Sekoitemallit ja EM-algoritmi klusteroinnissa

Ilkka Autio
Ilkka.Autio@cs.helsinki.fi

Klusterointimenetelmät-seminaari
Helsingin yliopisto, tietojenkäsittelytieteen laitos
Raportti C–2002–54, s. 127–138, marraskuu 2002

Tiivistelmä

Tässä raportissa käsitellään äärellisten sekoitemallien sovittamisen perusteita. Sekoitemallit ovat hyödyllisiä silloin kun perinteiset parametriset todennäköisyysmallit ovat liian yksinkertaisia mallinnettavaan ilmiöön nähden ja silloin kun parametriton mallintaminen on hankalaa esim. muuttujien jatkuvuuden ja datan ulottuvuuksien määrän takia.

Tämän esityksen puitteissa sekoitemallien sovittaminen suoritetaan EM-algoritmiksi kutsutulla iteraatiolla, joka perustuu uskottavuuden maksimoinnin periaatteeseen. EM-algoritmille esitetään yksityiskohtainen johto multinormaalijakaumien sekoitteiden tapauksessa.

Sovitettua sekoitemallia voidaan lopulta käyttää muun muassa klusterointiin. Tiettyjen oletusten vallitessa sekoitemallit tarjoavat matemaattisesti hyvin perustellun klusterointikriteerin: Kaikille datavektoreille saadaan todennäköisyysarvio kuhunkin klusteriin kuulumisesta ja tällöin vektori voidaan esimerkiksi olettaa kuuluvaksi siihen klusteriin, jonka todennäköisyys on suurin.

Lopuksi tarkastellaan EM-iteraation tunnettuja ominaisuuksia ja heikkouksia, jotta lukija voisi edes hiukan arvioida menetelmän soveltuvuutta käytäntöön.

1 Johdanto

Sekoitemalleissa oletetaan, että kukin datavektori on peräisin jostain "lähdejakaumasta". Nämä lähteet voidaan sopivissa olosuhteissa tulkita esimerkiksi klustereiksi. Reaaliarvoisen datan tapauksessa koko sekoitemalli ilmaistaan tiheysfunktiona, joka on lähdejakaumien tiheysfunktioiden painotettu summa. Tässä reaaliarvoisen datan tapauksessa lähdejakaumat ovat usein tavallisia ja hyvin tunnettuja parametrisiä jakaumia (esim. multinormaaleja).

Sekoitemallien käyttöä voidaan ainakin multinormaalien lähteiden tapauksessa motivoida seuraavasti: Jos sekoitteessa käytetään tarpeeksi suurta määrää multinormaaleja komponentteja, niin sekoitteella voidaan approksimoida mielivaltaisen hyvin mitä tahansa tiheysjakaumaa [1]. Toisin sanoen tällaisilla sekoitemalleilla on periaatteessa kapasiteettia matkia sitä jakaumaa, josta data on "oikeasti" peräisin jos "tosi" malli ei olekaan se valittu sekoite.

Nyt voidaan kysyä, että miksi reaaliarvoisen datan tapauksessa pitäisi käyttää parametrisia jakaumia edes lähteinä ? Vastaus on se, että mallin yleistyskyvyn kannalta on välttämätöntä tehdä oletuksia lopullisen tiheysjakauman "pehmeydestä". Jos datan "oikea jakauma" on jatkuva, niin mikään datavektori ei esiinny toistuvasti. Toisin sanoen mallin sovituksen jälkeisessä käytössä ei tulla periaatteessa enää näkemään niitä datavektoreita joiden perusteella malli sovitettiin. Jos ei tehdä oletuksia tiheysjakauman pehmeydestä, niin uusien datavektoreiden todennäköisyyksistä (tai klustereista) ei voida sanoa yhtään mitään vaikka ne olisivat mielivaltaisen "lähellä" koulutusdataa (koska tällöin ei tehdä mitään oletuksia siitä, että "läheisyys" merkitsisi mitään).

Yllä olevan perusteella voidaan väittää, että sekoitemallit ovat reaaliarvoisen datan tapauksessa hyvin perusteltu mallityyppi. Sekoitemallin sovituksesta seuraa, että datavektoreille voidaan aina laskea todennäköisin lähde. Kun samaistetaan lähteiden ja klustereiden käsitteet, datavektoreille voidaan aina osoittaa todennäköisin klusteri. Jos malli on "oikea" ja se on sovitettu hyvin, niin saadaan aikaan matemaattisesti hyvin perusteltu klusterointimenetelmä. Kannattaa kuitenkin huomata ettei yllä mainittu approksimointitulos kerro paljoakaan menetelmän soveltuvuudesta klusterointiin.

Alla kuvataan täsmällisemmin perusasiat äärellisistä sekoitemalleista, sekä kerrataan suurimman uskottavuuden periaate. Perusteellisempaa oppimateriaalia löytyy kirjallisuudesta (esim. [2]). Suurimman uskottavuuden periaate löytynee melkein mistä tahansa tilastomatematiikan oppikirjasta.

1.1 Äärelliset sekoitemallit

Oletetaan, että datavektorit $\mathbf{x}_i$ ovat d-ulotteisia reaalilukuvektoreita avaruudesta $\mathbb{R}^d$. Klusteroitava (tai muuten vain mallinnettava) data on riippumaton (i.i.d.) satunnaisotos $\mathbb{R}^d$:lle määritellystä todennäköisyysjakaumasta. Otoksen koko on N ja otos voidaan esittää joukkona $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Lisäksi dataa generoivalta prosessilta oletetaan seuraavat ominaisuudet:

- Kukin datavektori on peräisin yhdestä "lähteestä". Lähde määrää ainakin osittain datavektorin sisällön.
- Lähteitä on äärellinen määrä. Lähteiden määrä on ennalta tunnettu [1].
- Jotkut lähteet voivat tuottaa enemmän dataa kuin toiset.

> **Esimerkki:** Valitaan satunnaisesti $N = 30$ ihmistä jostain ravintolasta. Olkoon x_i skalaari, joka kertoo henkilön i painon. Oletetaan, että mittaukset ovat peräisin kahdesta lähteestä - naisten ja miesten painojen jakaumista. Jakaumat voivat olla eri odotusarvoihin keskittyneitä normaalijakaumia ja sukupuolten esiintymisfrekvenssit voivat olla erilaiset.

Äärellinen sekoitemalli voidaan nyt määritellä seuraavasti (M on lähteiden määrä, Y on lähteen ilmaiseva satunnaismuuttuja):

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{M} p(\mathbf{x}|Y = j, \theta) p(Y = j|\theta). \tag{1}$$

Kaavassa määritellään satunnaisvektorin $\mathbf{X}$:n tiheys $p(\mathbf{x}|\theta)$ käyttämällä M:ää lähdettä. Vektori θ sisältää mallin kaikki parametrit, eli kaikkien lähteiden j ehdollisten jakaumien $p(\mathbf{x}|j)$ parametrit[2] , sekä lähteiden prioritodennäköisyyksien $p(j)$ parametrit [3].

[1] vastaa k-means -menetelmän oletusta siitä, että klusterien määrä on ennalta tunnettu.

[2] Tässä käytetään tavallisia lyhennysmerkintöjä, eli $p(\mathbf{X} = \mathbf{x}|Y = j) = p(\mathbf{x}|j)$.

[3] Huomaa, että kussakin termissä ollaan oikeasti kiinnostuttu vain joistain vektorin θ komponenteista.

Esimerkki (jatkoa edelliseen): Olkoon $M = 2$ (miehet ja naiset), $1 = nainen$ ja $2 = mies$. Tarvitaan 6 parametriä $\theta = (\alpha_1, \alpha_2, \mu_1, \sigma_1, \mu_2, \sigma_2)$, missä parametrit α_j kertovat suoraan lähteiden prioritodennäköisyydet ja loput parametrit ovat normaalijakaumiksi oletettujen lähteiden parametrit (odotusarvot ja hajonnat). Nyt voidaan kirjoittaa:

$$\begin{aligned} p(x|\alpha_1, \alpha_2, \mu_1, \sigma_1, \mu_2, \sigma_2) &= p(x|nainen, \mu_1, \sigma_1)\alpha_1 + p(x|mies, \mu_2, \sigma_2)\alpha_2 \\ &= f(x; \mu_1, \sigma_1)\alpha_1 + f(x; \mu_2, \sigma_2)\alpha_2. \end{aligned}$$

Viimeisellä rivillä $f(x; \mu, \sigma)$ on normaalijakauman $N(\mu, \sigma^2)$ tiheys pisteessä x.

Yleisessä tapauksessa valmista mallia voidaan käyttää mm. klusterointiin. Lähteet voidaan tulkita klustereiksi ja vektorin $\mathbf{x}$ voidaan katsoa kuuluvan todennäköisimpään klusteriin, eli siihen klusteriin j, jolle

$$p(j|\mathbf{x}, \theta) = \frac{p(\mathbf{x}|j, \theta)p(j|\theta)}{p(\mathbf{x}|\theta)}. \tag{2}$$

on suurin [4].

1.2 Suurimman uskottavuuden periaate

ML-periaatetta (Maximum Likelihood -periaate, suurimman uskottavuuden periaate) voidaan luonnehtia seuraavasti: Kun käsillä on dataa, jonka oletetaan olevan peräisin tietystä parametroidusta jakaumasta, valitaan parametrit siten, että käsillä olevan datan todennäköisyys maksimoituu. Tämä vastaa sitä intuitiota, että saimme käsiimme "luultavasti" todennäköistä dataa emmekä epätodennäköistä dataa [5].

Olkoon $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ riippumaton otos [6]. Riippumattoman otoksen todennäköisyys on:

$$p(\mathcal{X}|\theta) = \prod_{i=1}^{N} p(\mathbf{x}_i|\theta) = L(\mathcal{X}|\theta). \tag{3}$$

Merkintä $L(\mathcal{X}|\theta)$ tarkoittaa otoksen uskottavuutta (likelihood) ja sen voi tulkita parametrivektorin θ funktioksi otoksen $\mathcal{X}$ pysyessä vakiona. Vektorin θ suurimman uskottavuuden estimaatti on sellainen arvovektori $\theta^\star$, joka maksimoi otoksen uskottavuuden:

$$\theta^\star = \arg\max_\theta L(\mathcal{X}|\theta). \tag{4}$$

Jos $p(\mathbf{x}_i|\theta)$ on funktionaaliselta muodoltaan yksinkertainen, parametrien θ suurimman uskottavuuden estimaattoreille voidaan usein johtaa yksinkertaisia ratkaisuja.

Esimerkki: Johdetaan tavallisen normaalijakauman suurimman uskottavuuden estimaattorit parametreille μ ja σ. Merkitään normaalijakauman tiheyttä f:llä:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

[4]Olisi myös mahdollista sanoa, että $\mathbf{x}$ ei kuulu mihinkään klusteriin, jos todennäköisyys olisi kaikilla j "liian pieni". Tällöin laskut olisivat hieman raskaampia kun termille $p(\mathbf{x}|\theta)$ olisi laskettava arvo.

[5]Huomaa, että jos malliperhe olisi erittäin "joustava", niin saatettaisiin päästä tilanteeseen, jossa havaitun datan todennäköisyys lähestyisi aina yhtä. Jokaista datajoukkoa voitaisiin pitää "väistämättömänä". Tällaista ylisovittamista tulisi välttää käyttämällä mahdollisimman yksinkertaista "riittävän hyvää" malliperhettä.

[6]Kirjallisuudessa X tarkoittaa usein satunnaismuuttujaa ja x sen saamaa arvoa. Kannattaa huomata, että nyt $\mathcal{X}$ ei tarkoita kumpaakaan.

Koska uskottavuus $L(\mathcal{X}|\theta)$ sisältää tulon, on helpompaa maksimoida:

$$\log L(\mathcal{X}|\theta) = \sum_{i=1}^{N} \log p(\mathbf{x}_i|\theta) = \sum_{i=1}^{N} \log f(x_i).$$

Nyt $\log L(\mathcal{X}|\theta)$ ääriarvokohdat löytyvät helposti derivoimalla:

$$\frac{\partial}{\partial \mu} \log f(x) = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \mu}(x^2 - 2x\mu + \mu^2) = \frac{x-\mu}{\sigma^2},$$

joten

$$\frac{\partial}{\partial \mu} \log L(\mathcal{X}|\theta) = \sum_{i=1}^{N} \frac{\partial}{\partial \mu} \log f(x_i) = \sum_{i=1}^{N} \frac{x_i - \mu}{\sigma^2}.$$

Asettamalla $\frac{\partial}{\partial \mu} \log L(\mathcal{X}|\theta) = 0$, saadaan:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Vastaavasti:

$$\frac{\partial}{\partial \sigma} \log f(x) = \frac{\partial}{\partial \sigma} \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2}(x-\mu)^2 \frac{\partial}{\partial \sigma}\frac{1}{\sigma^2} = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3},$$

jolloin

$$\frac{\partial}{\partial \sigma} \log L(\mathcal{X}|\theta) = \sum_{i=1}^{N} \frac{\partial}{\partial \sigma} \log f(x_i) = \sum_{i=1}^{N}(-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}) = 0.$$

Viimeisestä muodosta saadaan ratkaisu:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

Edellisestä esimerkistä voidaan huomata, että sekoitemallien tapauksessa

$$\log L(\mathcal{X}|\theta) = \sum_{i=1}^{N} \log p(\mathbf{x}_i|\theta) = \sum_{i=1}^{N} \log \sum_{j=1}^{M} p(\mathbf{x}_i|y_j,\theta)p(y_j|\theta),$$

jolloin logaritmien käyttö ei näytä erityisemmin helpottavan uskottavuuden derivointia esim. multinormaalijakaumien tapauksessa.

2 Uskottavuuden maksimointi EM-algoritmilla

Edellisen luvun lopussa nähtiin, että sekoitemallien tapauksessa uskottavuusfunktion suoraviivainen maksimointi voi vaikuttaa hankalalta. Tässä luvussa esitellään ns. Expectation Maximization -algoritmi, jolla voidaan löytää uskottavuuden lokaaleja maksimikohtia. Algoritmi on iteratiivinen, eli se tuottaa jonon parametrivektoreita $(\theta^1, \ldots, \theta^T)$ siten, että parametrien uskottavuus toivottavasti paranee iteraatiokertojen myötä. EM-algoritmi on jo vanha ja hyvin tunnettu menetelmä (alkup. [3]), joten kiinnostuneen lukijan on helppo löytää asiasta oppikirjoja (esim. [2, 4]).

2.1 EM-perusteet

Palautetaan mieleen yllä esitelty sekoitteen määritelmä (1):

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{M} p(\mathbf{x}|j,\theta)p(j|\theta).$$

Voidaan leikkiä seuraavalla ajatuksella: Jos datasta jotenkin etukäteen tiedettäisiin mihin joukkoon (klusteriin, luokkaan) j kukin $\mathbf{x}_i$ kuuluu, suurimman uskottavuuden parametrien estimoinnista tulisi huomattavasti helpompaa. Kun jakaumien funktionaaliset muodot olisi kiinnitetty, jokainen $p(\mathbf{x}|j,\theta)$ voitaisiin estimoida suoraan osaotoksesta $\{\mathbf{x}_i|\mathbf{x}_i \in R\{j\}\}$, missä x_i kuuluu j:nteen joukkoon joss $\mathbf{x}_i \in R\{j\}$. Joukkojen prioritodennäköisyyksien estimointi onnistuisi laskemalla luokkien suhteelliset frekvenssit. Ajatusleikissä täytyy olettaa, että kukin datavektori kuuluu "oikeasti" täsmälleen yhteen luokkaan. Oletus sopii sekoitemallien lähtöoletuksiin - oletettiin, että kukin datavektori oli peräisin tasan yhdestä "lähteestä".

EM-algoritmi perustuu tällaiseen ajatusleikkiin. Dataa periaatteessa täydennetään kuvitteellisilla luokkamuuttujilla. Koska kuvitteellisten luokkamuuttujien arvoja ei käytännössä tiedetä, luokkamuuttujia käsitellään satunnaismuuttujina. Koska otoksen uskottavuus riippuu näistä satunnaismuuttujista, uskottavuudestakin tulee satunnaismuuttuja. Uskottavuuden maksimoinnin sijaan joudutaankin maksimoimaan uskottavuuden odotusarvoa. Osoittautuu, että uskottavuuden odotusarvolle saadaan lauseke, joka riippuu sekä parametrien θ "oikeista" arvoista, että näiden parametrien aikaisemmista estimaateista. Lausekkeen perusteella voidaan rakentaa iteraatio, jossa parametrien aikaisempien estimaattien avulla pystytään laskemaan parametreille uudet estimaatit.

Tarkastellaan EM-algortimin toimintaperiaatetta hieman täsmällisemmin. Täydennetään havaintoja $\mathbf{x}_i$ niiden luokituksilla $y_i \in \{1,\ldots,M\}$:

$$\mathbf{z}_i = (\mathbf{x}_i, y_i). \tag{5}$$

Lisäksi oletetaan seuraavan tiheysfunktion olemassaolo:

$$p(\mathbf{z}|\theta) = p(\mathbf{x},y|\theta) = p(y|\mathbf{x},\theta)p(\mathbf{x}|\theta). \tag{6}$$

Määritelmän (5) perusteella voidaan määritellä "täydellinen" data:

$$\mathcal{Z} = (\mathcal{X},\mathcal{Y}) = \{(\mathbf{x}_i,y_i)|\mathbf{x}_i \in \mathcal{X}, \mathbf{x}_i \in R\{y_i\}\}. \tag{7}$$

Täydellisen otoksen uskottavuus määritellään seuraavasti:

$$L(\mathcal{Z}|\theta) = L(\mathcal{X},\mathcal{Y}|\theta) = p(\mathcal{X},\mathcal{Y}|\theta) = \prod_{i=1}^{N} p(\mathbf{x}_i,y_i|\theta) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i,\theta)p(\mathbf{x}_i|\theta). \tag{8}$$

Optimoitaessa uskottavuutta (8) havaittu otos $\mathcal{X}$ on tunnettu tekijä ja $\mathcal{Y}$ on tuntematon, satunnainen, tekijä. Oletetaan lisäksi, että otoksesta $\mathcal{X}$ on saatu hetkellä t irti edes jonkinlainen estimaatti θ^t (aluksi sokeasti arpomalla jos ei muuten).

Edellisen perusteella on periaatteessa olemassa suure $p(\mathcal{Y}|\mathcal{X},\theta^t)$, jonka perusteella voidaan laskea odotusarvoja jollekin funktiolle $h(\mathcal{Y})$ [7]:

$$E[h(\mathcal{Y})|\mathcal{X},\theta^t] = \int h(\mathcal{Y})p(\mathcal{Y}|\mathcal{X},\theta^t)d\mathcal{Y}. \tag{9}$$

[7]Huom: Kyseessä on tiheysfunktio suureelle $\mathcal{Y} = \{y_1,\ldots,y_N\}$, eikä millekään yksittäiselle Y. Tiheys on siis koko Y:n arvojen kokoelmalle.

Funktio $h(\mathcal{Y})$ voi aivan hyvin riippua myös suureesta $\mathcal{X}$ ja todellisista parametreista θ. Olkoon $h(\mathcal{Y}) = \log p(\mathcal{X}, \mathcal{Y}|\theta)$. Sijoittamalla h edelliseen yhtälöön saadaan:

$$E[\log p(\mathcal{X}, \mathcal{Y}|\theta)|\mathcal{X}, \theta^t] = \int \log p(\mathcal{X}, \mathcal{Y}|\theta) p(\mathcal{Y}|\mathcal{X}, \theta^t) d\mathcal{Y}. \tag{10}$$

Kaavoista (8,10) saadaan nyt muodostettua odotusarvo täydellisen otoksen log-uskottavuudelle s.e. odotusarvo lasketaan $\mathcal{Y}$:n suhteen käyttämällä apuna havaittua otosta $\mathcal{X}$ ja parametrien vanhoja estimaatteja θ^t. Edellisten yhtälöiden perusteella voidaan määritellä EM-algoritmin vaiheet:

1. Alusta $t = 0$ ja θ^t on satunnainen.
2. **E-askel.** Lasketaan täydellisen otoksen log-uskottavuuden odotusarvo:
$$Q(\theta, \theta^t) := E[\log p(\mathcal{X}, \mathcal{Y}|\theta)|\mathcal{X}, \theta^t] \tag{11}$$
3. **M-askel.** Maksimoidaan Q:
$$\theta^{t+1} := \arg\max_{\theta} Q(\theta, \theta^t). \tag{12}$$
4. Lopeta, jos $\theta^t \approx \theta^{t+1}$. Muutoin $t := t + 1$ ja jatka kohdasta 2.

Algoritmissa toistetaan E- ja M-askelia vuorotellen, kunnes algoritmi konvergoituu johonkin lokaaliin uskottavuuden maksimiin, eli kunnes iteraatio ei enää muuta parametrejä merkittävästi. Voidaan osoittaa, että iteraatiossa (epätäydellisen) todellisen otoksen log-uskottavuus ei voi pienentyä [3, 4]. Niin sanotussa yleistetyssä EM-menetelmässä (GEM, Generalized EM) M-askeleessa ei tarvitse maksimoida Q:ta - pieni parantaminen riittää [4].

2.2 EM multinormaalijakaumien sekoitteelle

Yllä esiteltiin EM-algoritmin teoriaa. Koska edellä liikuttiin yleisellä tasolla, esitetyn perusteella saattaa olla hankalaa hahmottaa EM-algoritmille konkreettista toteutusta esimerkiksi silloin kun haluttaisiin sovittaa dataan multinormaalijakaumien sekoite klusterointia varten.

Tässä johdetaan (esitystä [5] mukaillen) konkreettinen algoritmi mainittuun sovellukseen. Määritellään ensin malliperhe sovelluskohtaisemmin kuin kaavassa (1):

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{M} \alpha_j p_j(\mathbf{x}|\theta_j). \tag{13}$$

Kaavassa (13) prioritodennäköisyyksiä on yksinkertaistettu - ne ovat suoraan parametrejä. Ehdollisia tiheysfunktioita merkitään p_j:llä sen korostamiseksi, että sekoitteen eri lähteet j ovat eri tiheysfunktioita. Lisäksi kunkin ehdollisen tiheysfunktion p_j parametrejä on merkitty θ_j:llä sen korostamiseksi, että kussakin tiheysfunktiossa ollaan kiinnostuttu vain osasta parametrejä. Kaikki ehdolliset tiheysjakaumat ovat multinormaaleja $\theta_j = (\mu_j, C_j)$ [8]:

$$p_j(\mathbf{x}|\theta_j) = \frac{1}{(2\pi)^{d/2} \det(C_j)^{1/2}} \exp(-\frac{1}{2}(\mathbf{x} - \mu_j)^T C_j^{-1} (\mathbf{x} - \mu_j)). \tag{14}$$

Parametrivektori voidaan siis ilmaista muodossa: $\theta = (\alpha_1, \ldots, \alpha_M, \mu_1, \ldots, \mu_M, C_1, \ldots, C_M)$.

[8] odotusarvovektori ja kovarianssimatriisi

Täydellisen datan log-uskottavuus on (y_i jälleen piilomuuttuja, joka kertoo mistä jakaumasta $\mathbf{x}_i$ on peräisin):

$$\log L(\mathcal{X},\mathcal{Y}|\theta) = \log p(\mathcal{X},\mathcal{Y}|\theta) = \log \prod_{i=1}^{N} p(\mathbf{x}_i, y_i|\theta) = \sum_{i=1}^{N} \log p_{y_i}(\mathbf{x}_i|\theta_{y_i})\alpha_{y_i}. \tag{15}$$

Aiemmin (kaavassa 9) nähtiin, että E-vaihetta varten on oltava lauseke piilomuuttujien todennäköisyyksille ehdolla, että otoksen näkyvä osa tunnetaan ja on olemassa jokin arvaus parametrien arvoille. Pannaan merkille, että y_i:t eivät riipu toisistaan ja että kukin y_i riippuu vain $\mathbf{x}_i$:stä jolloin saadaan:

$$p(\mathcal{Y}|\mathcal{X},\theta^t) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i,\theta^t). \tag{16}$$

Tämän jälkeen tarvittavat todennäköisyydet $p(y_i|\mathbf{x}_i,\theta^t)$ saadaan Bayesin kaavalla:

$$p(y_i|\mathbf{x}_i,\theta^t) = \frac{\alpha_{y_i}^t p_{y_i}(\mathbf{x}_i|\theta_{y_i}^t)}{\sum_{k=1}^{M}\alpha_k^t p_k(\mathbf{x}_i|\theta_k^t)}, \tag{17}$$

missä kaikki tarvittavat $p_j(\mathbf{x}_i|\theta_j^t)$ saadaan suoraan kaavasta (14) sijoittamalla parametrien estimaatit.

Tämän jälkeen voidaan alkaa kirjoittamaan auki E-askeleen lauseketta (11,10):

$$Q(\theta,\theta^t) = E[\log p(\mathcal{X},\mathcal{Y}|\theta)|\mathcal{X},\theta^t] = \sum_{\mathcal{Y}} \log p(\mathcal{X},\mathcal{Y}|\theta)p(\mathcal{Y}|\mathcal{X},\theta^t). \tag{18}$$

Sijoittamalla (15,16) saadaan:

$$Q(\theta,\theta^t) = \sum_{\mathcal{Y}} \Big[\sum_{i=1}^{N} \log(p_{y_i}(\mathbf{x}_i|\theta_{y_i})\alpha_{y_i})\Big]\Big[\prod_{j=1}^{N} p(y_j|\mathbf{x}_j,\theta^t)\Big]. \tag{19}$$

Koska uloimmassa summassa käydään läpi kaikki $\mathcal{Y} = \{y_1,\ldots,y_N\}$, voidaan kirjoittaa:

$$Q(\theta,\theta^t) = \sum_{y_1=1}^{M}\cdots\sum_{y_N=1}^{M} \Big[\sum_{i=1}^{N} \log(p_{y_i}(\mathbf{x}_i|\theta_{y_i})\alpha_{y_i})\Big]\Big[\prod_{j=1}^{N} p(y_j|\mathbf{x}_j,\theta^t)\Big]. \tag{20}$$

Otetaan seuraavaksi käyttöön indikaattorimuuttuja I_{l,y_i}, joka saa arvon 1 joss $y_i = l$ (muutoin 0). Tarkoituksena on päästä eroon siitä, että yllä tuntematon muuttuja y_i määrää mitä jakaumaa p_{y_i} käytetään. Kaava (20) saadaan seuraavaan muotoon:

$$\begin{aligned} Q(\theta,\theta^t) &= \sum_{y_1=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{i=1}^{N}\sum_{l=1}^{M} I_{l,y_i}\log(p_l(\mathbf{x}_i|\theta_l)\alpha_l)\prod_{j=1}^{N} p(y_j|\mathbf{x}_j,\theta^t). \\ &= \sum_{l=1}^{M}\sum_{i=1}^{N}\log(p_l(\mathbf{x}_i|\theta_l)\alpha_l)\sum_{y_1=1}^{M}\cdots\sum_{y_N=1}^{M} I_{l,y_i}\prod_{j=1}^{N} p(y_j|\mathbf{x}_j,\theta^t) \end{aligned} \tag{21}$$

On turhaa käydä läpi sellaisia y_i:n arvoja, joilla indikaattori on nolla, joten pala

$$\sum_{y_1=1}^{M}\cdots\sum_{y_N=1}^{M} I_{l,y_i} = \sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M} 1,$$

jolloin

$$\sum_{y_1=1}^{M}\cdots\sum_{y_N=1}^{M} I_{l,y_i}\prod_{j=1}^{N} p(y_j|\mathbf{x}_j,\theta^t) = \sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1}^{N} p(y_j|\mathbf{x}_j,\theta^t)$$

$$= \Big(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1,j\neq i}^{N} p(y_j|\mathbf{x}_j,\theta^t)\Big)p(l|\mathbf{x}_i,\theta^t)$$

$$= \prod_{j=1,j\neq i}^{N}\Big(\sum_{y_j=1}^{M} p(y_j|\mathbf{x}_j,\theta^t)\Big)p(l|\mathbf{x}_i,\theta^t) = p(l|\mathbf{x}_i,\theta^t),$$

missä toiseksi viimeisessä askeleessa huomattiin, että kaikkien mahdollisten tulojen summa voidaan muodostaa myös kirjoittamalla auki kaikkien summien tulot. Viimeisessä askeleessa huomattiin, että todennäköisyyksien summa on yksi.

Tämän jälkeen (21) saadaan vihdoin kirjoitettua järkevään muotoon (muista, että $p(l|\mathbf{x}_i,\theta^t)$ on lyhennysmerkintä sille, että $p(Y_i = l|\mathbf{X}_i = \mathbf{x}_i,\theta^t)$):

$$Q(\theta,\theta^t) = \sum_{l=1}^{M}\sum_{i=1}^{N}\log(p_l(\mathbf{x}_i|\theta_l)\alpha_l)p(l|\mathbf{x}_i,\theta^t)$$

$$= \sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l|\mathbf{x}_i,\theta^t) + \sum_{l=1}^{M}\sum_{i=1}^{N}\log(p_l(\mathbf{x}_i|\theta_l))p(l|\mathbf{x}_i,\theta^t). \tag{22}$$

Lausekkeessa (22) vasen ja oikea puoli voidaan optimoida erikseen, koska α-termien ja θ_l-termien välillä ei ole vuorovaikutusta.

Yllä α-termien optimoinnista saadaan aikaan rajoitettu optimointiongelma (rajoitteena $\sum_l \alpha_l = 1$). Ongelman voi ratkaista ottamalla rajoitteen osalta mukaan sopivan Lagrange-sakkotermin $\lambda(\sum_l \alpha_l - 1)$ ja derivoimalla:

$$\frac{\partial}{\partial\alpha_k}\Big[\sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l|\mathbf{x}_i,\theta^t) + \lambda(\sum_{l}\alpha_l - 1)\Big] = 0. \tag{23}$$

$$\Leftrightarrow \sum_{i=1}^{N}\frac{1}{\alpha_k}p(k|\mathbf{x}_i,\theta^t) + \lambda = 0 \Leftrightarrow \sum_{i=1}^{N} p(k|\mathbf{x}_i,\theta^t) = -\lambda\alpha_k$$

$$\Leftrightarrow \sum_{i=1}^{N}\sum_{k=1}^{M} p(k|\mathbf{x}_i,\theta^t) = -\lambda\sum_{k=1}^{M}\alpha_k \Leftrightarrow N = -\lambda,$$

jolloin lausekkeen (23) derivaattaan voidaan sijoittaa $\lambda = -N$ ja saadaan:

$$\alpha_k = \frac{1}{N}\sum_{i=1}^{N} p(k|\mathbf{x}_i,\theta^t). \tag{24}$$

Vastaavalla tavalla voidaan derivoida lausekkeen (22) oikea puoli parametrien μ_k ja C_k suhteen. Tällöin ei tarvitse käyttää apuna Lagrange-termejä, koska käsiteltävillä optimointiongelmilla ei ole rajoitteita. Koska laskut ovat työläitä ja yllätyksettömiä, niitä ei käydä tässä läpi (kts. [5]). Saadaan seuraavat päivityskaavat:

$$\alpha_k^{new} = \frac{1}{N}\sum_{i=1}^{N} p(k|\mathbf{x}_i,\theta^t), \tag{25}$$

$$\mu_k^{new} = \frac{\sum_{i=1}^N \mathbf{x}_i p(k|\mathbf{x}_i, \theta^t)}{\sum_{i=1}^N p(k|\mathbf{x}_i, \theta^t)}, \tag{26}$$

$$C_k^{new} = \frac{\sum_{i=1}^N p(k|\mathbf{x}_i, \theta^t)(\mathbf{x}_i - \mu_k^{new})(\mathbf{x}_i - \mu_k^{new})^T}{\sum_{i=1}^N p(k|\mathbf{x}_i, \theta^t)}. \tag{27}$$

Pannaan merkille, että kaavoissa (25,26,27) suoritetaan E- ja M-askeleet rinnakkain. Todennäköisyydet $p(k|\mathbf{x}_i, \theta^t)$ saadaan kaavasta (17). Saadut kaavat voidaan tulkita seuraavasti:

- Kaavassa (25) uudet priorit saadaan laskemalla datan yli keskimääräinen todennäköisyys kuulua luokkaan k.
- Kaavassa (26) luokan k uusi luokkakeskiarvo μ_k on painotettu keskiarvo kaikista datavektoreista $\mathbf{x}_i$ s.e. $\mathbf{x}_i$:n paino on suuri silloin kun luokkaan kuuluminen on todennäköistä. Voidaan panna merkille, että jos luokkatodennäköisyydet olisivat aina joko ykkösiä tai nollia, niin kaava pelkistyisi $k - means$ -menetelmän päivityskaavaksi (luokkaan k kuuluvien $\mathbf{x}_i$ summa jaettuna luokan koolla).
- Kaavassa (27) päivitetään luokkien kovarianssimatriisien estimaatit. Lasku eroaa tavallisesta kovarianssiestimaatista siinä, että osia painotetaan jälleen luokkatodennäköisyyksillä, eli luokkaan todennäköisesti kuuluvat $\mathbf{x}_i$ painavat estimaatissa enemmän.

Koska kaava (26) muistuttaa tietyissä olosuhteissa $k - means$ -menetelmän päivitysaskelta, pieni vertailu on paikallaan. Klusterointikäytössä yo. multinormaalien sekoitteen sovitus eroaa $k - means$ -menetelmästä muun muassa siinä, että etäisyyksiä $||\mathbf{x} - \mu_j||_2$ skaalataan eksponenttifunktion ja kovarianssimatriisin käänteismatriisin C_j^{-1} avulla (14). Koska C_j muuttuvat iteraatiossa, "etäisyydet" muuttuvat tarpeen mukaan, eli ne riippuvat C_j:n määräämästä jakauman j leveydestä eri suuntiin kullakin ajan hetkellä. Kovarianssimatriisien erilaisuudesta eri luokille seuraa mm. se, että datavektori $\mathbf{x}$ ei välttämättä kuulu siihen klusteriin, jonka keskus on $\mathbf{x}$:ää lähinnä (tavallisella euklidisella normilla mitattuna).

3 EM -algoritmin ongelmia

Edellisissä luvuissa muodostettiin jonkinlainen käsitys EM-menetelmän teoriasta ja siitä, miten menetelmää voidaan käyttää esim. sovittamaan multinormaalijakaumien sekoite dataan jolloin syntyvää mallia voidaan käyttää klusterointiin (valmiista mallista katsotaan mikä on vektorin todennäköisin luokka). EM-iteraatiossa on kuitenkin seuraavia ongelmia [6]:

- Konvergoitumiseen tarvittava aika voi olla suuri jos data on korkeaulotteista ja sitä on paljon. *Esimerkki:* Päivityskaavassa 27 lasketaan matriisien painotettu summa. Matriiseja on yhtä paljon kuin dataa ja matriisien koko riippuu datan ulottuvuuksien määrästä).
- Konvergoituminen tapahtuu uskottavuuden lokaaliin maksimiin. Jos malliperheenä on esim. multinormaalijakaumien sekoite korkeaulotteisessa avaruudessa, huonoja lokaaleja maksimikohtia on luultavasti paljon (uskottavuus parametrien funktiona ei ole konveksi funktio ja on luultavasti pahasti "piikikäs"). Tästä johtuen EM on herkkä parametrien alkuarvauksien suhteen.
- Luokkien (lähteiden, klustereiden) määrää ei useinkaan voi arvata etukäteen. Tällöin algoritmia voidaan ajaa eri määrillä ja katsoa milloin saadaan parhaat tulokset. Tämä on erityisen ikävää jos yksittäisten ajojen konvergoituminen on hidasta.

- Ainakaan multinormaalijakaumien sekoite ei ole kovin robusti malliperhe silloin kun otoksessa on poikkeavia havaintoja (outlier), koska multinormaalijakaumien "hännät" ovat melko kapeita. *Esimerkki:* Luokan normaalijakauman varianssia voidaan kasvattaa liikaa, jos luokkaan otetaan mukaan kaukaisia pisteitä kun suuri osa pisteistä on kuitenkin lähellä odotusarvovektoria.

Alla kuvataan lyhyesti kirjallisuudessa esiintyneitä korjausehdotuksia muutamaan yllä mainituista ongelmista.

3.1 Konvergenssin nopeuttaminen

Konvergenssiä voidaan nopeuttaa suorittamalla E- ja M-askeleet "osittain". M-askeleen osittaista suorittamista käytetään mm. yleistetyssä EM-algoritmissa (GEM), jossa M-askel ei maksimoi uskottavuutta, vaan pelkästään kasvattaa sitä hiukan.

Jos piilomuuttujat ovat toisistaan riippumattomia (kuten aikaisemmin esitetyssä EM-menetelmässä multinormaalijakaumien sekoitteelle), ne vaikuttavat uskottavuuteen siinä mielessä yksinkertaisella tavalla, että voidaan menetellä seuraavasti: E-askeleessa ei johdeta tuntemattomalle datalle jakaumaa $p(\mathcal{Y}|\mathcal{X}, \theta^t)$ (16) joka kierroksella, vaan rajoitutaan tietyn $p(Y_i|\mathcal{X}, \theta^t)$ päivitykseen kierroksen aikana (muut $p(Y_j|\mathcal{X}, \theta^t), j \neq i$ jätetään ennalleen). Tämän jälkeen voidaan välittömästi suorittaa M-askeleen korvike.

> **Esimerkki:** Muuttujat Y_i ovat multinormaalijakaumien sekoitteelle tarkoitetussa EM-algoritmissa toisistaan riippumattomia. Kun valitaan i ja muutetaan yhtä jakaumaa $p(Y_i = k|\mathbf{X}_i = \mathbf{x}_i, \theta^t)$ ja muut jätetään ennalleen, niin päivityskaavoista (25,26,27) huomataan, että tarvittavia summia ei tarvitse laskea kokonaan uudelleen. Riittää, että vanhoista summista vähennetään termi ja lisätään uusi tilalle. Kun toimitaan näin, tullaan suorittaneeksi osittaisia E- ja M-askeleita.

Osittainen askellus saattaa joskus nopeuttaa konvergenssiä [7]. Osittaiseen askellukseen perustuvia menetelmiä voidaan perustella teoreettisesti muotoilemalla E- ja M-askeleet hieman toisella tavalla ja optimoimalla uskottavuuden sijaan ns. "negatiivista vapaata energiaa" [7].

Perustelujen monimutkaisuuden takia tässä riittänee todeta, että "negatiivinen vapaa energia" on tässä yhteydessä täydellisen datan log-uskottavuuden odotusarvo (kuten aiemmin) plus $p(\mathcal{Y}|\mathcal{X}, \theta)$:n entropia. On mahdollista todistaa, että negatiivisen vapaan energian lokaalit maksimit ovat myös tavallisen täydellisen datan log-uskottavuuden odotusarvon lokaaleja maksimeja [7].

3.2 Konvergoituminen lokaaliin maksimiin

Artikkelissa [6] esitellään simuloituun jäähdytykseen perustuva EM-klusteroinnin variaatio, jossa pyritään ottamaan huomioon kaikki yllä mainitut perus- EM:n ongelmat. Lähestymistapa perustuu artikkelissa [7] esiteltyyn "negatiiviseen vapaaseen energiaan" (H on jakauman entropia):

$$F(\theta^1, \theta^2, \beta) = \beta E[\log p(\mathcal{X}, \mathcal{Y}|\theta^1)|\mathcal{X}, \theta^2] + H(p(\mathcal{Y}|\mathcal{X}, \theta^2)). \tag{28}$$

Energiassa (28) esiintyvä β on [6]:ssä tehty lisäys. Simuloitu jäähdytys tapahtuu siten, että aluksi β on hieman nollaa suurempi ja kasvaa hieman jokaisen E- ja M-askelparin jälkeen [9]. Intuitiivisesti ajatellen pieni β kasvattaa entropian painoa log-uskottavuuteen nähden.

[9] Energialla iteraatiot suoritetaan hieman eri tavalla: E-askeleessa θ^1 on kiinnitetty jolloin optimoidaan jakaumaa $p(\mathcal{Y}|\mathcal{X}, \theta^2)$ ja M-askeleessa jakauma on kiinnitetty ja optimoidaan θ^1:htä.

Koska entropia on aina konveksi funktio, entropian suuri painoarvo säätää energiafunktiota “konveksimpaan suuntaan”. Kun $\beta = 1$, kyseessä on tavallinen EM [6, 7]. Intuitiivisesti ajatellen jäähdytyksellä voidaan siis pienentää lokaaliin maksimiin juuttumisen todennäköisyyttä laskennan alkuvaiheessa.

4 Yhteenveto

Yllä esitettiin johdatus EM-algoritmin teoriaan, sekä johdettiin toteutus multinormaalijakaumien sekoitteelle. EM-algoritmin käytännön ongelmia tarkasteltiin lyhyesti ja muutaman ongelman osalta esiteltiin valittuja paloja kirjallisuudessa esiintyneistä ratkaisuehdotuksista. Kaksi ongelmaa jätettiin käsittelemättä seuraavista syistä:

- Luokkien määrän ennalta arvaamisen ongelma ei ole erityisesti EM-algoritmin ongelma.
- Multinormaalijakaumien robustisuuden puute poikkeavien havaintojen (outlier) tapauksessa voidaan nähdä malliperheen valintaan liittyvänä sovelluskohtaisena ongelmana. Artikkelissa [6] esitetään normaalijakaumien korvaamista t-jakaumilla, mutta yleisessä tapauksessa voidaan tuskin väittää, että mitä tahansa dataa voidaan aina mallintaa paremmin t-jakaumien sekoitteella kuin normaalijakaumien sekoitteella.

Jos EM-algoritmilla sovitettua mallia aiotaan käyttää klusterointiin, niin voidaan panna merkille seuraavat asiat (artikkelin [8] terminologiaa mukaillen):

- Klusteroinnissa käytettävä *induktioperiaate* (nyt suurimman uskottavuuden periaate) on avoimesti esillä kaavoissa. Myös malliperheen muoto, sekä tehtävät oletukset näkyvät eksplisiittisesti kaavoissa.
- Induktioperiaatteen käyttöön liittyvää optimointiongelmaa voidaan yrittää ratkaista EM-algoritmilla.

Malliperheen ja oletusten eksplisiittinen esittäminen on siinä mielessä terveellistä, että valistunut käyttäjä voi yrittää valita malliperheen klusterointiongelmaan liittyvän taustatietämyksen perusteella [10]. Lisäksi käyttäjä voi säästyä turhalta työltä silloin kun datasta voi “suoraan nähdä”, että tietyn muotoisten mallien sovittamista ei edes kannata yrittää.

Klusterointikäytössä jää kuitenkin arveluttamaan seuraava asia: Esitelty klusterointitapa vaatii datan todennäköisyysjakauman eksplisiittistä estimointia. On tunnettu tosiasia, että korkeaulotteisissa avaruuksissa jakauman estimointi vaatii suuren otoksen (erityisesti jos sovitettava malli on joustava). Voisi kysyä onko todennäköisyysjakauman eksplisiittinen sovittaminen aidosti klusterointiin kuuluva alitehtävä, vai ratkaistaanko tässä helpompi ongelma vaikeamman ongelman ratkaisun avulla. Tarvitaanko esimerkiksi klusterien välisten rajojen löytämiseen välttämättä globaalisti pätevää tiheysestimaattia.

Viitteet

[1] J.Li, A.Barron. Mixture density estimation. Advances in neural information systems 12, The MIT Press, 2000.

[2] G.McLachlan, D.Peel. Finite mixture models. Wiley, 2000.

[10] Toki esim. k-means menetelmässäkin voi epäsuorasti käyttää taustatietämystä mm. normin valinnassa tai datan kuvaamisessa suotuisempaan piirreavaruuteen.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(B), 1977.

[4] G.J. McLachlan, T. Krishnan. The EM algorithm and extensions. Wiley, 1997.

[5] J.A.Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. TR-97-021, U.C.Berkeley, USA, 1998.

[6] Shy Shoham. Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. *Pattern Recognition*, 35(5):1127-1142, May 2002.

[7] R.M.Neal, G.E.Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. Teoksessa M.I.Jordan(Ed.), *Learning in Graphical Models*, Kluwer, 1998.

[8] Vladimir Estivill-Castro. Why so many clustering algorithms - a position paper. *SIGKDD Explorations*, 4(1):65-75, June 2002.