Klusterointi: esimerkki

Määritelmiä

Klusterointi:

- "Aineiston ryhmittely samankaltaisuuden perusteella"
- "Heterogeenisen aineiston jako homogeenisempiin osiin"

- Usein aineiston ositus, paljon muitakin vaihtoehtoja

Sovelluksia

- "Millaisiin luonteviin ryhmiin aineistoni voidaan jakaa?"
  (etsi klusterit)
- "Millaiset havainnot ovat tyypillisiä tai edustavia?"
  (anna tyypilliset eri klustereiden edustajat)
- "Miten aineistoani voi luonnehtia lyhyesti?"
  (kuvaa klusterit koko aineiston asemesta)

- Usein osa eksploratiivista dataan tutustumista

Ongelmat

1. Algoritminen optimointiongelma:
   k^n vaihtoehtoista ositusta klustereihin
   (k: klusterien lkm, n: datapisteiden lkm)
2. Klusteroinnin tavoitteen määrittely:
   Mikä on optimoitava funktio?!
   Miten klusteroinnin laatua mitataan?

K:n keskiarvon klusterointimenetelmä

1. Yksinkertainen esimerkki
2. Klusteroinnin tavoite: optimoitava funktio
3. Klusterointialgoritmi

Esimerkki

Klusteroinnin tavoite

- Syöte: joukko datapisteitä $d_1, \ldots, d_n$
  Parametri: k, klusterien lukumäärä
- Tavoite: sijoita k klusterikeskipistettä c_j siten, että

$$\sum_i \min_j (dist(d_i, c_j)^2)$$

  minimoituu (d_i: datapiste, c_j: klusterikeskipiste, $dist(d_i, c_j)$: pisteiden välinen (euklidinen) etäisyys
- Eli:
  - klusterit kuvataan niiden keskipisteinä
  - kukin datapiste kuuluu lähimpään klusteriin
  - tavoitteena sijoittaa klusterit (keskipisteet) niin, että ne ovat keskimäärin mahdollisimman lähellä datapisteitä

Algoritmi

K means clustering
Input:
– datapoints $\{d_1, \ldots, d_n\} = D$, Euclidean distance function $dist$ – k: number of clusters
Output:
– k cluster means (and a partitioning of D to k clusters) such that $\sum_i \min_j (dist(d_i, c_j)^2)$ is approximately minimized (a local optimum)
Method:

1. let $c_1, \ldots, c_k$ be random data points (in the domain of D)
2. repeat while cluster means c_k change:
   2.1. assign each data point d_i to the nearest cluster c_j
   2.2. recompute cluster means