhteinen seuraaja, ja poistamalla kaarien suunnat. Menetelmän toisessa vaiheessa syntynyt suuntaamaton verkko *kolmioidaan (triangulation)* lisäämällä verkkoon suuntaamattomia kaaria kunnes kaikissa verkossa sijaitsevien yli kolmen kaaren pituisissa silmukoissa on joitakin kahta silmukan solmua yhdistävä kaari. Jos kolmioidussa verkossa esiintyvien maksimaalisten *klikkien* (suurimpien sellaisten verkosta löytyvien solmujoukkojen, joissa kaikki solmut on kytketty toisiinsa) solmut yhdistetään yhdeksi muuttujaksi, niin syntyvien yhdistettyjen muuttujien avulla on mahdollista luoda alkuperäisen monipolkuisen Bayes-verkon kanssa ekvivalentti yksipolkuinen verkko. Syntyvää yksipolkuista verkkoa kutsutaan *liittymäpuuksi (junction tree)*. Nimitys johtuu siitä, että liittymäpuun solmusta voi olla kaari toiseen solmuun vain silloin, jos solmuja vastaavien klikkien leikkaus ("liittymä") ei ole tyhjä. Liittymäpuiden konstruoimisessa käytettävä algoritmi on melko monimutkainen, eikä sen esittäminen tämän raportin yhteydessä ole mielekästä. Algoritmin yksityiskohdat selviävät mm. lähteistä [108, 100, 62, 18].

Liittymäpuihin perustuvat menetelmät tarjoavat siis mahdollisuuden muuntaa annettu monipolkuinen Bayes-verkko puurakenteiseksi Bayes-verkoksi siten, että syntyvä liittymäpuu esittää samaa todennäköisyysjakaumaa kuin alkuperäinen monipolkuinen Bayes-verkko. Koska syntyvä liittymäpuu on puurakenteinen, siis yksipolkuinen Bayes-verkko, voidaan probabilistisen päättelyn suorittamisessa käyttää yksipolkuisille verkoille kehitettyjä tehokkaita viestinvälitysalgoritmeja, jolloin päättely voidaan suorittaa lineaarisessa ajassa syntyneen verkon parametrien lukumäärään nähden. Tässä yhteydessä on kuitenkin tärkeää huomata, että syntyvän liittymäpuun parametrien lukumäärä ei ole sama kuin alkuperäisen monipolkuisen Bayes-verkon parametrien lukumäärä, vaan se voi olla paljon suurempi, pahimmassa tapauksessa eksponentiaalisen suuri alkuperäiseen lukumäärään verrattuna. Voidaan itse asiassa osoittaa, minimaalisen määrän liittymäpuuparametreja tuottavan kolmioinnin löytäminen on NP-täydellinen ongelma [147]. Alkuperäisen ongelman vaikeutta ei siis voi paeta tiedon esitystapaa muuttamalla (yhdistämällä muuttujia), mikä on tietenkin muutenkin itsestään selvää. Becker ja Geiger [7] ovat esittäneet "hajota-ja-hallitse"-periaatteeseen perustuvan algoritmin, jonka tuottama kolmiointi on aina vähintään tietyn vakioetäisyyden päässä optimaalisesta. Kjærulff [68] on tutkinut stokastisten optimointimenetelmien soveltamista tämän ongelman ratkaisemisessa.

Käytännön sovelluksista saadut kokemukset osoittavat, että liittymäpuihin perustuvat algoritmit toimivat yleensä varsin hyvin — useimmat luvussa 5.2 kuvatuista ohjelmistoista perustuvat tähän lähestymistapaan. Pearlin *ehdollistamismenetelmä (conditioning)* [107] on läheistä sukua solmujen yhdistämiseen perustuville tekniikoille: ehdollistamisessa monipolkuista verkkoa ei kuitenkaan eksplisiittisesti muuteta yksipolkuiseksi verkoksi, vaan ongelmia aiheuttavat polut poistetaan implisiittisesti kiinnittämällä (ehdollistamalla) polulla sijaitsevia muuttujia johonkin arvoon. Probabilistisen päättelyn toteuttamisessa voidaan myös luopua laskennan deterministisyydestä, ja käyttää stokastisia päättelyalgoritmeja. Teoksessa [96] kuvataan, kuinka tämän ongelman ratkaisemisessa voidaan käyttää Boltzmannin koneena tunnettua stokastista neuroverkkoarkkitehtuuria. Muita vaihtoehtoisia lähestymistapoja probabilistisen päättelyn toteuttamiseksi monipolkuisissa Bayes-verkoissa on kuvattu lähteissä [18, 58, 31, 121, 87, 148].

4.4 Bayes-verkkojen rakentaminen

Kuten luvussa 4.1 näimme, Bayes-verkkomallien komponenteille voidaan antaa intuitiivisesti selkeä semanttinen tulkinta: Bayes-verkkostruktuuri kuvaa ongelmakentän mallintamisessa käytettyjen attribuuttien välisiä riippuvuuksia, Bayes-verkon parametreina käytetyt ehdolliset todennäköisyydet taas riippuvuuksien voimakkuuksia. Tämäntyyppinen tieto on usein helposti saatavilla ongelmakentän asiantuntijoilta, jolloin Bayes-verkkojen konstruointi on helppoa. Bayes-verkkokehittimiä (katso luku 5.2) käyttäen Bayes-verkkojen soveltaminen ei edellytä syvempää Bayes-verkkoteorian ymmärtämistä, jolloin asiantuntijat voivat myös konstruoida Bayes-verkkomalleja suoraan itse, ja siirtyä samalla testaamaan kehitettyjen mallien toimivuutta. Eric Horvitzin mukaan[3] United Airlinesin insinöörit testasivat ensimmäisiä itse kehittämiään Bayes-verkkomalleja jo muutaman tunnin kuluttua siitä, kun Bayes-verkkomallien käsite ja niiden konstruoimiseen soveltuvat kehittimet oli heille esitelty.

Toisaalta, kuten luvussa 2.2 todettiin, bayesiläisen mallinnuksen yksi suurimpia vahvuuksia on se, että lähestymistapa mahdollistaa asiantuntijatietämyksen yhdistämisen tilastolliseen oppimiseen esimerkkiaineistosta. Mallien oppiminen jaettiin edellä kahteen vaiheeseen: mallistruktuurin oppiminen ja malliparametrien oppiminen. Kuvaamme seuraavassa kuinka laskea luvussa 2.2.2 esitetty yleinen oppimiskriteeri eri Bayes-verkkostruktuureille, ja kuinka edelleen kiinnittää Bayes-verkon parametrit, annettuna struktuuri. Samassa yhteydessä selvitämme kuinka asiantuntijatietämys voidaan priorijakaumia käyttäen ottaa huomioon oppimisprosessissa.

[3] Katso URL: http://www.auai.org/BN-Testimonial.html

4.4.1 Verkkostruktuurin oppiminen

Luvussa 2.2.2 näimme, että bayesiläisessä oppimisessa mallistruktuurit evaluoidaan vertailemalla mallien posterioritodennäköisyyksiä,

$$P(M \mid \mathcal{D}) \propto P(\mathcal{D} \mid M)P(M),$$

ja luvussa 2.2.4 näimme, kuinka asiantuntijatietämys voidaan integroida oppimisprosessiin priorijakaumien kautta. Mallistruktuurin oppimisen päämääränä on siis iteratiivisen etsintäalgoritmin kautta etsiä se mallistruktuuri M, joka maksimoi tämän todennäköisyyden. Bayes-verkkojen tapauksessa etsintäavaruus on valtava: erilaisten syklittömien suuntaamattomien verkkojen lukumäärä saadaan lähteessä [119] esitetystä rekursioyhtälöstä, josta näemme että jo kymmenellä solmulla saadaan aikaan noin 4.2×10^{18} erilaista Bayes-verkkostruktuuria! Tässä valossa ei olekaan lainkaan yllättävää, että Bayes-verkkojen oppimisongelman on osoitettu olevan NP-kova ongelma [25]. On kuitenkin jälleen kerran korostettava, että kyseessä on pahimman tapauksen analyysi, joka ei kerro mitään ongelman keskimääräisestä vaikeudesta. Bayes-verkkojen oppimisessa onkin sovellettu hyvällä menestyksellä tunnettuja heuristisia etsintäalgoritmeja — viitteitä käytettyihin algoritmeihin löytyy esimerkiksi teoksista [49, 127].

Jos ongelmakentästä ei ole saatavilla priori-informaatiota, on järkevää olettaa eri struktuureille tasainen priorijakauma. Todennäköisyyttä $P(M)$ voidaan tällöin pitää vakiona, ja käyttää oppimiskriteerinä siis yksinomaan kokonaisuskottavuutta $P(\mathcal{D} \mid M)$. Mikäli priori-informaatiota on saatavilla, voidaan tätä tietämystä käyttää hyväksi priorijakauman $P(M)$ määräämisessä. Koska mahdollisten Bayes-verkkojen määrä on kuitenkin tavattoman suuri, ei ole käytännössä mahdollista määrätä prioritodennäköisyyttä $P(M)$ erikseen kullekin Bayes-verkkostruktuurille. Priori-informaatiota voidaan kuitenkin käyttää hyväksi yhden (tai muutaman) ns. prioriverkon konstruoimisessa. Prioriverkoilla tarkoitetaan verkkostruktuureja, joiden prioritodennäköisyden arvellaan olevan hyvin suuri (verrattuna muihin verkkostruktuureihin). Prioriverkkoja voidaan käyttää hyödyksi mallistruktuurin etsimisessä esimerkiksi heurististen etsintäalgoritmien aloituspisteinä, tai varsinaisen priorijakauman estimoinnissa lähteessä [51] esitetyllä tavalla.

Bayes-verkkojen kokonaisuskottavuuden $P(\mathcal{D} \mid M)$ laskemiseksi on tehtävä joitakin teknisiä oletuksia, joista tärkein koskee parametrien priorijakaumaa $P(\theta \mid M)$. Kuten luvussa 4.1 näimme, Bayes-verkkostruktuuriin liitettävät parametrit θ ovat muotoa $P(X \mid F(X))$ olevia ehdollisia todennäköisyyksiä, missä $F(X)$ on muuttujan X edeltäjien joukko. Numeroidaan seuraavassa joukon $F(X)$ kaikki mahdolliset arvokombinaatiot, ja oletetaan että niitä on q_i kappaletta. Merkinnällä θ_{ijk} tarkoitamme jatkossa parametria $P(X_i = k \mid F(X_i) = j)$. Jos oletamme nyt, että priorijakauma $P(\theta_{ij} \mid M)$ on ns. *Dirichlet-jakauma* (ks. esim. [40]) *hyperparametrein*[4] $N'_{ij1}, \ldots, N'_{ijr_i}$, niin voidaan osoittaa [28, 51] että kokonaisuskottavuus voidaan laskea kaavalla (4.4).

Kokonaisuskottavuuden laskeminen:

$$P(\mathcal{D} \mid M) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}, \tag{4.4}$$

missä

- Γ on ns. Gamma-funktio, tavallisen kertomafunktion yleistys myös muille kuin kokonaisluvuille (minkä ansiosta hyperparametrien ei tarvitse välttämättä olla kokonaislukuarvoisia),
- n on muuttujien lukumäärä,
- q_i on solmun X_i edeltäjien mahdollisten arvokombinaatioiden lukumäärä.
- r_i on muuttujan X_i arvojen lukumäärä,
- N_{ijk} on niiden tapausten lukumäärä opetusjoukossa $\mathcal{D}$, joissa muuttujan X_i arvo on k, ja muuttujan X_i edeltäjät $F(X_i)$ ovat arvokombinaatiossa j,
- N'_{ijk} on parametria θ_{ijk} vastaavan priorijakauman määräämisessä käytetty hyperparametri,
- $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, ja
- $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$.

Kaavassa (4.4) esiintyvät tilastolliset tunnusluvut N_{ijk} on helppo laskea opetusjoukosta $\mathcal{D}$, ja kokonaisuskottavuus voidaan laskea ajassa $O(Nn^2r)$, missä N on joukon $\mathcal{D}$ koko, n muuttujien lukumäärä, ja r suurin mahdollinen muuttujien arvojen lukumäärä. Käyttäjän tehtäväksi jää määrätä hyperparametrit N'_{ijk}. Mikäli käyttäjällä ei ole ongelmakenttää koskevaa prioritietämystä, on tällöin luontevaa käyttää tasaista priorijakauma, joka saadaan

[4]Nimitys 'hyperparametri' johtuu siitä, että kyseessä on "toisen kertaluvun" parametri: muuttujien yhteistodennäköisyysjakauman määrittelemisessä käytettävän parametrin jakauman määrittelemisessä käytetty parametri.

asettamalla $N'_{ijk} = 1$ kaikilla i:n j:n ja k:n arvoilla. Jos prioritietämystä on saatavilla, antaa kaava (4.4) vihjeen siitä kuinka asiantuntijatietämys voidaan koodata hyperparametreiksi N'_{ijk}: hyperparametrit ja tilastolliset tunnusluvut N_{ijk} ovat käytettävässä kaavassa täysin identtisessä asemassa, joten on helppo päätellä, että jos käyttäjä asettaa hyperparametrin N'_{ijk} arvoksi esimerkiksi viisi, on tämän asetuksen vaikutus kokonaisuskottavuuteen sama kuin jos käyttäjä olisi lisännyt opetusjoukkoon $\mathcal{D}$ viisi esimerkkiä, joissa kussakin muuttujan X_i arvo olisi k, ja muuttujan edeltäjät $F(X_i)$ olisivat arvokombinaatiossa j. Käyttäjä voi siis muotoilla henkilökohtaisen näkemyksensä muuttujan X_i arvojen jakaumasta muuttamalla hyperparametreja N'_{ijk}, ja hyperparametrien määräämisessä käytetyn arvoalueen skaalaa muuttamalla saadaan vaihdeltua prioritietämyksen vaikutuksen suhteellista voimakkuutta opetusjoukkoon vaikutukseen verrattuna. Tätä lähestymistapaa kutsutaan *ekvivalentin otoskoon (equivalent sample size)* menetelmäksi: asiantuntija voi painottaa tietämyksensä merkitystä ajattelemalla tietämystään "prioriopetusjoukkona", joukkona opetusjoukon ulkopuolisia tapauksia joista asiantuntija on oman kokemusperäisen tietämyksensä muodostanut.

Kaavassa (4.4) esitetty menetelmä kokonaisuskottavuuden laskemiseksi pätee diskreeteille muuttujille, kun muuttujien arvot otosjoukossa $\mathcal{D}$ oletetaan jakautuneiksi multinomijakauman mukaan. Vastaavanlainen tulos voidaan kuitenkin itse asiassa johtaa hyvin laajalle joukolle erilaisia jakaumia (ns. eksponentiaaliselle perheelle), myös jatkuville jakaumille, ja jatkuvien ja diskreettien muuttujien muodostamille sekajakaumille [10]. Esimerkki Bayes-verkon kokonaisuskottavuuden laskemisesta jatkuvien jakaumien tapauksessa löytyy lähteestä [41].

Yllä kuvattu lähestymistapa perustuu Bayes-verkkostruktuurien vertailuun niiden posterioritodennäköisyyden (tai käytännössä yleensä kokonaisuskottavuuden) perusteella. Luvussa 4.1 esitetty Bayes-verkkojen määritelmä tarjoaa myös vaihtoehtoisen lähestymistavan tähän ongelmaan: koska kukin Bayes-verkkostruktuuri kuvaa tiettyä joukkoa riippumattomuusoletuksia, voidaan opetusjoukosta ensin yrittää tilastollisesti päätellä ongelmakentässä vallitsevat riippumattomuudet, ja konstruoida sitten pääteltyjä riippumattomuuksia vastaava Bayes-verkkostruktuuri. Tähän lähestymistapaan perustuvia algoritmeja on esitetty esimerkiksi lähteessä [135]. Täydellisten riippumattomuuksien sijaan voidaan yrittää myös tilastollisesti päätellä muuttujien osittaisia järjestyksiä siten, että syntyvässä Bayes-verkossa muuttujasta X voi olla kaari muuttujaan Y vain jos X on Y:n edeltäjä kiinnitetyssä järjestyksessä. Tilastollisesti pääteltyjä osittaisia järjestyksiä voidaan myös yhdistää yllä esitettyyn kokonaisuskottavuusmetriikkaan perustuviin oppimisalgoritmeihin [130]. Osittaisen järjestyksen määräämiseksi voidaan tilastollisten testien sijaan käyttää tietenkin myös priori-informaatiota. Itse asiassa monet kokonaisuskottavuusmetriikkaa soveltavat oppimisalgoritmit, kuten esimerkiksi Cooperin ja Herskovitsin K2-algoritmi [28], edellyttävät tällaisen osittaisen järjestyksen määräämistä ennen oppimisprosessin alkua.

4.4.2 Bayes-verkon parametrien oppiminen

Suurimman todennäköisyyden parametrit. Kun Bayes-verkkostruktuuri M on kiinnitetty, on seuraava vaihe mallin oppimisessa mallin parametrien θ valinta. Kuten luvussa 2.2.3 näimme, bayesiläisessä lähestymistavassa parametrit kiinnitetään MAP-arvoonsa, siis siihen arvoon $\hat{\theta}$, joka maksimoi posterioritodennäköisyyden $P(\theta \mid \mathcal{D}, M)$:

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}, M).$$

Samojen teknisten oletusten vallitessa, joita käytettiin edellisessä luvussa esitetyn kokonaisuskottavuuden kaavan (4.4) johtamisessa, voidaan osoittaa että malliparametrien posterioritodennäköisyys maksimoituu asettamalla

$$\hat{\theta}_{ijk} = \frac{N_{ijk} + N'_{ijk} - 1}{N_{ij} + N'_{ij} - r_i}, \tag{4.5}$$

missä merkinnät ovat kuten edellä kaavassa (4.4). MAP-parametriarvojen laskemisessa ei siis tarvita iteratiivista oppimisalgoritmia, vaan ne voidaan määrätä suoraan käyttäen kaavaa (4.5)!

Suurimman uskottavuuden parametrit. On tärkeää huomata, että yllä annettu laskukaava MAP-parametrien määräämiseksi poikkeaa klassisesta tilastotieteestä tutusta *suurimman uskottavuuden (maximum likelihood, ML)* parametriasetuksesta $\tilde{\theta}$,

$$\tilde{\theta}_{ijk} = \frac{N_{ijk}}{N_{ij}} = \frac{N_{ijk}}{\sum_{k=1}^{r_i} N_{ijk}}, \tag{4.6}$$

joka maksimoi datan uskottavuuden $P(\mathcal{D} \mid \theta, M)$. Koska

$$P(\theta \mid \mathcal{D}, M) \propto P(\theta \mid M) P(\mathcal{D} \mid \theta, M),$$

näemme, että ML- ja MAP-parametrit ovat samat siinä erikoistapauksessa, että priorijakauma $P(\theta \mid M)$ on tasainen (siis vakio). Suurimman uskottavuuden parametrit saadaan siis erikoistapauksena kaavasta (4.5), jos käytämme tasaisia priorijakaumia, eli asetamme kaikkien hyperparametrien N'_{ijk} arvoksi yksi.

Odotusarvoparametrit. Parametrien θ asettaminen MAP-arvoihinsa on bayesiläisen perusidelogian mukainen vastaus kysymykseen kuinka kiinnittää parametrit annettuna mallistruktuuri. Mikäli syntyvää mallia (M, θ) kuitenkin käytetään probabilistisessa päättelyssä (eikä siis ainoastaan ongelmakentän analysoinnissa), näemme hiukan yllättäen että MAP-parametreja käyttävälle mallille löytyy vaihtoehto, joka tuottaa probabilistisessä päättelyssä tarkempia ennustuksia. Olkoon nimittäin $P(X_1, \ldots, Xn \mid \theta, M)$ mallin (M, θ) tuottama jakauma. Koska emme tiedä varmasti minkä mallin θ tuottama jakauma vastaa ongelmakentän todellista todennäköisyysjakaumaa parhaiten, arvioimme eri mallien hyvyyttä niiden posterioritodennäköisyyden $P(\theta \mid \mathcal{D}, M)$ mukaan käyttämällä apuna ongelmakentästä saatua otosta $\mathcal{D}$. Yllä esitetyssä MAP-lähestymistavassa valittiin siis parametrit, jotka maksimoivat tämän todennäköisyyden. Tarkempi jakauma saataisiin kuitenkin aikaan ottamalla esimerkiksi kaksi tässä mielessä parasta mallia θ_1 ja θ_2, ja muodostamalla niiden painotettu keskiarvo:

$$P(X_1, \ldots, Xn \mid \theta_1, M)P(\theta_1 \mid \mathcal{D}, M) + P(X_1, \ldots, Xn \mid \theta_2, M)P(\theta_2 \mid \mathcal{D}, M).$$

Seuraamalla tätä päättelyketjua eteenpäin, näemme että optimaalisessa tapauksessa meidän pitäisi käyttää *kaikkia (äärettömän monia) mahdollisia parametriasetuksia*, jolloin saamme jakauman

$$\int P(X_1, \ldots, Xn \mid \theta, M)P(\theta \mid \mathcal{D}, M)d\theta = P(X_1, \ldots, Xn \mid \mathcal{D}, M).$$

Voidaan kuitenkin osoittaa, että syntynyt jakauma on identtinen jakauman kanssa, joka saadaan käyttäen yhtä mallia jossa kukin parametriarvo on kiinnitetty odotusarvoonsa $E(\theta_{ijk} \mid \mathcal{D}, M)$. Odotusarvomallin käyttäminen vastaa siis, paradoksaalista kyllä, tilannetta jossa emme kiinnitä parametriarvoa lainkaan vaan käytämme kaikkia äärettömän monia parametriasetuksia!

Odotusarvoparametrit saadaan laskettua kaavasta

$$\hat{\theta}_{ijk} = E(\theta_{ijk} \mid \mathcal{D}, M) = \frac{N_{ijk} + N'_{ijk}}{N_{ij} + N'_{ij}}. \tag{4.7}$$

Odotusarvoihin perustuvien mallien on empiirisesti havaittu tuottavan paljon tarkempia estimaatteja ongelmakentän todennäköisyysjakaumasta kuin MAP- tai ML-parametrien, etenkin silloin kun käytettävissä olevan opetusjoukon $\mathcal{D}$ koko on pieni [76, 77, 138].

Kolme laskukaavaa malliparametrien määräämiseksi:

- Suurimman todennäköisyyden parametrit (4.5):

$$\theta_{ijk} = \frac{N_{ijk} + N'_{ijk} - 1}{N_{ij} + N'_{ij} - r_i}.$$

- Suurimman uskottavuuden parametrit (4.6):

$$\theta_{ijk} = \frac{N_{ijk}}{N_{ij}}.$$

- Odotusarvoparametrit (4.7):

$$\theta_{ijk} = \frac{N_{ijk} + N'_{ijk}}{N_{ij} + N'_{ij}}.$$

4.5 Bayes-verkkojen variaatioita

4.5.1 Päätösverkot (influence diagrams)

Luvussa 2.4 näimme, kuinka probabilistisessa päätelyssä tuotettuja ehdollisia todennäköisyysjakaumia voidaan käyttää hyväksi päätöksentekoprosessissa päätösteoriana tunnetun formalismin kautta. *Päätösverkko (influence diagram)* on Bayes-verkko, jossa päätöksentekoon liittyvät vaihtoehdot koodataan erityisinä päätössolmuina, ja päätössolmujen eri arvoihin liittyvä odotettavissa olevan hyödyn määrä lasketaan suoraan probabilistisen päättelyn yhteydessä (katso esim. [128, 104]). Päätössolmujen käsittelymahdollisuus on liitetty mukaan useisiin Bayes-verkkoja manipuloiviin ohjelmistoihin (katso luku 5.2).

4.5.2 “Noisy-or”-malli

Kuten luvussa 4.3 näimme, probabilistisen päättelyn eri muodot voidaan toteuttaa yksipolkuisilla (joko aidosti yksipolkuisilla, tai monipolkuista verkkoa vastaavalla liittymäpuulla) Bayes-verkoilla ajassa, joka on suoraan verrannollinen verkkoa vastaavien parametrien määrään. Luvussa 4.1 näimme että nämä parametrit ovat muotoa $P(X \mid F(X))$, missä $F(X)$ on solmun X edeltäjien joukko annetussa verkossa. Tätä muotoa olevia parametreja on eksponentiaalinen määrä solmun X edeltäjien arvojen lukumäärän suhteen, joten jos solmulla on hyvin suuri joukko edeltäjiä, saattaa tarvittavien parametrien lukumäärä olla liian suuri käytännön tarpeita ajatellen[5]. Ns. "noisy-or"-mallissa muotoa $P(X \mid F(X))$ olevia parametreja approksimoidaan kaavan

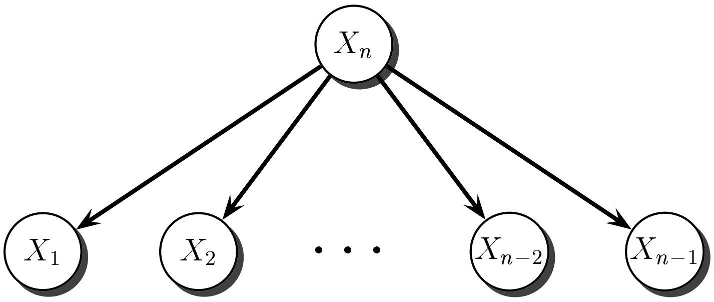
$$P(X \mid Y_1, \ldots, Y_n) = 1 - \prod_{i=1}^{n}(1 - P(X \mid Y_i))$$

mukaan. "Noisy-or"-mallia käyttäen voidaan tarvittavien parametrien lukumäärää pienentää siis huomattavasti: kutakin solmua X kohden riittää tallettaa $|F(X)|$ parametria, missä $|F(X)|$ on solmun edeltäjien lukumäärä. "Noisy-or"-approksimaation on empiirisesti havaittu toimivan useissa käytännön tilanteissa hyvin, vaikka toisinaan sen sisältämät implisiittiset riippumattomuusoletukset eivät pidä paikkaansa, jolloin approksimaatio ei ole riittävän tarkka käytännön sovelluksia ajatellen [54, 100].

4.5.3 Jatkuva-arvoiset muuttujat

Tässä raportissa on havainnollisuuden vuoksi keskitytty tarkastelemaan diskreettejä Bayes-verkkoja, joissa erilaisia mahdollisia muuttujien arvoja on vain äärellinen määrä. Koska Bayes-verkko kuitenkin viime kädessä on yhteistodennäköisyysjakauman hajotelman graafinen esitys, joka kertoo kuinka muuttujajoukon yhteistodennäköisyys saadaan laskettua ehdollisten todennäköisyyksien tulona (katso kaava 4.3), ei mikään estä käyttämästä tässä hajotelmassa myös jatkuvia ehdollisia tiheysfunktioita: sekä luvussa 4.3 hahmoteltu probabilistisen päättelyn toteuttamiseen käytetty viestinvälitysalgoritmi [85, 3], että luvussa esitetty Bayes-verkkojen rakentamisessa käytetty mallinvalintakriteeri [41] voidaan määritellä myös jatkuvien muuttujien tapauksessa, ja myös sekatapauksessa, jossa käytetään yhtäaikaa sekä jatkuvia että diskreettejä muuttujia. Malliperheen laajennus tällä tavalla saattaa kuitenkin joissakin tapauksissa aiheuttaa teknisiä ongelmia — yksinkertaisinta jatkuvien muuttujien käsittely on ns. äärellisissä sekajakaumamalleissa, joita käsittelemme luvussa 4.5.5. On myös huomattava, että käytännön sovelluksissa reaaliarvoisten muuttujienkaan arvot eivät todellisuudessa ole äärettömän tarkkuuden reaalilukuja, vaan jollakin äärellisellä tarkkuudella esitettyjä numeroita. Niinpä monissa sovelluksissa jatkuvia arvoja voidaankin yhtä hyvin käsitellä suoraan diskreetteinä muuttujina, tai ne voidaan ensin diskretisoida. Bayesiläistä lähestymistapaa diskretisointiin on käsitelty lähteissä [37, 72].

[5]Tosin käytännön sovelluksissa käytettävät Bayes-verkot ovat yleensä suhteellisen harvoja, jolloin tätä ongelmaa ei esiinny.



Kuva 4.8: Naiivi Bayes-mallin Bayes-verkkoesitys. Juurimuuttuja X_n on luokkamuuttuja, jonka suhteen muiden muuttujien riippumattomuusoletus tehdään.

4.5.4 Naiivi Bayes-malli

Luokitteluongelmissa tehtävänä on laskea diskreetin luokkamuuttujan jakauma annettuna muiden muuttujien (mahdollisesti osittainen) arvokombinaatio $\mathbf{S}_2 = \mathbf{s}_2$. Niin sanotussa *Naiivi Bayes*-mallissa muut muuttujat oletetaan riippumattomiksi, jos luokkamuuttujan arvo on tunnettu. Tästä oletuksesta seuraa, että muuttujien yhteistodennäköisyysjakauma voidaan esittää muodossa

$$P(X_1, \ldots, X_n) = P(X_1, \ldots, X_{n-1} \mid X_n)P(X_n) = P(X_n) \prod_{i=1}^{n-1} P(X_i \mid X_n),$$

missä muuttujat on oletettu järjestetyiksi siten, että luokkamuuttujan indeksi on n. Tämän hajotelman Bayes-verkkoesitys on kuvassa 4.8.

Naiivi Bayes-mallia on sovellettu monissa erilaisissa luokitteluongelmissa varsin hyvällä menestyksellä. Yksi syy mallin suosioon on probabilistisen päättelyn tehokkuus: koska Bayes-verkko on tässä tapauksessa kaksitasoinen puu, on laskenta äärimmäisen tehokasta. Lisäksi Naiivi Bayes-mallissa tapahtuva probabilistinen päättely voidaan toteuttaa feedforward-neuroverkon tapaisena rinnakkaisarkkitehtuurina [98, 97]. Mallin toinen merkittävä etu on se, että koska mallin arkkitehtuuri on kiinnitetty, ei mallistruktuurin oppimisongelmaa ole, ja mallin parametrit voidaan asettaa suoraan luvussa 4.4.2 esitettyjä kaavoja käyttäen.

Vaikka Naiivi Bayes-malliin johtava riippumattomuusoletus saattaa tuntua melko epärealistiselta, toimii Naiivi Bayes-malli käytännössä usein yllättävän hyvin [138, 84]. Itse asiassa monet Bayes-verkkojen kaupallisista sovelluksista käyttävät tätä yksinkertaista mallia (esimerkiksi Microsoftin Office

Assistant, ks. luku 5.1). Hyvä esimerkki Naiivi Bayes-mallin sovellettavuudesta saatiin "Knowledge Discovery and Data Mining"-konferenssin yhteydessä kesällä 1997 julkistetussa maailmanlaajuisessa ennustuskilpailussa, jossa kaksi kolmesta parhaasta ratkaisusta käytti juuri Naiivi Bayes-mallia [106].

Tämän yksinkertaisen mallin menestykselle voidaan antaa seuraava intuitiivinen selitys: monissa luokitteluongelmissa (esimerkiksi lääketieteellisissä ongelmissa) on muuttujien arvojen määrittäminen kallista, koska se edellyttää esimerkiksi erityisten laboratoriokokeiden suorittamista. Niinpä kahden toisistaan riippuvan muuttujan mukaanottaminen on resurssien tuhlausta, koska jo toinenkin näistä muuttujista sisältää tarvittavan informaation. Vaikuttaa siltä, että monet reaalimaailman datajoukot on kerätty resursseja säästäen, minkä seurauksena Naiivi Bayes-mallin riippumattomuusoletus saattaa olla hyvinkin järkevä.

Naiivi Bayes-mallin on toisaalta havaittu luokittelevan hyvin myös sellaisissa (keinotekoisissa) tilanteissa, joissa käytetty datajoukko on konstruoitu sellaiseksi, että Naiivi Bayes-mallin riippumattomuusoletus rikkoutuu varmasti [5]. Mallin luokittelutarkkuus on näissäkin tilanteissa ollut suuri — malli pystyy siis erottelemaan suurella tarkkuudella annetut esimerkit vastaaviin luokkiin — mutta jos tarkastellaan mallin antamaa jakaumaa luokkamuuttujalle, nähdään että mallin estimoima jakauma ei ole kovinkaan tarkka. Naiivi Bayes-mallin riippumattomuusoletuksesta seuraa siis ilmeisen konsistetti vinouma (bias) estimoituun luokkamuuttujan jakaumaan. Vinouma estää luokkamuuttujan jakauman tarkan estimoinnin, mutta koska oletuksista seurannut vinouma on kaikissa tilanteissa johdonmukainen, saattaa menetelmän erottelukyky olla edelleen hyvä. Tämä merkitsee siis sitä, että Naiivi Bayes-malli saattaa olla hyvä vaihtoehto silloin, jos luokittelussa riittää erotella annetut tapaukset eri luokkiin, mutta Naiivi Bayes-mallin tuottaman luokkamuuttujan jakauman käyttöön riskianalyysissä kannattaa suhtautua varauksella.

4.5.5 Latentit muuttujat ja äärelliset sekajakaumat

Standardit Bayes-verkkomallit perustuvat siihen oletukseen, että mallintamisessa käytettävät muuttujat $X_1, \ldots, X_n$ on valittu siten että ongelmakentälle voidaan konstruoida toimiva malli valittuja muuttujia käyttäen. Mikäli näin ei ole, voidaan tietysti kyseenalaistaa koko tehtävänasettelu, ja valita käytettävät muuttujat uudelleen. Jos muuttujajoukon uudelleenmäärittely ei kuitenkaan tällaisessa tilanteessa ole mahdollista, saattaa Bayes-verkkojen (tai minkä tahansa vain annettuja muuttujia käyttävän malliperheen) soveltaminen tuottaa epätyydyttäviä tuloksia. Seuraava yksinkertainen esimerkki havainnollistaa tätä seikkaa.

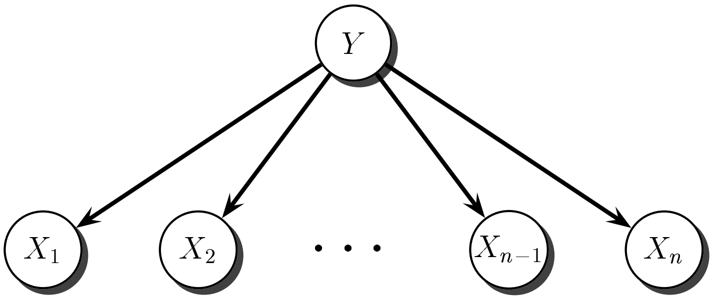
Oletetaan, että käytettävissä olevia muuttujia on kaksi, X_1, ja X_2, ja muuttujan X_1 arvo kertoo hukkumiskuolemien viikottaisen lukumäärän Suomessa, ja muuttujan X_2 arvo Suomessa viikottain myydyn jäätelön määrän. Tilastojen perusteella on helppo nähdä, että muuttujilla on selvästi jonkinlainen riippuvuussuhde, joten sellainen Bayes-verkkostruktuuri, jossa muuttujat eivät ole yhdistettyjä toisiinsa, ei mallinna ongelmakenttää tyydyttävällä tavalla. Jäljelle jää vain kaksi mahdollista Bayes-verkkostruktuuria: $X_1 \rightarrow X_2$ ja $X_1 \leftarrow X_2$. Intuitiivisesti ottaen nämä tapaukset näyttäisivät kuitenkin johtavan siihen, että jompi kumpi muuttuja olisi toisen muuttujan syy, mikä ei varmaankaan pidä paikkaansa todellisuudessa. Jos meidän nyt sallitaan lisätä malliimme uusi muuttuja Y, niin näemme että Bayes-verkkorakenne $X_1 \leftarrow Y \rightarrow X_2$ tuottaa muuttujien X_1 ja X_2 välille *epäsuoran riippuvuussuhteen*, joka mallintaa tyydyttävästi intuitiivista kuvaamme ongelmakentästä.

Edellisessä esimerkissä käytettyä muuttujaa Y kutsutaan *latentiksi muuttujaksi* tai *piilomuuttujaksi*, koska muuttuja ei ollut mukana alkuperäisessä muuttujajoukossa, eikä sen arvoja siten annettu opetusjoukossa $\mathcal{D}$. Yksinkertaisessa esimerkissämme on käytetylle latentille muuttujalle helppo keksiä semanttinen tulkinta (vuodenaika), mutta aina ei näin ole asianlaita, eikä se bayesiläisen mallintamisen kannalta ole välttämättä tarpeenkaan: latenttien muuttujien arvoja opetusjoukossa voidaan pitää puuttuvana datana, joka marginalisoidaan pois käytetyistä laskukaavoista.

Mikäli latentteja muuttujia on vain yksi, ja latentista muuttujasta on kaari kaikkiin muihin muuttujiin joiden välillä ei puolestaan ole keskinäisiä kaaria, kutsutaan syntyvää Bayes-verkkomallia *äärelliseksi sekajakaumamalliksi (finite mixture model)* [35, 141]. Syntyvä yhteistodennäköisyysjakauma on tässä tapauksessa

$$\begin{aligned}
P(X_1, \ldots, X_n) &= \sum_y P(X_1, \ldots, X_n, Y = y) \\
&= \sum_y P(X_1, \ldots, X_n \mid Y = y) P(Y = y) \\
&= \sum_y P(Y = y) \prod_{i=1}^{n} P(X_i \mid Y = y),
\end{aligned}$$

missä summa käy yli kaikkien mahdollisten piilomuuttujan Y arvojen. Vastaava Bayes-verkkostruktuurin graafinen esitys on kuvassa 4.9. On tärkeää huomata, että vaikka syntyvä mallistruktuuri on muodollisesti identtinen Naiivi Bayes-mallin (kuva 4.8) kanssa, on juurimuuttuja sekajakaumamallien tapauksessa latentti muuttuja (ja mahdollinen luokkamuuttuja on siis puun lehti siinä missä muutkin muuttujat), joten malli ei tee samoja riippumattomuusoletuksia kuin Naiivi Bayes-malli. Itse asiassa on helppo nähdä, että



Kuva 4.9: Äärellisen sekajakaumamallin Bayes-verkkoesitys. Juurimuuttuja Y on latentti piilomuuttuja, jonka suhteen muiden muuttujien riippumattomuusoletus tehdään.

mikä tahansa Bayes-verkkomalli voidaan muuttaa äärelliseksi sekajakaumamalliksi, eli sekajakaumamallit muodostavat universaalin malliperheen.

Bayes-verkkoihin verrattuna sekajakaumamalleilla on etuna se, että koska syntyvä Bayes-verkkostruktuuri on yksinkertainen puu, on sekajakaumamalleilla suoritettava päättely laskennallisesti aina erittäin tehokasta. Lisäksi mahdollisia verkkostruktuureja on vain yksi, joten mallistruktuuriavaruus ei ole kovin suuri: ainoa vapaa parametri on latentin muuttujan arvojen lukumäärä, jonka voidaan olettaa olevan pienempi kuin joukon $\mathcal{D}$ koko. Juurimuuttujan latenttiudesta seuraa kuitenkin, että kokonaisuskottavuus $P(\mathcal{D} \mid M)$, joka Bayes-verkkojen tapauksessa lasketaan helposti kaavaa (4.4) käyttäen, ei ole sekajakaumien tapauksessa laskettavissa käytännössä tarkasti: kokonaisuskottavuus saadaan marginalisoimalla yli kaikkien mahdollisten piilomuuttujan arvoista muodostuvien vektorien Z yli,

$$P(\mathcal{D} \mid M) = \sum_{Z} P(\mathcal{D}, Z \mid M),$$

joita on eksponentiaalinen määrä. Sekajakaumamallien kokonaisuskottavuuden laskemiseksi on kuitenkin kehitetty monia tehokkaita approksimaatiomenetelmiä (katso [74, 83, 80, 73, 26]).

Kokonaisuskottavuuden laskemisen lisäksi tekee piilomuuttujaoletuksesta implisiittisesti seuraava puuttuva informaatio myös sekajakaumamallien malliparametrien laskemisen vaikeammaksi kuin normaalien Bayes-verkkojen tapauksessa. Tähänkin tehtävään on kehitetty tehokkaita approksimaatiomenetelmiä, joista yleisin tunnetaan nimellä *Expectation-Maximization (EM)* [33, 94, 64, 13, 81]. Koska latenttia piilomuuttujaa Y voidaan ajatella eräänlaisena “ryhmittelymuuttujana”, joka jakaa annetut esimerkkitapaukset erillisiin ryhmiin (clusters), voidaan EM-algoritmia pitää probabilistisena ryhmittelyalgoritmina. Algoritmin tuottamia tapausryhmiä voidaan soveltaa ongelmakentän probabilistisessa analysoinnissa ja visualisoinnissa (data mining) samalla tavoin kuin esimerkiksi SOM-neuroverkkomallia [69]. Algoritmi tarjoaa siis bayesiläisen lähestymistavan *numeerisena taksonomiana* tunnetulle tutkimusalueelle [44].

Kuten luvussa 3.3 todettiin, sekajakaumamalliperhe tarjoaa yhtenäisen teoreettisen kehikon monille näennäisesti erilaisille lähestymistavoille, kuten esimerkkeihin perustuva päättely [75, 140, 139, 98, 71], ydinestimaattorit, ydinkantafunktiot (radial basis functions), sekä probabilistiset neuroverkot [12, 112], itseorganisoivat kartat [14], sekä piilo-Markov-mallit (hidden Markov models) [131, 30, 111]. Lähteissä [138, 140] raportoidut laajat empiiriset testitulokset osoittavat, että mallistruktuurin näennäisestä yksinkertaisuudesta huolimatta sekajakaumallit (ja myös jopa Naiivi Bayes-mallit) tuottavat konsistentisti erittäin hyviä tuloksia vaihtoehtoisilla menetelmillä (kuten esimerkiksi neuroverkot ja päätöspuut) saavutettuihin tuloksiin verrattuna. Erityisen kiinnostavaa on, että bayesiläiset mallit näyttävät suoriutuvan hyvin myös tilanteissa, joissa opetusdataa on hyvin vähän: saatavilla olevan datan määrän vähentäminen jopa 90 prosentilla ei näytä juurikaan huonontavan mallien antamia tuloksia [76, 77].

4.5.6 Kvalitatiiviset Bayes-verkot

Kvalitatiivisissa Bayes-verkoissa (qualitative Bayesian networks) [145, 34] luovutaan verkon parametrien (ehdollisten todennäköisyyksien) esittämisestä kvantitatiivisina reaalilukuina, ja siirrytään äärelliseen joukkoon mahdollisia kvalitatiivisia arvoja, esimerkiksi joukkoon $\{-, 0, +\}$. Verkon kaaret kuvaavat tässä tilanteessa kuinka tietyn solmun edeltäjät vaikuttavat vastaavaan muuttujaan X: vähentääkö ('-') vai lisääkö ('+') edeltäjien arvokombinaatio muuttujan X todennäköisyyttä, vai onko arvokombinaatio X:n suhteen irrelevantti ('0'). Vaikuttaa siltä, että kvalitatiivisten Bayes-verkkojen tapauksessa on mahdollista suorittaa määrättyjä probabilistisen päättelyn muotoja hyvin tehokkaasti myös monipolkuisten verkkojen tapauksessa, mutta tyhjentäviä teoreettisia tuloksia ei asiasta vielä ole. Suomenkielinen katsaus kvalitatiivisia verkkoja käsittelevän tutkimuksen nykytilaan löytyy lähteestä [132].

4.5.7 Bayes-verkkojen ja neuroverkkojen yhteyksiä

Kuten edellä on jo mainittu, bayesiläisiä tekniikoita voidaan periaatteessa soveltaa myös muiden malliperheiden kuin Bayes-verkkojen tapauksessa. Mackay [90] ja Neal [99] kuvaavat väitöskirjoissaan bayesiläisten tekniikoiden soveltamista suunnattujen neuroverkkojen oppimisessa. Mackay voitti kehittämällään ohjelmistolla vuonna 1993 järjestetyn laajan aikasarjaennustamiskilpailun [91]. Sekä Mackayn[6] että Nealin[7] kehittämät ohjelmistot ovat vapaasti saatavilla kehittäjiensä kotisivuilta.

Paitsi neuroverkkojen oppimisalgoritmien kehittämisessä, Bayes-verkkojen ja neuroverkkojen välille on luotu myös lukuisia muita yhteyksiä. Michael Jordanin tutkimusryhmä MIT:ssa on tutkinut Bayes-verkkojen ja suunnattujen neuroverkkojen välimuotoa, jossa Bayes-verkkojen muuttujien välisten riippuvuuksien voimakkuuksia kuvataan neuroverkoista tutulla sigmoidfunktiolla. Malliperheestä käytetään nimitystä *sigmoid belief networks* [122]. Bishop [12] ja Ripley [112] kuvaavat kirjoissaan ydinkantafunktioiden (radial basis functions), probabilististen neuroverkkojen, ydinestimaattoreiden ja sekajakaumamallien välisiä yhteyksiä. Myllymäki kuvaa väitöskirjassaan [96] kuinka Bayes-verkoissa suoritettava probabilistinen päättely voidaan suorittaa massiivisesti rinnakkaisella Boltzmannin kone-neuroverkkoarkkitehtuurilla. Bayesiläinen kehikko itseorganisoiville kartoille [69] on kuvattu artikkelissa [14].

[6] URL: http://wol.ra.phy.cam.ac.uk/mackay/homepage.html
[7] URL: http://www.cs.utoronto.ca/~radford/