Mallipohjainen klusterointi

Marko Salmenkivi

Johdatus koneoppimiseen, syksy 2008

Luentorunko perjantaille 5.12.2008

Johdattelua mallipohjaiseen klusterointiin, erityisesti gaussisiin sekoitemalleihin

- Uskottavuusfunktio ja suurimman uskottavuuden estimaatti
- Sekoitemallit
- Sekoitemallin parametrien estimointi EM-algoritmilla

Mallipohjainen klusterointi

- mallipohjainen klusterointi (probabilistic model-based clustering): datavektorien ajatellaan olevan peräisin K:sta eri lähteestä
- lähteet = klusterit
- jokaiselle klusterille tilastollinen malli (todennäköisyysjakauma), josta klusteriin kuuluvat datavektorit ovat otoksia
- lähteen ilmaiseva luokkamuuttuja on piilomuuttuja (ts. sitä ei voida suoraan havaita)
- luokkamuuttujia voidaan mallintaa puuttuvana datana
- tehtävänä on ratkaista kunkin datavektorin kohdalla, mistä lähteestä se on peräisin

Parametrinen ja ei-parametrinen estimointi

- generatiivinen lähestymistapa: mallinnetaan aineistoa todennäköisyysjakaumalla, josta aineisto on peräisin
- pyritään aineiston perusteella tekemään päätelmiä jakaumasta

1. parametriset mallit
   - kiinnitetään etukäteen generoivan jakauman funktionaalinen muoto (esim. normaalijakauma)
   - estimoidaan aineiston perusteella jakauman parametreja (esim. 1-ulotteisen normaalijakauman tapauksessa odotusarvo ja varianssi, multinormaalijakauman tapauksessa odotusarvo ja kovarianssimatriisi)
2. ei-parametriset mallit
   - vältetään jakaumaan liittyvien oletusten tekemistä
   - pyritään estimoimaan generoivaa jakaumaa "suoraan" aineiston perusteella

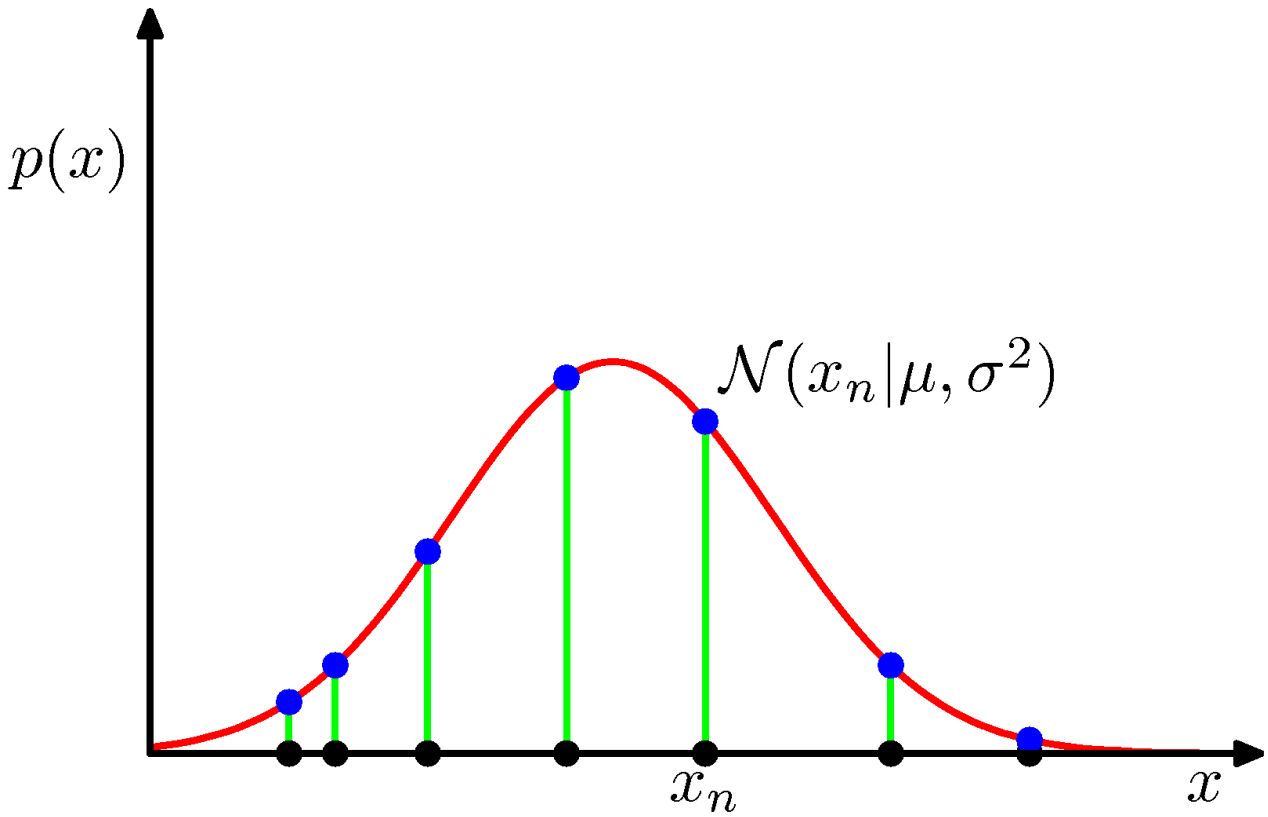
Uskottavuusfunktio

- kun aineistosta estimoidaan jakauman parametreja, yleisimmin käytetty kustannusfunktio ratkaisun hyvyydelle on uskottavuusfunktio (likelihood function)
- esimerkki: parametriseksi malliksi on valittu 1-ulotteinen normaalijakauma, josta havainnot $x_1, \ldots, x_n$ oletetaan riippumattomiksi otoksiksi

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right\}$$

- kun funktiota tarkastellaan siten, että parametrit ovat muuttujia ja aineisto kiinnitetty, sitä nimitetään uskottavuusfunktioksi
- funktion arvon maksimoivia parametrien arvoja nimitetään (aineiston) suurimman uskottavuuden estimaateiksi (maximum likelihood estimates)

Esimerkki

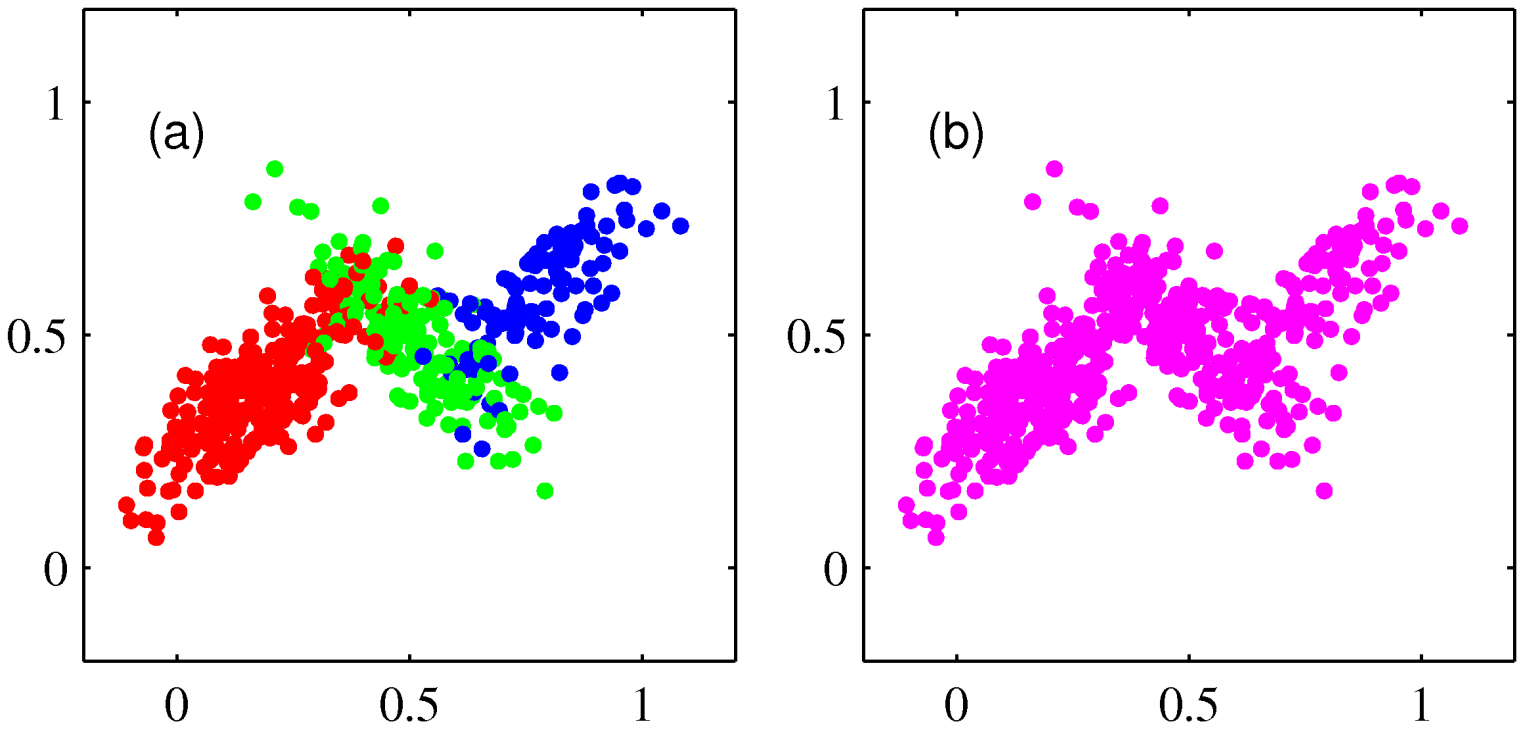


- datapisteiden tn-tiheydet palkkien korkeudet
- suurimman uskottavuuden estimointi: löydettävä sellainen normaalijakauma (ts. sellaiset μ ja σ^2), että tiheyksien tulo maksimoituu

Sekoitemallit

- sekoitemallien tapauksessa oletetaan, että aineisto on peräisin K:sta parametrisesta jakaumasta
- esimerkiksi K:sta (multi)normaalijakaumasta, joilla on erilaiset parametrit (odotusarvo, kovarianssimatriisi)
- mutta emme tiedä jakaumien parametreja emmekä sitä, mikä havainto on peräisin mistäkin jakaumasta!

Esimerkki



Sekoitemallit

- tarkastellaan mielivaltaista datavektoria $\mathbf{x}$; oletuksemme on, että se on peräisin yhdestä lähteestä
- otetaan käyttöön lähteen ilmaiseva muuttuja Cl (arvoalue kokonaisluvut $1, \ldots, K$)
- voimme ilmaista yhteistodennäköisyyden $p(\mathbf{x}, Cl) = p(\mathbf{x}|Cl)p(Cl)$
- siis esim. todennäköisyys, että $\mathbf{x}$ on peräisin lähteestä $Cl = k$, on $p(\mathbf{x}|Cl = k)p(Cl = k)$
- saamme (marginaali)todennäköisyyden $\mathbf{x}$:lle summaamalla eri lähteistä peräisin olemisen todennäköisyydet

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|Cl = k)p(Cl = k)$$

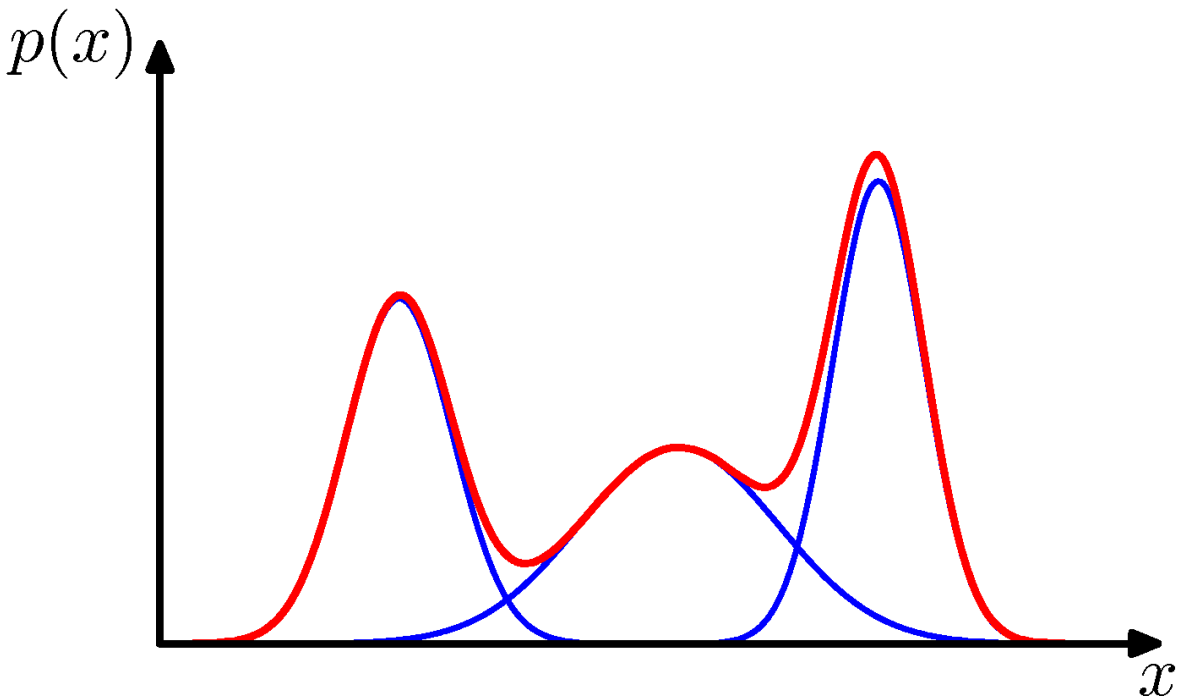
Sekoitemallit

- jakaumia $p(\mathbf{x}|Cl)$ mallinnetaan jollakin parametrisella jakaumalla, kuten multinormaalijakaumalla
- olkoon parametrinen jakauma k:nnelle lähteelle $f_k(\mathbf{x}; \theta_k)$, missä θ_k on jakauman parametrien vektori
- tällaista komponenteista koostuvaa mallia nimitetään äärelliseksi sekoitemalliksi (*finite mixture model*); se on muotoa

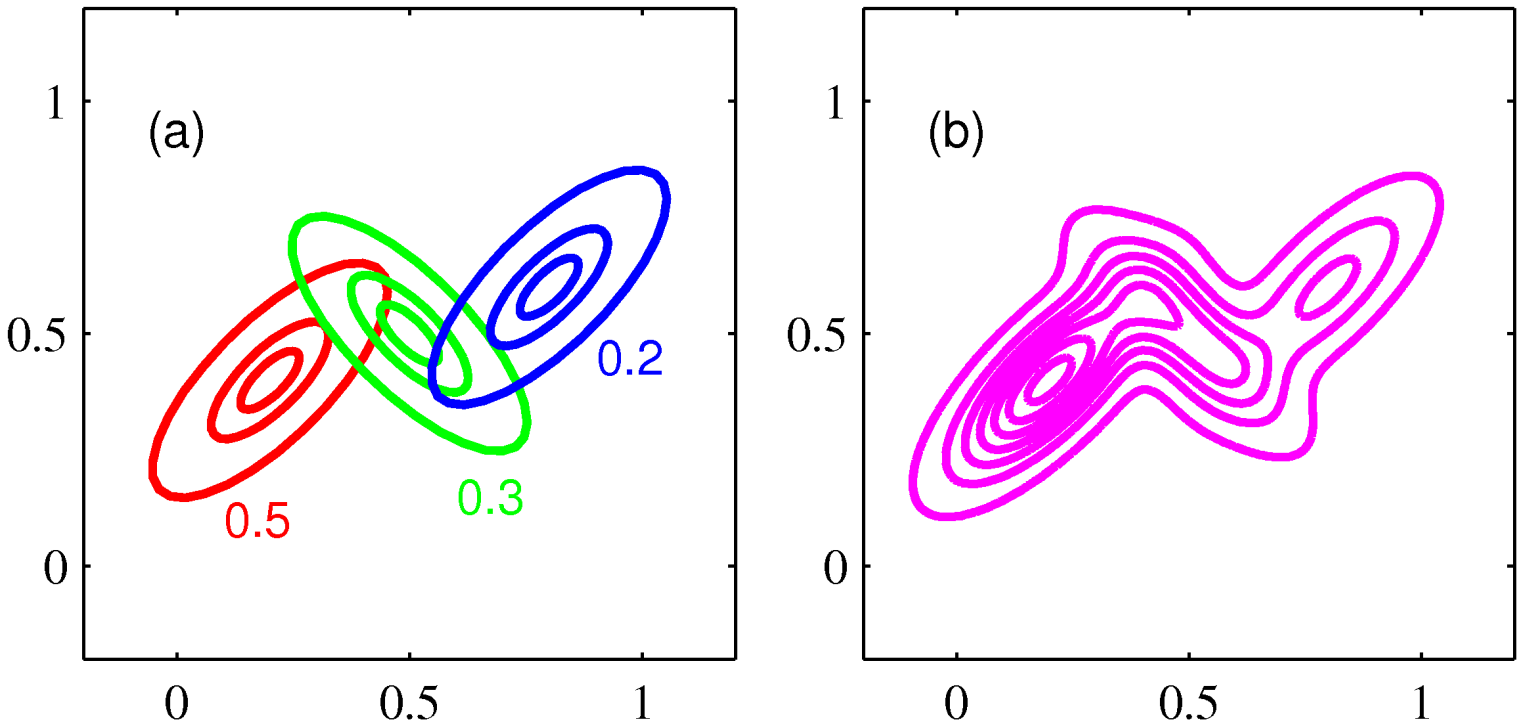
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}, \theta_k)$$

  - $f_1, \ldots, f_K$ ovat komponenttijakaumat
  - $\pi_1, \ldots, \pi_K$ ovat sekoitussuhteet tai painot; π_k ilmaisee todennäköisyyden, että komponentti k on tuottanut *satunnaisesti valitun datavektorin*
  - mallin parametrit Θ: komponenttijakaumien parametrivektorit ja sekoitussuhteet $\pi_1, \ldots, \pi_K$

Esimerkki, 1-d



Esimerkki, 2-d



Klusterointi

- oletetaan (vain hetkeksi) että tunnemme komponenttijakaumien parametrien arvot ja komponenttien sekoitussuhteet
- voimme arvioida havaintomme lähteen todennäköisyyttä soveltamalla Bayesin kaavaa

$$p(Cl|\mathbf{x}, \Theta) = \frac{p(\mathbf{x}|Cl, \Theta)p(Cl|\Theta)}{p(\mathbf{x}|\Theta)}$$

- tässä $p(\mathbf{x}|\Theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}; \theta_\mathbf{k})$
- $\mathbf{x}$:n lähteeksi voidaan veikata todennäköisintä vaihtoehtoa
- emme kuitenkaan tunne mallin parametreja
- aineistosta on estimoitava samanaikaisesti komponenttimallien parametreja ja luokkamuuttujia

Uskottavuuden maksimointi

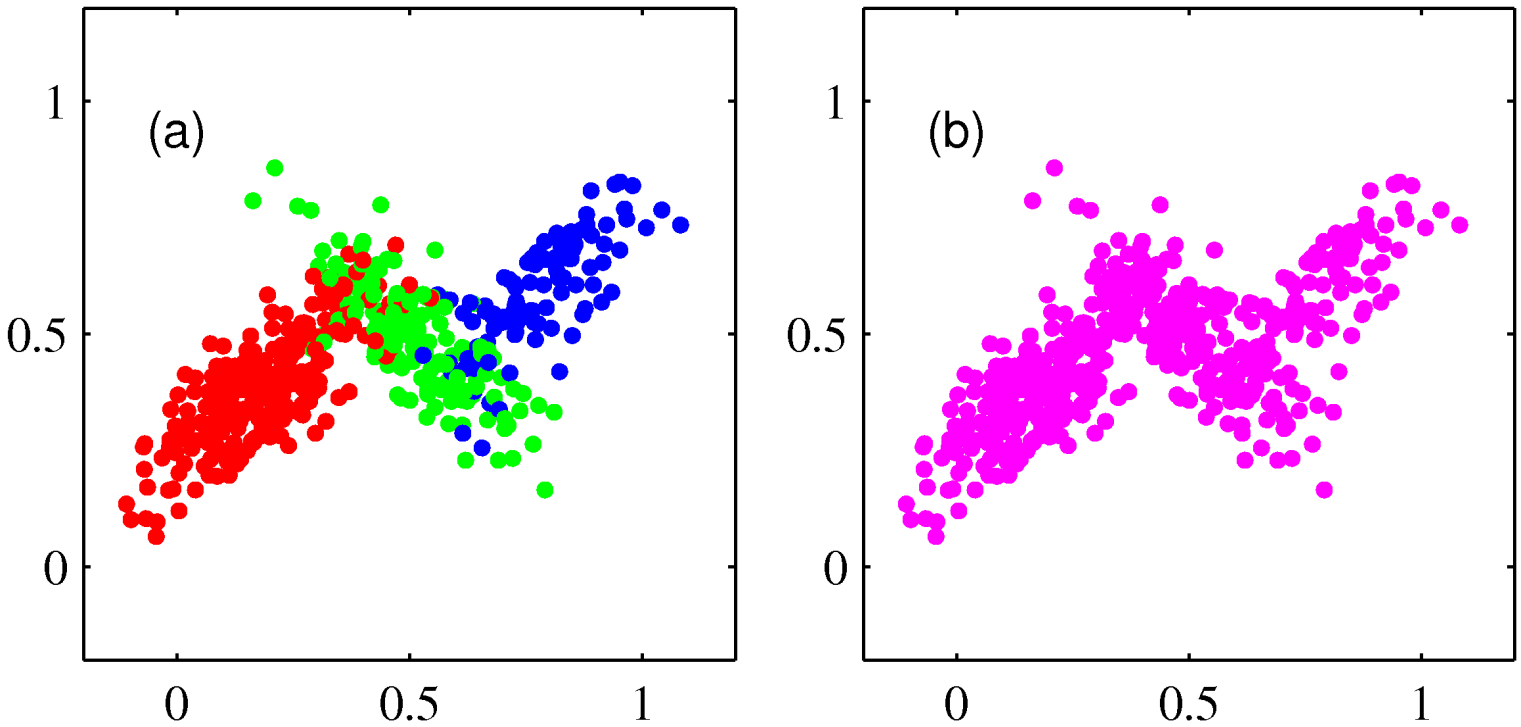
- kun malli sisältää vain yhden komponentin ja aineiston otokset ovat riippumattomia, log-uskottavuusfunktio on useiden parametristen mallien tapauksessa sellaista muotoa, että suurimman uskottavuuden estimaattien laskeminen voidaan tehdä suljetussa muodossa
- sekoitemallien sisältämä summatermi hankaloittaa huomattavasti log-uskottavuusfunktion maksimoimista

$$l(\Theta) = \ln(\sum_{Cl} p(\mathbf{x}_1, \ldots, \mathbf{x}_n, Cl | \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K))$$

Uskottavuuden maksimointi

- tehdään “ajatusleikki”: jos joku kertoisi, mistä lähteestä mikäkin datavektori on peräisin, summatermistä päästäisiin eroon
- aineisto voitaisiin jakaa K osaan ja kunkin komponenttijakauman $f_k(\mathbf{x}; \theta_k)$ parametrien suurimman uskottavuuden estimaatit voitaisiin etsiä erikseen kyseisestä lähteestä peräisin olevien datavektorien avulla
- (samalla klusterointikin olisi tullut hoidettua ja voitaisiin lopettaa tyytyväisinä ;-)

Esimerkki



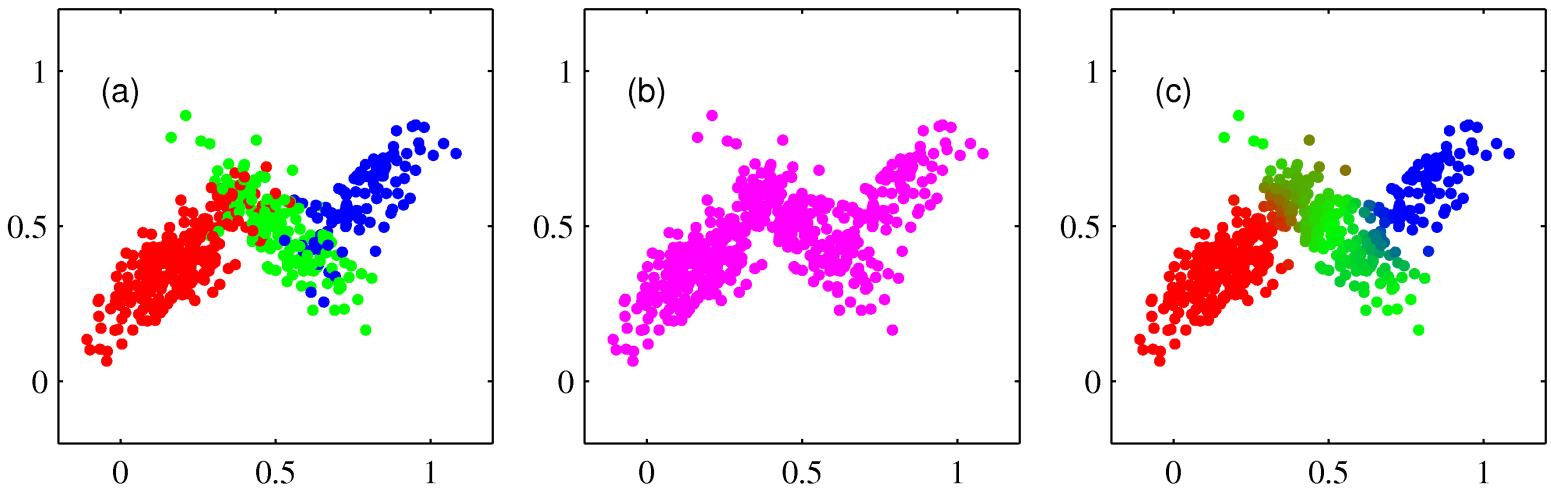
- “täydellinen” data ja “epätäydellinen” data

Uskottavuuden maksimointi

- todellisuudessa ei tietenkään ole tietoa luokkamuuttujien arvoista
- käytettävissä on ainoastaan aineistolla ja parametrivektorien kiinnitetyillä arvoilla ehdollistetut posterioritodennäköisyydet

$$p(Cl|\mathbf{x}, \Theta) = \frac{p(\mathbf{x}|Cl, \Theta)p(Cl|\Theta)}{p(\mathbf{x}|\Theta)}$$

Esimerkki



EM-algoritmi

- on leikitelty vuoroin ajatuksella, että
  - tiedämme luokkamuuttujien arvot, jolloin voisimme estimoida helpohkosti parametrien arvot
  - tiedämme parametrien arvot, jolloin voisimme arvioida kunkin datavektorin kohdalla luokkamuuttujan eri arvojen luultavuutta, ts. havainnon todennäköisintä lähdettä
- EM-algoritmi perustuu näiden molempien ajatusten vuorottelulle

EM-algoritmi

- alustetaan muuttujille $\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K$ alkuarvot
- Sen jälkeen iteroidaan seuraavia askelia:
  1. E-askel: estimoidaan piilomuuttujien ehdollista jakaumaa, kun parametrivektorin arvot on kiinnitetty
  2. M-askel: valitaan parametreille uudet arvot siten, että ne maksimoivat uskottavuuden, kun piilomuuttujien jakauma on kiinnitetty
- ...kunnes lopettamiskriteeri täyttyy (parametrien arvot eivät enää muutu tai log-uskottavuusfunktion arvo muuttuu enää "vähän")

E-askel

- hetki sitten todettiin, että jos $\Theta = (\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K)$ tunnetaan, niin datavektorin $\mathbf{x}$ kuhunkin klusteriin kuulumisen (posteriori)todennäköisyys saadaan Bayesin kaavasta

$$p(Cl|\mathbf{x}, \Theta) = \frac{p(\mathbf{x}|Cl, \Theta)p(Cl|\Theta)}{p(\mathbf{x}|\Theta)}$$

- tässä $p(\mathbf{x}|\Theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}, \theta_\mathbf{k})$
- tämä on E-askel

Esimerkki: E-askel yksiulotteisten normaalijakaumien sekoitteessa

- E-askel: estimoidaan luokkamuuttujien ehdolliset todennäköisyydet kullekin havainnolle x_i

$$\hat{P}(Cl = k|x_i, \mu, \sigma^2) = \frac{\pi_k f_k(x_i|\mu_k, \sigma_k)}{\sum_{i=1}^{K} \pi_i f_i(x_i|\mu_i, \sigma_i)}$$

M-askel yksiulotteisten normaalijakaumien sekoitteessa

maksimoidaan $\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K$

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{P}(Cl = k|x_i)$$

$$\hat{\mu}_k = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} x_i \hat{P}(Cl = k|x_i)$$

$$\hat{\sigma}_k = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} (x_i - \hat{\mu}_k)^2 \hat{P}(Cl = k|x_i)$$

EM-algoritmin toimivuudesta

- voidaan osoittaa, että EM-algoritmin askeleet eivät voi pienentää uskottavuusfunktion arvoa
- algoritmi konvergoi aina (vähintään) uskottavuuden lokaaliin maksimikohtaan
- valittu alkutila vaikuttaa tulokseen

Aikavaativuudesta

- multinormaalijakaumien sekoitteella EM-algoritmin aikavaativuutta dominoi kovarianssimatriisin parametrien estimointi
- K lähdettä, d dimensiota, $O(Kd^2)$ kovarianssiparametria
- kunkin kohdalla tarvitaan n havainnon läpikäymistä
- yhden askeleen aikavaativuus luokkaa $O(Kd^2n)$

Ongelmia

- konvergoituminen voi kestää kauan, jos paljon dataa ja/tai dimensioiden määrä on suuri
- korkeaulotteisissa avaruuksissa lokaaleja maksimikohtaia on yleensä paljon (uskottavuusfunktio ei ole konveksi ja on luultavasti 'piikikäs') $\longrightarrow$ herkkyys valitulle alkutilalle
- lähteiden määrää vaikea arvata etukäteen $\longrightarrow$ tarvitaan ajoja eri K:n arvoilla $\longrightarrow$ hidasta
- poikkeavien havaintojen vaikutus
- uskottavuuden maksimointi voi joskus johtaa ongelmiin: jos komponentin odotusarvoestimaatti osuu täsmälleen jonkin datapisteen kohdalle, uskottavuuden arvo voi kasvaa rajattomasti (uskottavuuden maksimoinnin sijasta voidaan siirtyä Bayes-päättelyyn ja maksimoida MAP (Maximum A Posteriori) -funktiota)

EM-algoritmi ja K:n keskipisteen klusterointi

- tarkastellaan seuraavanlaisten multinormaalimallien sekoitetta:

$$p(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi\epsilon}} \exp\left\{ -\frac{1}{2\epsilon}\|\mathbf{x} - \mu_k\|^2 \right\}$$

(kovarianssimatriisi on diagonaalimatriisi, jonka jokainen nollasta poikkeava arvo on kiinteä ϵ)

EM-algoritmi ja K:n keskipisteen klusterointi

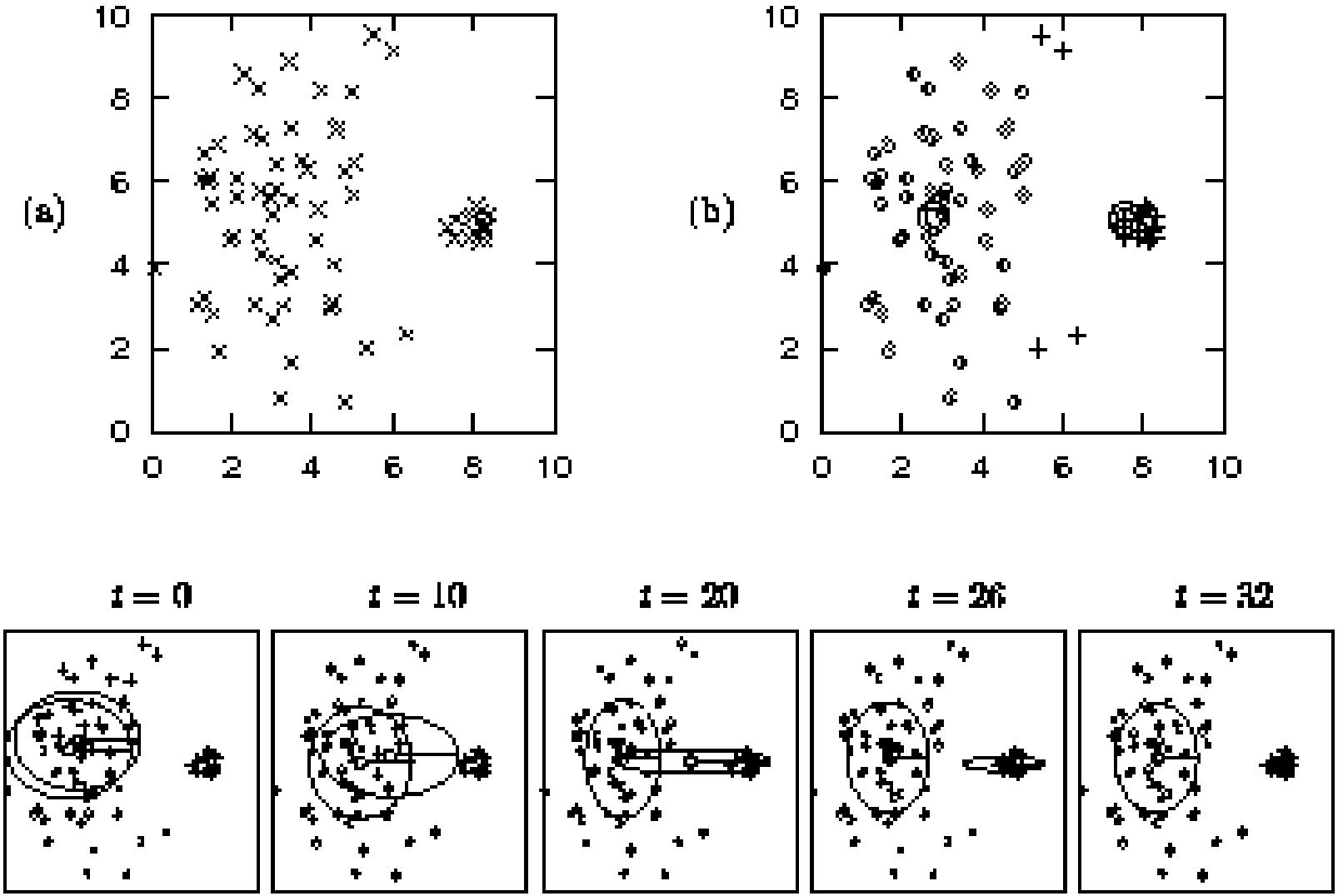
- K:n edellä mainittua muotoa olevan multinormaalijakauman sekoitus + uskottavuuden maksimointiin EM-algoritmilla perustuva ryvästys
- klusterijäsenyyksien posterioritodennäköisyydet datavektorille $\mathbf{x}$:

$$p(Cl = k|\mathbf{x}, \mu_k) = \frac{\pi_k \exp\{-\|\mathbf{x}_n - \mu_k\|^2/2\epsilon\}}{\sum_j \pi_j \exp\{-\|\mathbf{x}_n - \mu_j\|^2/2\epsilon\}}$$

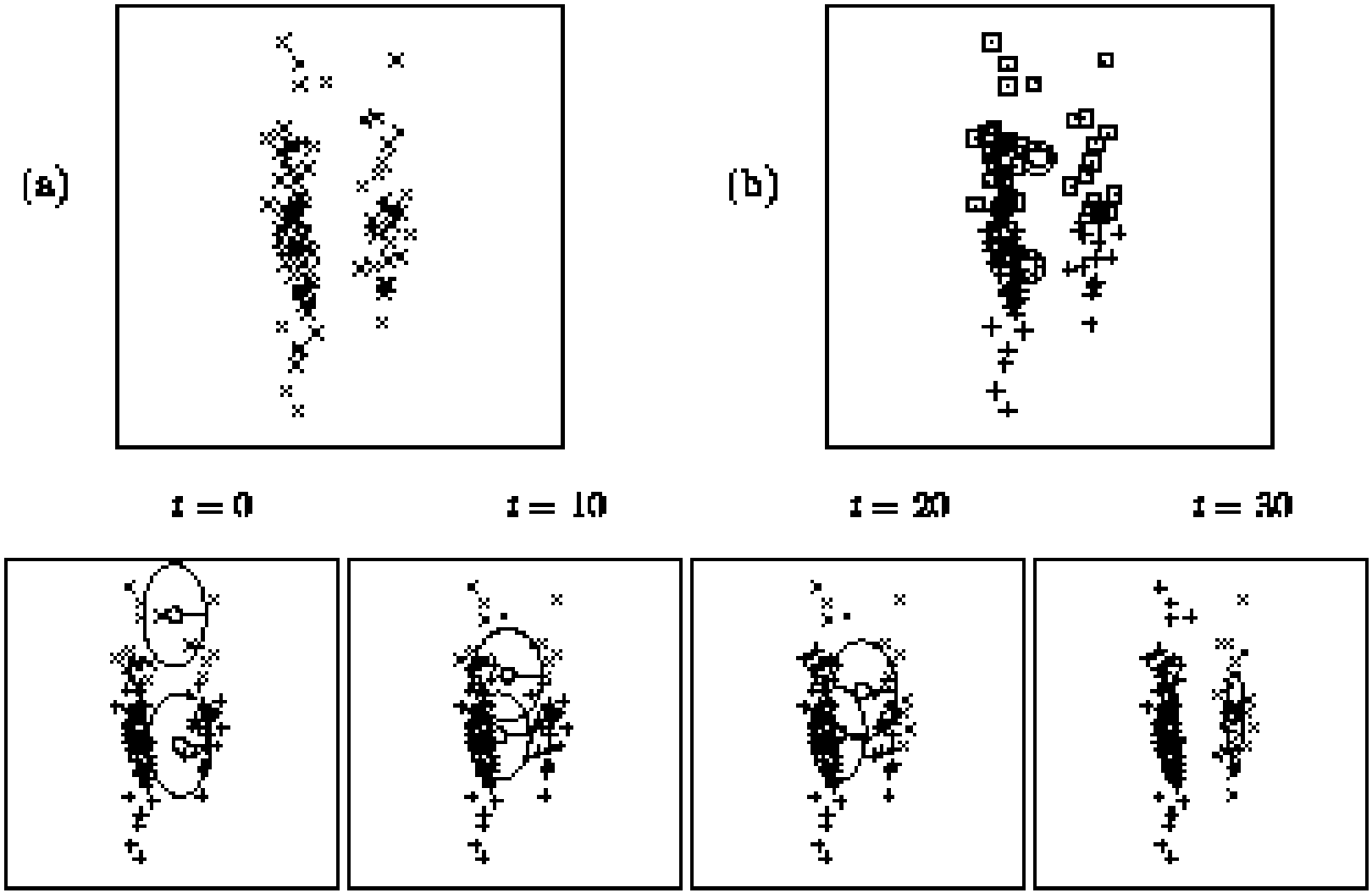
- kun $\epsilon \to 0$, kaikille $k \in \{1, \ldots, K\}$ yhtä lukuunottamatta $p(Cl = k|\mathbf{x}, \mu_k) \to 0$
- siinä yhdessä jäljelle jäävässä tapauksessa todennäköisyys lähestyy ykköstä

$\to$ EM-algoritmi yhtyy tällä rajalla K:n keskipisteen ryvästysalgoritmiin

Esimerkki: K-means ja gaussinen sekoitemalli



Esimerkki: K-means ja gaussinen sekoitemalli



Mallipohjainen klusterointi

Näin helppoa :-)

- määritä kuinka monta klusteria (K)
- valitse parametrinen malliperhe (f), esim. multinormaalijakauma
- sovella EM-algoritmia, jolla estimoidaan parametrit θ_k, π_k aineistosta

Lopuksi

Ei kun klusteroimaan (tai ryvästämään) . . .
. . . mutta niin mukavaa kuin ryvästäminen onkin, niin pidetään pää kylmänä :-)