Luento 6: Klusterointi

- Klusterointiongelma
  - etäisyys-/erilaisuusfunktiot
  - dimensionaalisuuden kirous (curse of dimensionality)
- Menetelmien pääluokat
  1. Osittavat
  2. Hierarkiset
  3. Malliperustaiset (tiheysfunktioihin perustuvat)
- Mikon esitys tasapainoisesta klusteroinnista

Klusterointiongelma

Intuitio: Halutaan löytää datan ryhmittely K:hon klusteriin $c_1, \ldots, c_K$, siten että klusterin sisäiset pisteet ovat lähekkäisiä/samankaltaisia mutta klusterien välillä on eroja/etäisyyksiä.

Tarvitaan ainakin:

- K=klusterien lkm
- d=pisteiden etäisyys- tai erilaisuusfunktio
- D=klusterien etäisyys- tai erilaisuusfunktio
- kriteeri jolla arvioida klusteroinnin hyvyyttä (score function) (ei tarvita aina)
- klusterointimenetelmä

Usein suotavia ominaisuuksia

- Ei rajoittavia oletuksia klusterin muodosta (eri klusterit voivat olla erimuotoisia)
- Samalla olisi hyvä tunnistaa ulkopuoliset
- Jos datassa ei selvästi erottuvia ryhmiä, niin pitäisi esittää jotenkin havainnollisesti klusterien päällekäisyys
- Klusterien lukumäärä haluttaisiin määrittää datasta
- Tasapainoisuus joskus suotavaa (esim. diskretointi), mutta dataa ei pidä väkisin pakottaa siihen (esim. ulkopuoliset, vähemmistöt)
- Lisäinformaationa mahdollisesti klusterien väliset suhteet (klusterihierarkia) tai datan tiheysjakauma

Usein suotavia ominaisuuksia

- Jos datassa ei oikeasti ole ryhmiä, sekin olisi tärkeää saada selville!

Vaara: Klusterointi voi luoda dataan (katsojan silmään) ryhmittelyn, vaikkei sitä olisikaan!

Etäisyysfunktio numeeriselle datalle

Merk. datapiste $x = (x_1, \ldots, x_k)$ (kaikkien k:n muuttujan arvot)

- yleisin d on L_p-metriikka eli Minkowski-etäisyys (perhe etäisyysmittoja):

$$L_p(x,y) = \left[\sum_{i=1}^{k} (x_i - y_i)^p \right]^{1/p}$$

- $p \in \mathbb{R}^+$ kertoo mistä mitasta on kyse
- kun $p = 2$, euklidinen etäisyys
- kun $p = 1$, Manhattan-etäisyys
- robustimpi mitta, kun pieni p

Etäisyysfunktio numeeriselle datalle

- L_p:t toimivat hyvin, kun klusterit tiiviitä ja hyvin erottuvia
- mutta eivät toimi, jos muuttujat eri skaalassa!
- Mitä jos yksi muuttuja grammoja, toinen milligrammoja ja kolmas senttejä??

Dimensionaalisuuden kirous (curse of dimensionality)

Data-avaruuden tilavuus kasvaa eksponentiaalisesti dimension funktiotan → avaruudesta tulee hyvin harva. Kaikki naapuripisteet kaukana eikä data riittäisi enää minkään mallin määritykseen.

- muitakin merkityksiä (kaikenlaiset suuresta dimensiosta seuraavat vaikeudet)
- klusteroinnin erityisongelma: ei enää selkeitä tihentymiä
- etäisyysmitta menettää merkityksenä, kun kaikkien pisteiden etäisyydet jokseenkin yhtä suuria
- euklidinen etäisyys L_2 kärsii erityisesti → L_1 (tai $L_{0.5}$ tms. $p < 1$) parempi

Dimensionaalisuuden kirous (jatk.)

- epärelevantit muuttujat alkavat helposti dominoida mittaa
  Mitä jos $k = 20$ mutta vain 2 muuttujaa relevantteja?

Huom! aikavaativuuden kannalta datarivien lkm n kriittisempi ongelma.

- n pistettä voi jakaa K:hon klusteriin $\frac{K^n}{K!}$ eri tavalla
- yleensä klusterointiogelmat *NP*-kovia, mutta riippuu tietysti klusteroinnin määritelmästä!

1. Osittavat menetelmät (partitioning)

Idea: Siirrellään datapisteitä klusterista toiseen, kunnes hyvyysfunktio ei enää parane.

- Klusterien esitys=centroidit ("keskipisteet")
  - centroidin muuttuja-arvot yleensä keskiarvoja tai mediaaneja klusterien pisteiden muuttuja-arvoista
- tunnetuin menetelmä K-**means**

K-means

Minimoitava funktio $SSE = \sum_{j=1}^{k} \sum_{x_i \in c_j} d^2(x_i, centroid(c_j))$

- yleensä $d = L_2$ (euklidinen etäisyys)
- Oletus: klusterit erillisiä, tiiviitä hyperpalloja
- kärsii epärelevanteista muuttujista (erit. suuridimensioisessa datassa) → voi tulla virheellisiä klustereita
- alkucentroidit valitaan satunnaisesti → eri ajokerroilla eri tuloksia → kannattaa ajaa useita kertoja!
- jos juuttuu lokaaliin optimiin tai oikeat klusterit eivät oletuksen mukaisia, tulee epäintuitiivisia tuloksia (usein huono "oikean elämän" datalla)

Etuja ja haittoja

+ varsin tehokas
+ helppo toteuttaa
− hyvin rajoittunut oletus klustereista
− voi tuottaa epäintuitiivisia tai selvästi vääriä klusterointeja
− L_2 hyvin herkkä ulkopuolisille → k-medians parempi
− L_2 ei sovi suuridimensioiselle datalle → esim. L_1 parempi
− vaikea soveltaa kategoriselle tai sekadatalle
− parametrien valinta kriittistä

Muita osittavia menetelmiä

Muuttamalla klusteroinnin hyvyysfunktiota, voi tunnistaa myös hyperellipsoidit tai tuottaa samankokoisia tai suurinpiirtein smanamuotoisia klustereita.

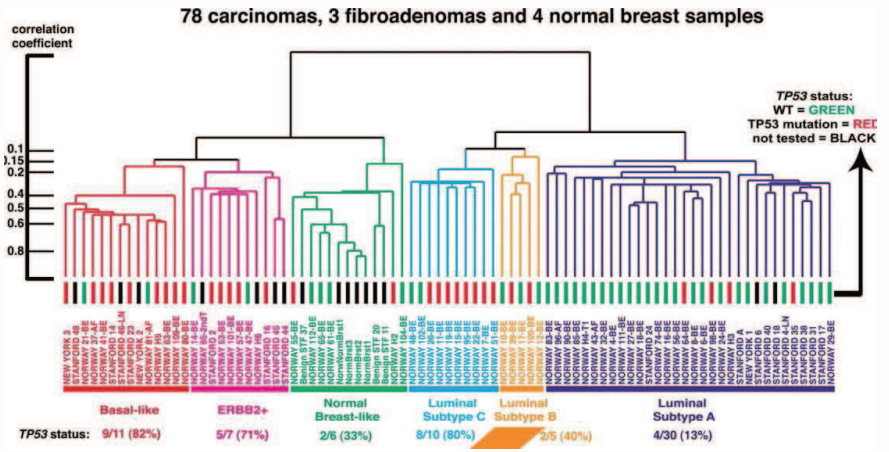
→ Smyth et al, luku 9.4.1!

2. Hierarkiset menetelmät

Idea: Yhdistetään (tai ositetaan) klustereita hierarkisesti, kunnes jäljellä sopiva määrä klustereita.

- tuloksena saadaan **dendrogrammi**, joka kuvaa klusterien hierarkian
- dendrogrammin voi myös rakentaa loppuun asti (lehdissä 1-alkioiset klusterit ja juuressa koko datajoukko) ja valita sopivan lukumäärän jälkikäteen

Dedrogrammeilla voi kuvata lajien geneettisiä sukulaissuhteita!



Lähde: Sörlie et al: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS vol. 98 no. 19, 2001.

Dedrogrammin rakennus

- kaksi rakennustapaa:
  1. Yhdistävät (agglomerative) menetelmät rakentavat dendrogrammia lehdistä juureen (yleisempi ja tehokkaampi)
  2. Jakavat (divisive) menetelmät rakentavat juuresta lehtiin
- ei eksplisiittistä hyvyysmittaa
- Monia vaihtoehtoisia mittoja klusterien etäisyydelle $D(c_1, c_2) \rightarrow$ **hyvin erilaisia tuloksia!**

Klusterien etäisyysmittoja $D(c_1, c_2)$

1. Single-link: $D(c_1, c_2) = \min_{x,y}\{d(x,y) \mid x \in c_1,\ y \in c_2\}$ (lähimpien pisteiden etäisyys)
2. Complete-link: $D(c_1, c_2) = \max_{x,y}\{d(x,y) \mid x \in c_1,\ y \in c_2\}$ (kauimpien pisteiden etäisyys)
3. Centroid-mitta: $D(c_1, c_2) = d(centroid(c_1), centroid(c_2))$
4. Average-link: $D(c_1, c_2) = \frac{\sum_{x \in c_1} \sum_{y \in c_2} d(x,y)}{|c_1||c_2|}$ (kaikkien pisteparien keskimääräinen etäisyys)
5. Minimum variance: $D(c_1, c_2) = \frac{|c_1||c_2| d^2(centroid(c_1), centroid(c_2))}{|c_1|+|c_2|}$ (varianssin kasvu jos yhdistetään)

Klusterien etäisyysmitat: Vertailua

Mitta	Klusterin muoto	Huomioita
Single link	pitkulaisia, epäsäännöllisiä jopa sisäkkäisiä	sopii erillisille epäsymmetrisille Epävakaa! (kohina)
Complete link	pieniä, tiiviitä	Ei löydä pitkulaisia Ulkopuoliset sotkevat.
Centroid-mitta	tiiviitä, erillisiä hyperpalloja	Joskus huonoja tuloksia. Tehokas laskea Aika robusti.

Vertailua (jatkuu)

Mitta	Klusterin muoto	Huomioita
Average link	melko tiiviit, suuremmat voivat olla harvempia	Kehuttu! Robusti
Minimum Variance	samanmuotoisia, jokseenkin samankokoisia, symmetrisiä melko tiiviitä	Herkkä ulkopuolisille! Tehokas laskea

Etuja ja haittoja

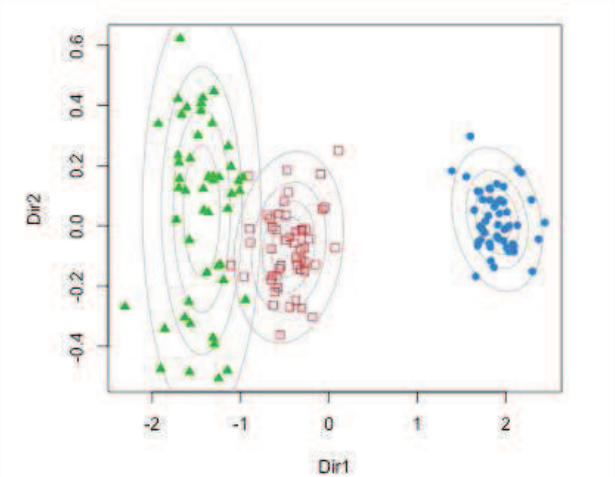
+ käytännössä toimivat yleensä hyvin (joskin riippuu mitasta!)
+ dendrogrammi lisää informaatiota
+ klusterien lkm:n voi valita dendrogrammista
+ sovellettavissa myös ei-numeeriselle datalle, kunhan klusterien etäisyydelle mitta
– useimmilla mitoilla tulos riippuu datan järjestyksestä! (em:ista complete-, centroid-, average-link ja min. var)
– skaalautuvat huonosti suurille datoille (kun paljon rivejä)
– menetelmät eivät osaa korjata aiempia tasoja

3. Malliperustaiset (tiheysfunktioihin perustuvat) klusterointimenetelmät

Idea: Määritetään koko datan tiheysjakauma, jolla data uskottavin. Samalla saadaan klusteroinnin probabilistinen kuvaus.

- sama idea kuin probabilistisessa luokittelussa!
- datapistettä ei määrätä deterministisesti vain yhteen klusteriin, vaan se voi kuulua eri todennäköisyyksillä eri klustereihin
- erittäin informatiivinen (enemmän kuin vain klusterointi)
- mutta vaikeaa tehdä hyvin...

Tuloksena saatu datan “tiheyskartta”



Lähde: mclust, R-paketti tällaiseen!
`http://www.stat.washington.edu/mclust/`

[Extra] Malliperustaisen klusteroinnin idea tarkemmin

Jokaiselle klusterille c_i ($i = 1, \ldots, K$) määritetään

- tiheysfunktio $P(x|c_i, \theta_i)$
  - kertoo todennäköisyyden, että piste x kuuluu klusteriin c_i
  - θ_i sisältää kaikki parametrit (esim. keskiarvo ja keskihajonta, jos normaalijakauma) – nämä siis **tuntemattomia!**
- prioritodennäköisyys $P(c_i)$ (klusterin todennäköisyys, jos dataa ei tunnettaisi)

[Extra] Datan uskottavuus annettuna klusterointi

- Merk. $\theta = (\theta_1, \ldots, \theta_K)$
- Datapisteen x (posteriori) todennäköisyys koko mallissa on

$$P(x|\theta) = \sum_{i=1}^{K} P(c_i)P(x|c_i, \theta_i)$$

- Jos todennäköisyys $P(x|\theta)$ kovin pieni, piste ei sovi malliin ja se voidaan tulkita sen ulkopuoliseksi
- Koko datan uskottavuus annettuna klusterointi on silloin

$$P(data|\theta) = \Pi_{j=1}^{n} P(x_j|\theta)$$

(eli kaikkien datapisteiden todennäköisyyksien tulo)

[Extra] Miten ratkaista θ?

- Maximum likelihood-periaate: Valitaan sellainen θ, joka maksimoi $P(data|\theta)$:n (eli data on uskottavin)
- vaihtoehtoisesti voisi mallintaa bayesläisittäin:

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{P(data)}$$

- Käytännössä ratkaistaan yleensä iteratiivisesti EM-algoritmilla

Millainen todennäköisyysjakauma?

- Data yleensä moniulotteista, joten tarvitaan monen muuttujan tiheysfunktio!
- Yksinkertaisuuden vuoksi oletetaan yleensä, että **muuttujat ovat toisistaan riippumattomia**
- Epärealistista, mutta silloin yhteinen tiheysfunktio helppo esittää (yhden muuttujan tiheysfunktioiden tulona)
- Muuttujat voisi tehdä ensin riippumattomiksi (ainakin korreloimattomiksi) PCA:lla
- Numeerisille muuttujille oletetaan yleensä normaalijakauma

Etuja

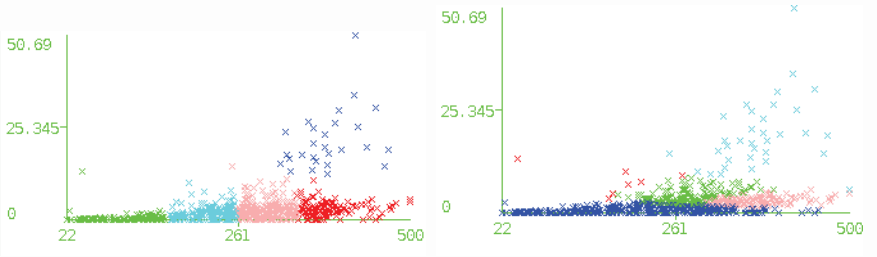
+ hyvin informatiivinen
+ voi hyvin erilaisia klustereita, myös päällekkäisiä
+ erittäin joustava (joka klusteri voisi noudattaa jopa erityyppistä jakaumaa)
+ tunnistaa samalla ulkopuoliset (ainakin periaatteessa)
+ toimii myös kategorisella datalla (mutta vaikeampaa)

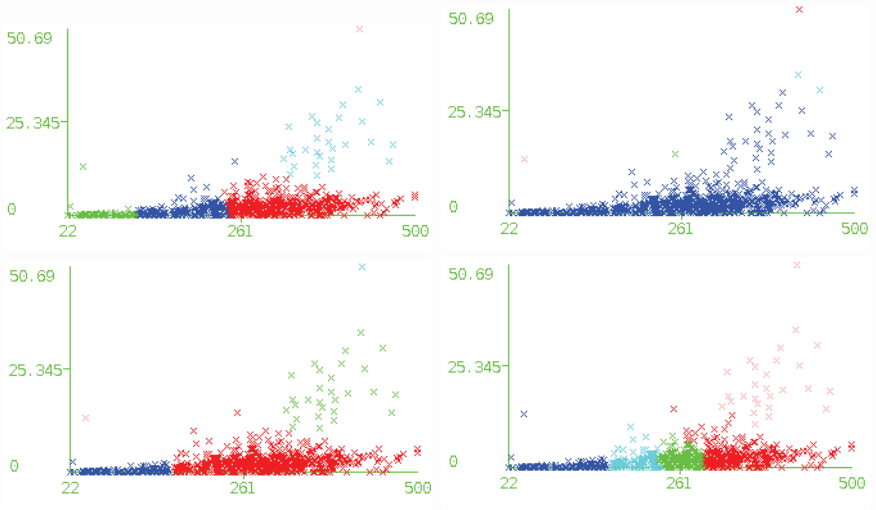
Haittoja

– tulokset riippuvat vahvasti oletetusta jakaumasta ja K:sta
– vaikea tietää ennalta “oikeita” jakaumatyyppejä!
– jos ol. normaalijakauma, ei robusti ulkopuolisille
– muuttujien riippumattomuuden oletus rajoittava! → voi tulla huonoja tuloksia (ellei hoida esim. PCA:lla)
– vaikea käsitellä sekaisin numeerisia ja kategorisia muuttujia
– algoritmit hitaita
– EM-algoritmi suboptimaalinen ja voi joskus tuottaa huonoja tuloksia

Esimerkkejä: Wekan K-means ja EM

(Muuttujat rotan paino/g ja gon-rasva/mg, $K = 5$)



Onko EM-klusterointi järkevä?? (ei kuvassa)

Complete-, single-, average-link ja min. variance