Ohjaamaton oppiminen

Marko Salmenkivi

Johdatus koneoppimiseen, syksy 2008

Luentorunko keskiviikolle 26.11.2008

- Ohjaamaton oppiminen
  - Mikä erottaa ohjatusta oppimisesta?
  - Esimerkkejä
- Johdattelua ryvästämiseen eli klusterointiin
- Aineiston esikäsittely ja esitysmuodot

Luentomateriaali perustuu huomattavassa määrin Jarmo Hurrin ja Juho Rousun kurssimateriaaliin kahtena edellisenä lukuvuonna luennoidulle kurssille Laskennallinen data-analyysi I

- Tähän mennessä kurssilla on käsitelty ohjattua oppimista: tavoitteena ennustaa piirrettä y, annettuna x.
- Tämä asetelma ei kuitenkaan sovi kaikkiin data-analyysitehtäviin
- Usein on tarpeen määritellä data-analyysitehtävä siten, että datajoukossa ei ole erikseen annettua ennustettavaa piirrettä $\longrightarrow$ ohjaamaton oppiminen

Ohjaamaton vs. ohjattu oppiminen

- ohjaamattomassa oppimisessa sana “ohjaamaton” viittaa siihen, ettei ole määritelty ennustettavaa piirrettä
- ohjaamaton oppiminen on hyvin heterogeeninen kategoria
- ohjaamattomassa oppimisessa pyritään kuvaamaan aineiston rakennetta oppimalla jokin aineistoon sopiva malli
- malli tässä hyvin yleinen käsite
- mallin tarkoitus kuvata aineiston oleellisia piirteitä (eikä esim. kohinaa, vrt. ylisovittaminen)
- ohjaamattomassa oppimisessa (välittömänä) tavoitteena ei yleensä ole ennustaminen

Ohjaamattomuudesta

- sananmukaisesti täysin ohjaamaton oppiminen on käytännössä mahdotonta
- vaikka ennustettavaa piirrettä ei ole kiinnitetty etukäteen, joudutaan moniin muihin asioihin ottamaan kantaa

Ohjaamaton vs. ohjattu oppiminen

Ohjaamaton oppiminen:

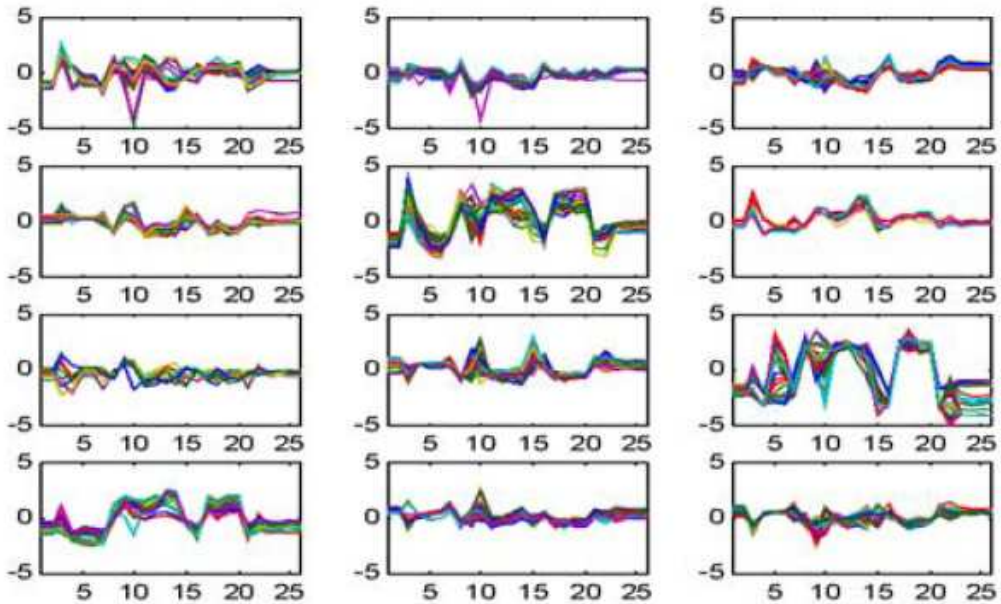
1. (tietyn piirteen) ennustamisen sijasta ollaan kiinnostuneita kuvaamaan aineiston rakennetta
2. ohjatun oppimisen vaatimien opetusesimerkkien hankkiminen on liian kallista / vaivalloista / haitallista / vaarallista.
3. ennustettava piirre on vaikeasti formalisoitavissa ja siten ennustustehtävä on vaikea määritellä, esim. mikä on relevantti dokumentti Google-haussa
4. aineistossa voi olla lukuisia piirteitä ja niiden yhdistelmiä, joiden ennustamisesta voidaan periaatteessa olla kiinnostuneita $\longrightarrow$ aineistoa pitää tutkia, ennen kuin tiedetään tarkalleen, mikä ennustustehtävä halutaan ratkaista

Esimerkki: prosessinvalvonta

- Tarkastellaan tuotantoprosessia, jota valvotaan jatkuvilla mittauksilla
- Haluttaisiin rakentaa työkalu, joka mittauksien perusteella antaisi varoituksen prosessin ajatumisesta pois halutusta toimintatilasta
- Luokittelijan oppiminen edellyttäisi esimerkkejä normaalista ja epänormaalista toimintatilasta.
- Epänormaalia toimintatilaa edustavien esimerkkien generoiminen tarkoittaa tuotantomenetyksiä eli kallista hintaa
- Olisi suotavaa rakentaa työkalu pelkästään normaalitilaa kuvaavien esimerkkien perusteella

Esimerkki: prosessinvalvonta

- Prosessinvalvontatehtävä voidaan ratkaista keräämällä mittausdataa normaaleista toimintaoloista
- Mittauksista saadut profiili pyritään ryhmittelemään samankaltaisiin ryppäisiin
- Kutakin ryvästä asetetaan vastaamaan prototyyppiprofiili
- Poikkeustilanteeksi tulkitaan mittausprofiili, joka poikkeaa kaikista prototyypeistä "liikaa"



Esimerkki: kielimallit

- tilastollisessa konekääntämisessä tarvitaan kielimalleja huolehtimaan tuotetun käännöksen sujuvuudesta ja oikeakielisyydestä
- käännöksen sujuvuutta ja oikeakielisyyttä on vaikea lähestyä luokittelutehtävänä: negatiivisia esimerkkejä, "huonoa kieltä", on melko vaikeaa hankkia
- tavallisesti käytetään malleja, joissa tarkastellaan peräkkäisiä kolmen sanan ryhmiä
- lauseen $s_1 \cdots s_n$ todennäköisyys

$$P(s_1 \cdots s_n) = P(s_3|s_2 s_1) \cdot P(s_4|s_3 s_2) \cdots P(s_n|s_{n-1} s_{n-2}),$$

  missä s_i ovat lauseen sanat.
- kunkin sanan esiintymisen todennäköisyyttä (ja sitä kautta sanan esiintymisen kielellistä mielekkyyttä) tarkastellaan siis siinä valossa, mitkä ovat sitä edeltävät kaksi sanaa

Esimerkki: kielimallit

- Kielimalli

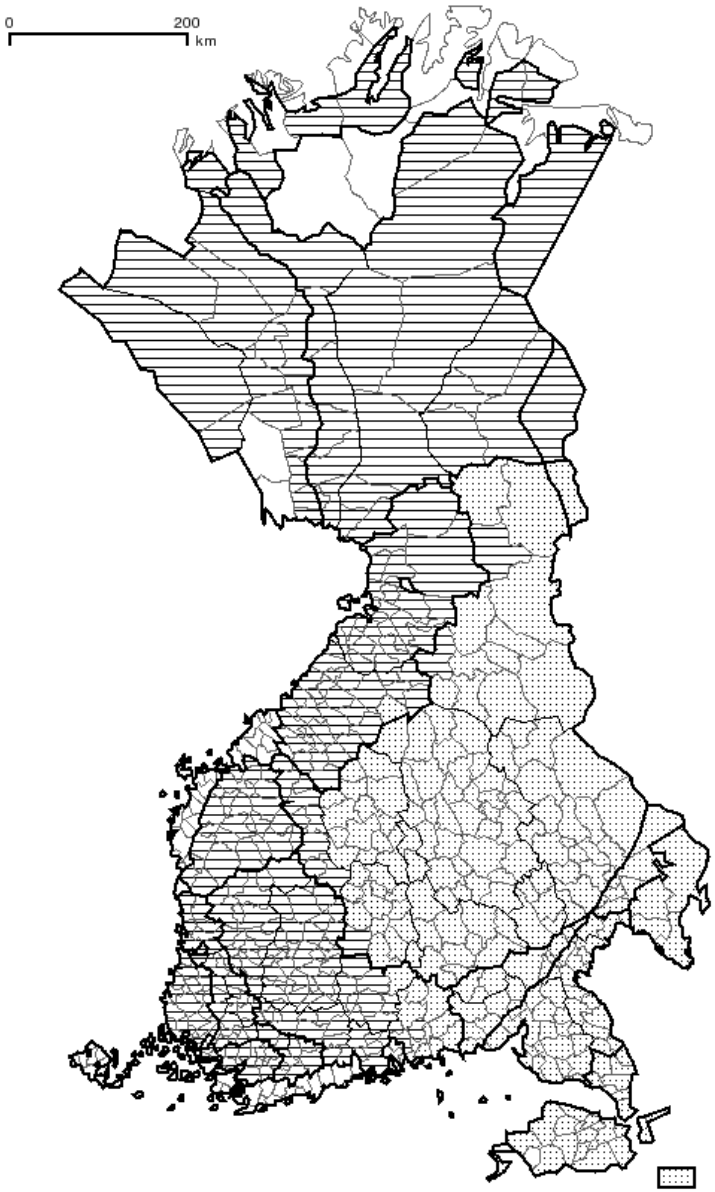
$$P(s_1 \cdots s_n) = P(s_3|s_2 s_1) \cdot P(s_4|s_3 s_2) \cdots P(s_n|s_{n-1} s_{n-2}),$$

  voidaan oppia yksinkertaisesti laskemalla sanakolmikkoja suuresta kohdekielen tekstiaineistosta
- Google Language model on rakennettu keräämällä tätä tietoa www:stä usean DVD:n verran

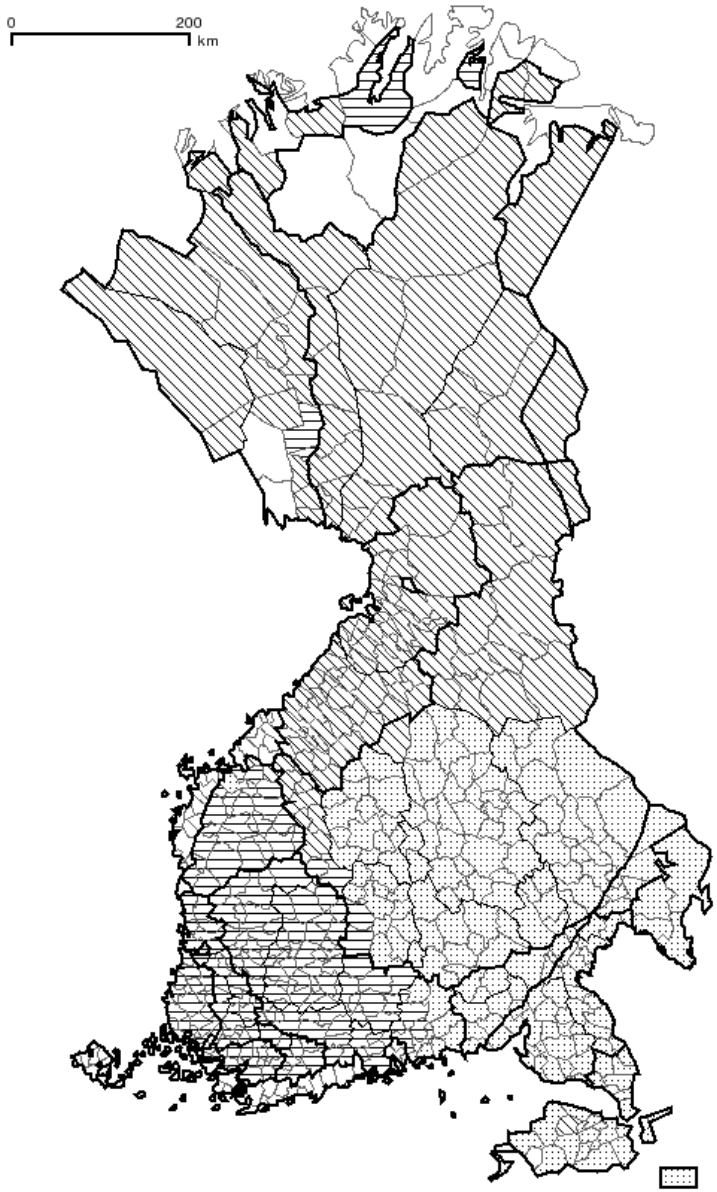
Esimerkki: kielitieteellinen data-analyysi

- aineistona suuri joukko suomen kielen murresanoja ja kuhunkin sanaan liittyen joukko pitäjiä, joissa sanaa on havaittu käytetyn
- onko sanojen maantieteellisten jakaumien perusteella hahmoteltavissa murteeltaan samankaltaisten pitäjien ryhmiä?
- ennustaminen ei selvästikään ole mielekäs kysymyksenasettelu

Esimerkki: murresanasto



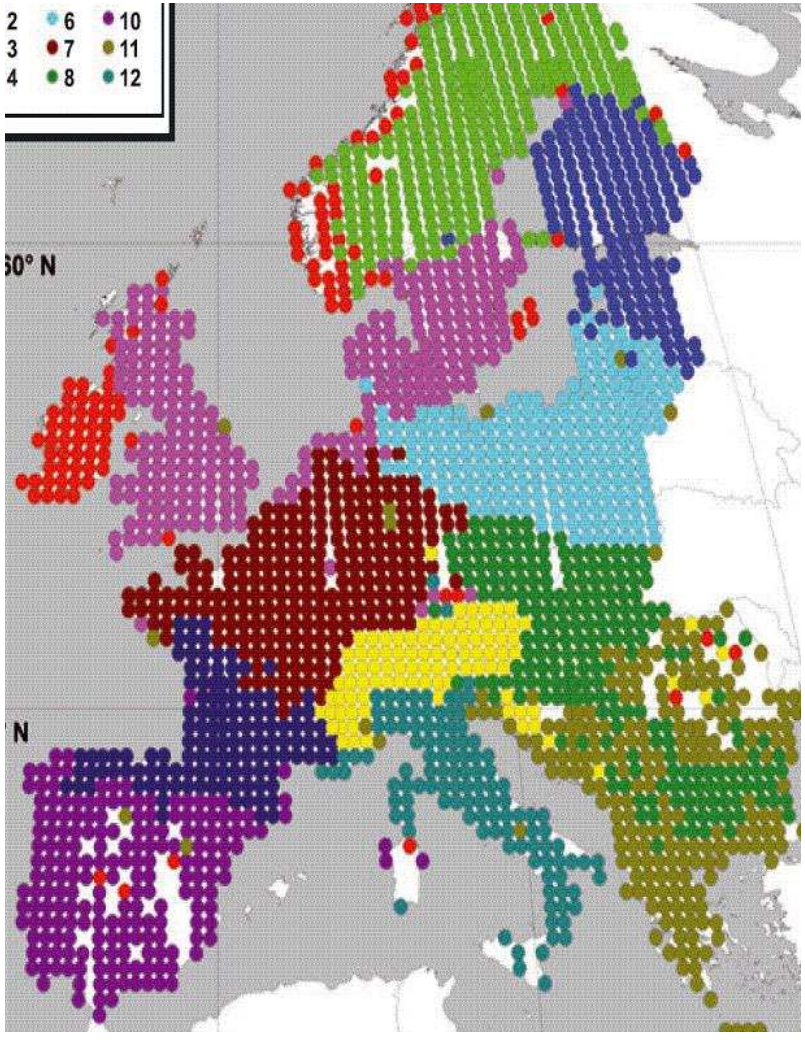
Pohjakartta © Genimap oy, lupa L6199/05-11



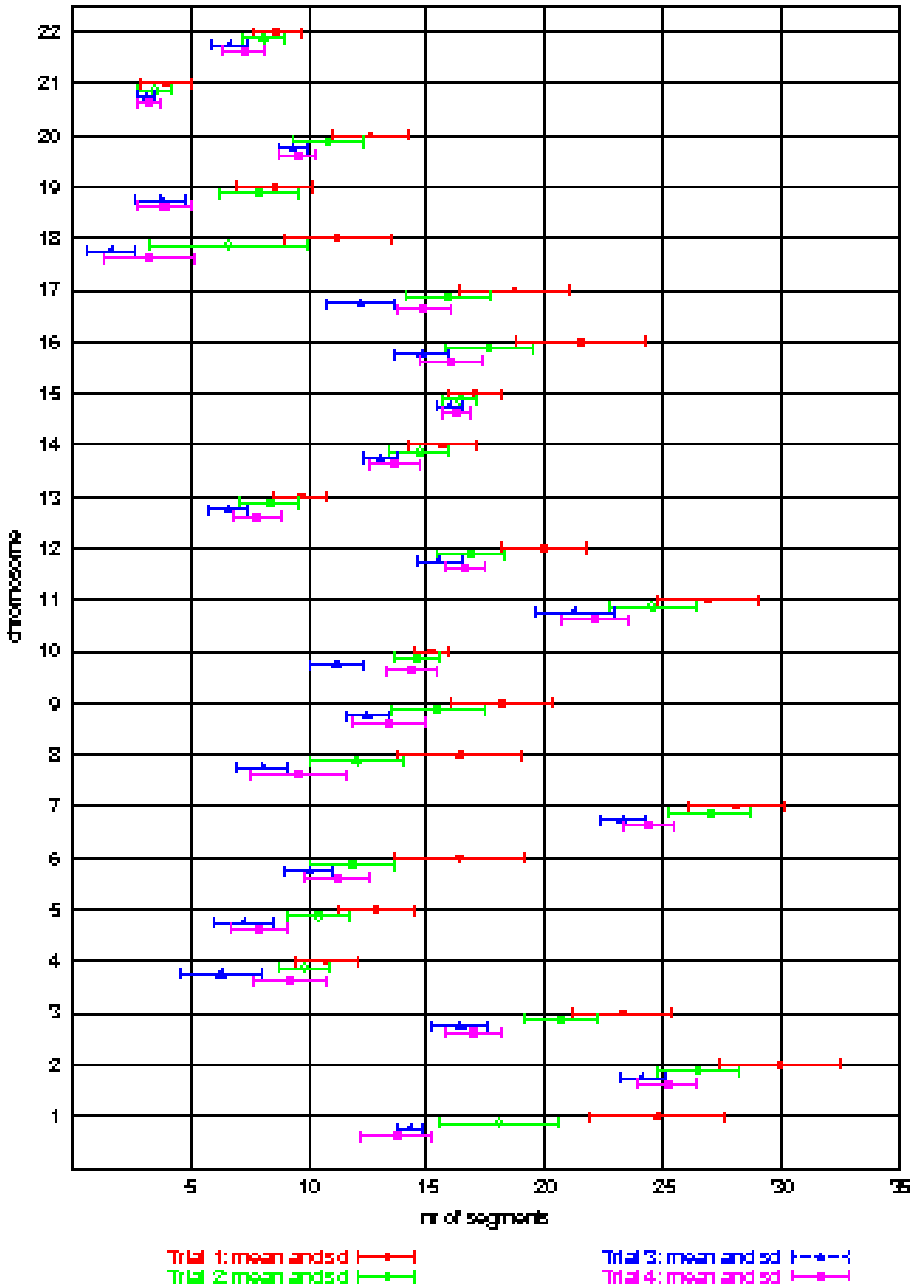
Pohjakartta © Genimap oy, lupa L6199/05-11

Esimerkki: luonnontieteellinen data-analyysi

- Tieteellisessä data-analyysissä halutaan usein löytää datasta aikaisemmin tuntemattomia riippuvuuksia (tai todentaa aikaisemmin tunnettuja riippuvuuksia uudella aineistolla ja/tai menetelmällä)
- Kuvassa on tutkittu nisäkkäiden esiintymistä 50x50 km ruuduissa
- Kutakin ruutua vastaa 124-ulotteinen binäärivektori (laji "i"esiintyy/ei esiinny)
- Väritys kuvaa samankaltaisten lajiprofiilien esiintymäalueita



Esimerkki: DNA-sekvenssin segmentointi



Esimerkki: sääntöjen louhinta

- NBA-koripalloliigassa pidetään tarkkaa kirjaa pelitapahtumista ja pelaajien tekemisistä
- Tuloksena on suuri tietokanta, josta voidaan etsiä riippuvuuksia, jotka jäisivät ehkä muuten huomaamatta
- Advanced Scout -järjestelmä [1] etsii sääntöjä kuten "`Kun pelaaja` X `on kentällä, pelaajan` Y `heittotarkkuus putoaa 75 prosentista 30 prosenttiin`"
- Tämän tyyppistä data-analyysia käsitellään lisää keväällä kurssilla Tiedon louhinta

[1] Bhandari I., Colet, E., Parker, J., Pines Z., Pratap R., Ramanujam K. (1997): Advanced Scout: datamining and knowledge discovery in NBA data. Data Mining and Knowledge Discovery, 1 (1), 121–125

Signaalien erottaminen

- Havaitaan signaali, joka on yhdistelmä useasta riippumattomasta lähteestä
- Tavoitteena on erottaa lähdesignaalit toisistaan
- Riippumattomien komponenttien analyysi (ICA) on eräs menetelmä tälläisen ongelman ratkaisemiseksi

`www.cis.hut.fi/projects/ica/cocktail/cocktail_en.cgi`

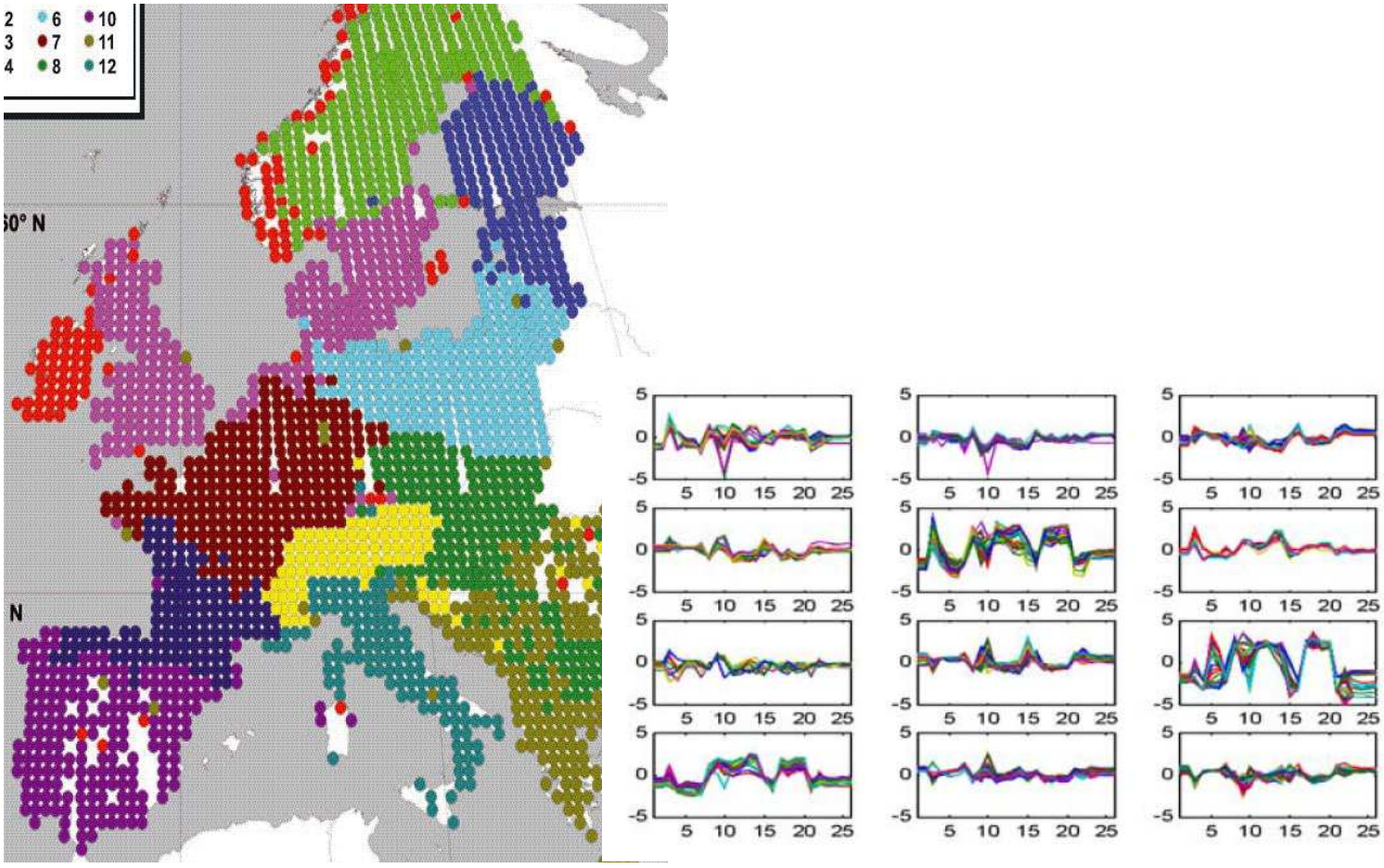
Ryvästäminen (klusterointi, engl. clustering)

- ohjaamattoman oppimisen menetelmistä keskitymme tällä kurssilla vain ryvästämiseen eli klusterointiin
- ryvästämisessä tehtävä on jakaa data erillisiin osajoukkoihin siten, että kukin osajoukko on niin homogeeninen kuin mahdollista
- Esimerkkejä klusteroinnin käyttökohteista
  - digitaalisten kuvien segmentointi
  - markkina-analyysi (asiakassegmentit)
  - geenien ryhmittely vaikutusten perusteella
  - tekstidokumenttijoukkojen ryhmittely

Ryvästämismenetelmien komponentit

- erilaisia ryvästämismenetelmiä on paljon
- yleisellä tasolla ryvästämismenetelmissä voidaan erottaa seuraavat komponentit:
  1. Kustannusfunktio, joka mittaa esimerkkiryppäiden homogeenisuuden
  2. Valintakriteeri ryppäiden määrälle
  3. Algoritmi, jolla esimerkit jaetaan ryppäisiin

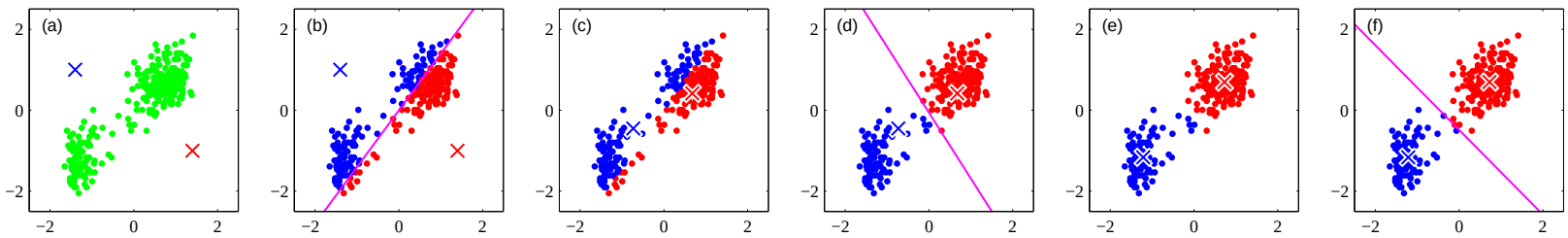
Kaksi ryvästämistehtävää



Ryvästämisalgoritmit voivat tuottaa

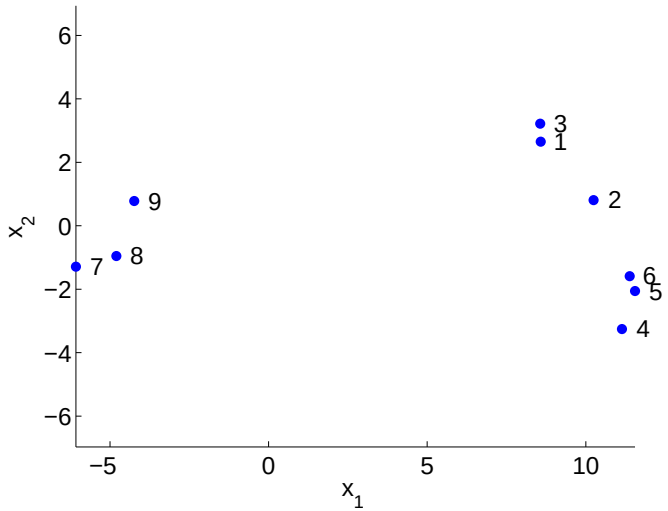
- pistejoukon ryhmittelyn
  - kohteiden “kovan” ryhmittelyn: kukin piste voi kuulua vain yhteen ryhmään
    - esimerkiksi K-means
  - “pehmeän” tai probabilistisen ryhmittelyn (kohteet voivat kuulua eri määrin useaan ryhmään)
    - esimerkiksi gaussinen sekoitusmalli (ei käsitellä)
- hierarkkisen ryhmittelyn (ryhmittelypuun)
- tasoesityksen datasta, siten että datan klusterit erottuvat tason eri alueina (ei käsitellä)
  - esimerkiksi itseorganisoiva kartta

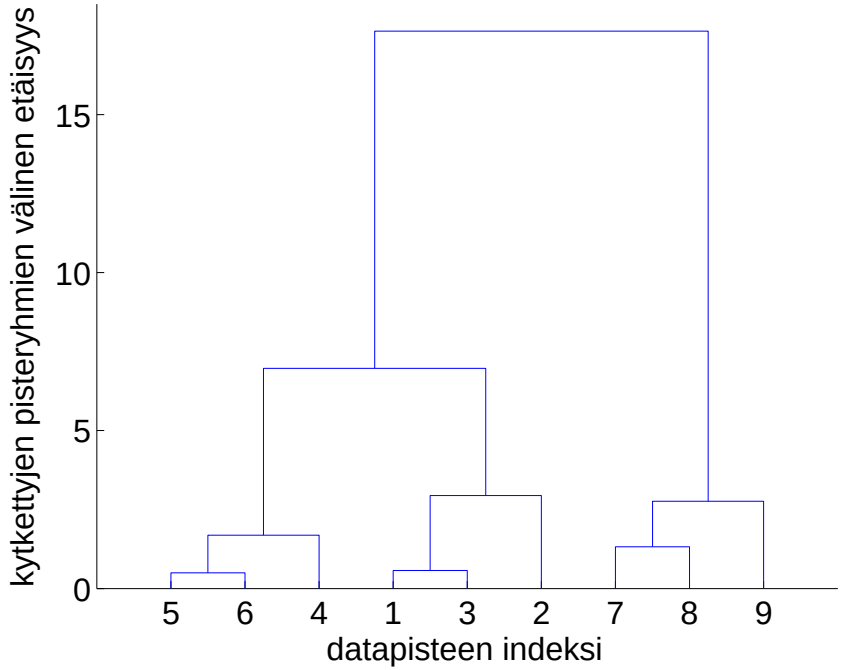
“Kova” klusterointi: esimerkki (K-means)



- data Old Faithful -nimisestä kuumasta lähteestä; vaaka-akseli purkauksen kesto, pystyakseli aika seuraavaan purkaukseen (nollakeskiarvoistettuina ja skaalattuina)

Hierarkkinen klusterointi: esimerkki





- dendrogrammipuuta (oikealla) luetaan seuraavasti:
  - puun lehtinä kaikki datapisteet eli yhden pisteen pisteryhmät
  - alhaalta ylöspäin edetessä yhdistetään lähimmät pisteryhmät toisiinsa; tässä pisteryhmien välinen etäisyys ryhmien kauimmaisten pisteiden etäisyys
  - yhdistämistaso (vaakaviiva) kytkettyjen pisteryhmien välisen etäisyyden tasolla

Esikäsittely ja esitysmuodot

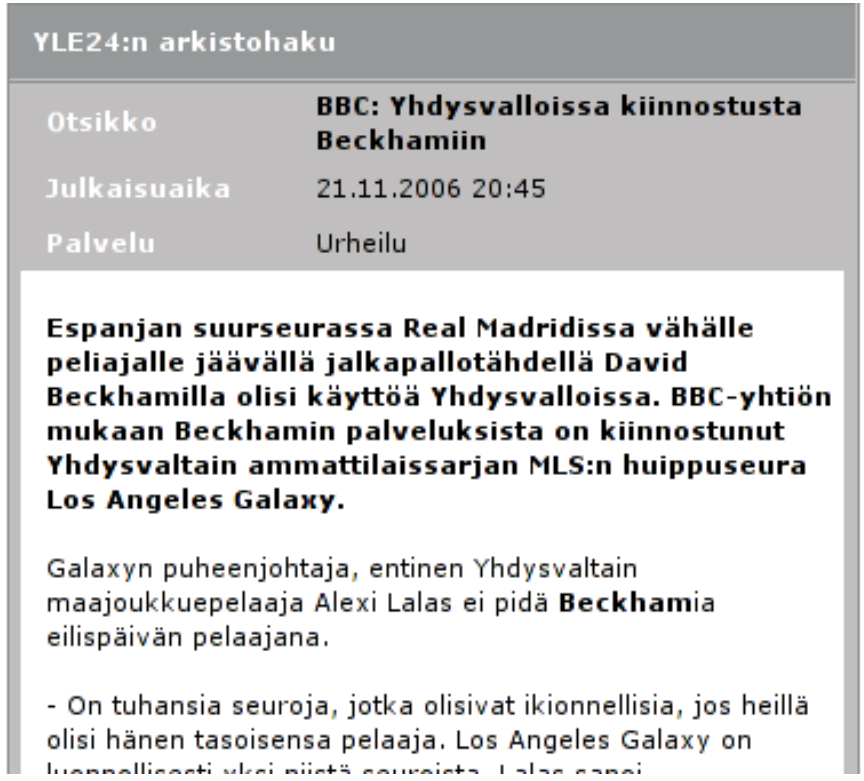
- Tähän asti kurssilla data on oletettu tupsahtaneeksi jostain sopivasti esikäsiteltynä numeerisiksi piirrevektoreiksi
- Käytännössä data ilmenee kuitenkin moninaisissa esitysmuodoissa (kuva, teksti, signaalit, monivalintakysymysten vastaukset, . . .)
- Ohjatun oppimisen menetelmät ovat vahvasti riippuvaisia datan esitysmuodosta
- Kustannusfunktiot perustuvat useimmiten datapisteiden välisen etäisyyden mittaamiselle
- Datan esikäsittelyn yksi tavoite on saada data sellaiseen muotoon, että etäisyyksiä voidaan mitata mielekkäästi

Esikäsittely ja esitysmuodot

- Kaksi lähestymistapaa datan esitysmuotojen suhteen:
  - Esikäsittely + yleiskäyttöinen oppimisalgoritmi, syötteenä (yleensä numeerinen) piirrevektori
  - Esitysmuotospesifinen oppimisalgoritmi; oma menetelmä kuville, oma tekstille, jne.
- Halutaan tuottaa piirre-esitys, jolla pystytään mittaamaan merkityksellisiksi ajateltujen hahmojen tai ominaisuuksien esiintymistä aineistossa

Esimerkki: tekstinhaku

- Tehtävä: Halutaan etsiä uutistietokannasta artikkelit, jotka kertovat David Beckhamin siirtymisestä Real Madridista LA Galaxyyn
- Piirreesityksenä sanasäkki (bag of words): $\phi_{Beckham}(\mathbf{b})$ kertoo montako kertaa *Beckham* esiintyy dokumentissa $\mathbf{b}$; sanan esiintymien sijainnista ei olla kiinnostuneita.
- $\phi_{Beckham}(\mathbf{b}) = 4$, $\phi_{Real}(\mathbf{b}) = 1$, $\phi_{Madrid} = 1$, $\phi_{Galaxy} = 3$, $\phi_{BBC} = 2$, ...



YLE24:n arkistohaku

Otsikko	**BBC: Yhdysvalloissa kiinnostusta Beckhamiin**
Julkaisuaika	21.11.2006 20:45
Palvelu	Urheilu

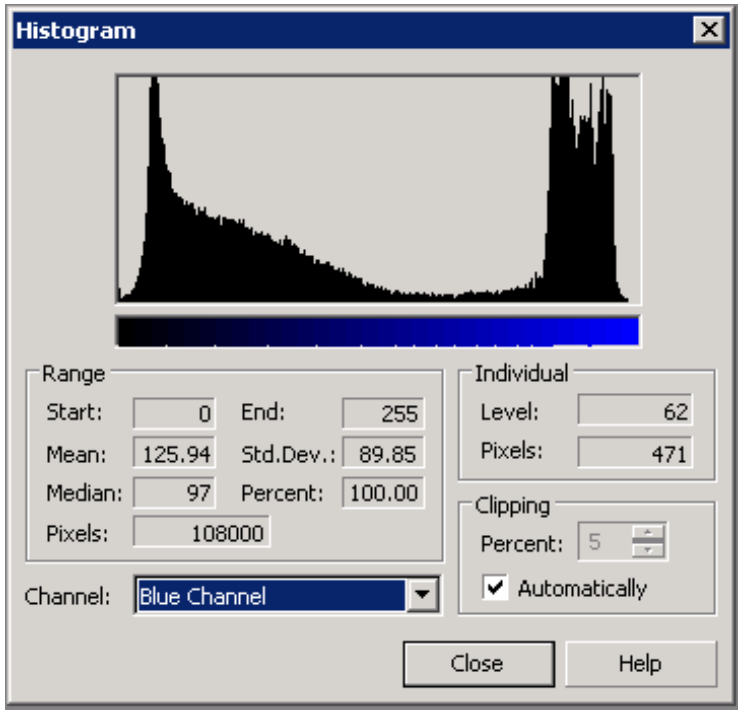
Espanjan suurseurassa Real Madridissa vähälle peliajalle jäävällä jalkapallotähdellä David Beckhamilla olisi käyttöä Yhdysvalloissa. BBC-yhtiön mukaan Beckhamin palveluksista on kiinnostunut Yhdysvaltain ammattilaissarjan MLS:n huippuseura Los Angeles Galaxy.

Galaxyn puheenjohtaja, entinen Yhdysvaltain maajoukkuepelaaja Alexi Lalas ei pidä **Beckham**ia eilispäivän pelaajana.

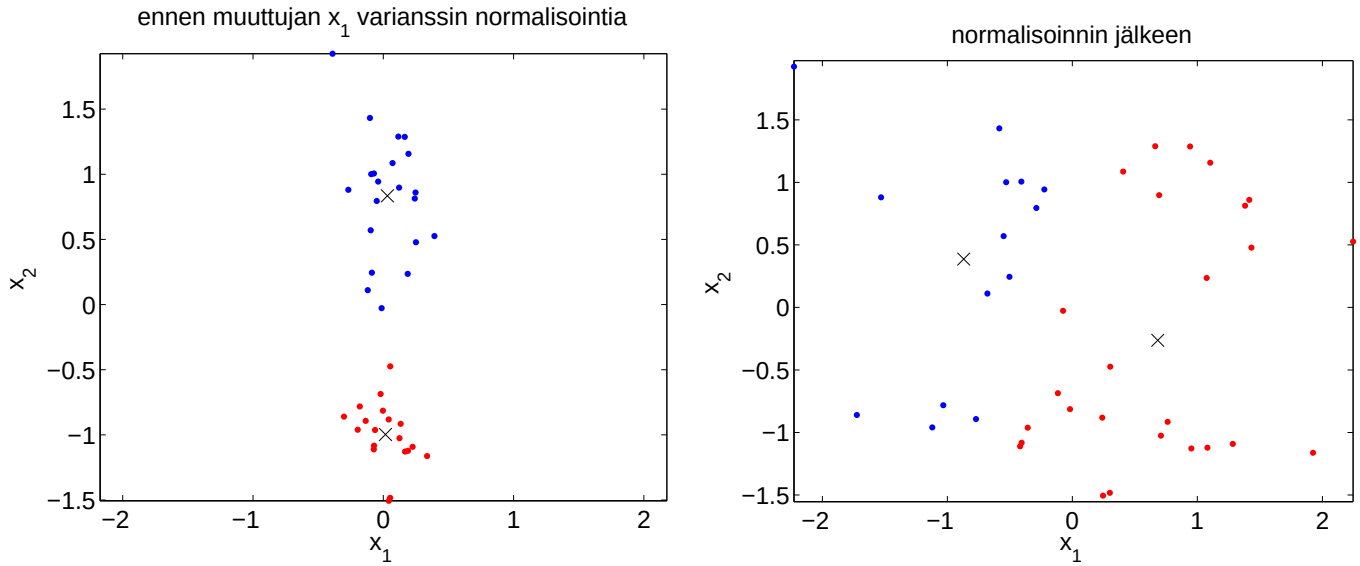
- On tuhansia seuroja, jotka olisivat ikionnellisia, jos heillä olisi hänen tasoisensa pelaaja. Los Angeles Galaxy on

Esimerkki: kuvanhaku

- Tavoite: Halutaan etsiä järvimaisemia kuvatietokannasta
- Piirre-esityksenä kuvien värihistogrammit: $\phi_i(\mathbf{b})$ on värisävyn i pikselien lukumäärä kuvassa $\mathbf{b}$
- Pikselien sijainnista ei olla kiinnostuneita: kuvakulman kiertäminen (rotation) tai siirtäminen (translation) suhteen ei vaikuta



Esimerkki, muuttujien skaalaaminen



- muuttujien skaalaamisella voi olla dramaattinen vaikutus tuloksiin
- muuttujien standardointi/normalisointi: keskiarvoksi nolla ja varianssiksi 1 (vähennetään keskiarvo ja jaetaan keskihajonnalla)
- molemmissa kuvissa varianssi muuttujan x_2 suuntaan 1

Muuttujien esikäsittely: numeerinen data

Numeerisen datan esitysmuodoissa ongelmia voivat aiheuttaa:

- Erilaiset arvoalueet/yksiköt: piirre x_1 mitattu kilogrammoina, x_2 grammoina, euklidisessa etäisyydessä grammoina mitattu piirre saa 1000-kertaisen painoarvon
- Poikkeava varianssi: piirre x_1 vaihtelee absoluuttisesti vähemmän kuin piirre x_2 , tällöin pieni muutos x_1:ssä voi olla yhtä tärkeää kuin suuri muutos x_2:ssä

Muuttujien esikäsittely: numeeriset piirteet

Piirteiden erilaisista skaaloista ja variansseista päästään eroon normalisoimalla

1. Keskitys ja jakaminen keskihajonnalla:

$$\phi_j(x) = (x_j - \mu_j)/\sigma_j,$$

   μ_j on piirteen j keskiarvo datajoukossa, σ_j keskihajonta; käy kaikille numeerisille piirteille

2. Jos arvot sijoittuvat välille $[x_{min}, x_{max}]$

$$\phi_j(x) = (x_j - x_{min})/(x_{max} - x_{min})$$

Esimerkki: nominaaliarvoiset syötemuuttujat

- Monissa data-analyysitehtävissä data ei ole valmiiksi numeerista, vaan joudumme muuntamaan datan numeeriseksi käyttämällä piirrefunktioita
- Oletetaan syötemuuttuja $x_j \in V_j$, missä arvojoukko $V_j = \{v_1, \ldots, v_r\}$ on nominaalinen (alkioilla ei järjestysrelaatiota)
- Muodostetaan piirrefunktio muuttujan x_j kullekin mahdolliselle arvolle $v_h \in v_j$:

$$\phi_{j,v_h}(x) = \begin{cases} 1, & x = v_j \\ 0, & x \neq v_j \end{cases}$$

Esimerkki: nominaaliarvoiset syötemuuttujat

- esim. klassisessa 'Mushrooms' (sienien luokittelu) aineistossa muodostettaisiin piirrefunktiot

  $\phi_{capshape,bell}, \phi_{capshape,conical}, \ldots$

---

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?: bruises=t,no=f