582631 Johdatus koneoppimiseen (4 op)
Syksy 2008
Marko Salmenkivi

Harjoitukset 6
3.12.2008

1. Olkoot x_1 ja x_2 d-ulotteisia bittivektoreita. Hammingin etäisyys on vektorien toisistaan poikkeavien bittien lukumäärä.

   Osoita, että
   $$d_{Ham}(x_1, x_2) = ||x_1 - x_2||^2$$
   missä $||\cdot||$ on Euklidinen etäisyys.

2-3. K:n mediaanin ryvästysalgoritmi (K-medoids algorithm) on K:n keskiarvon ryvästysalgoritmin variantti, jossa ryppään edustajavektoriksi valitaan *medoidi*

$$r_k = \text{argmin}_{x \in C_k} \sum_{z \in C_k} d(x, z),$$

   eli esimerkki, josta on pienin yhteenlaskettu etäisyys ryppään muihin esimerkkeihin.

   Osoita, että K mediaanin algoritmi pysähtyy, kun etäisyysmitaksi valitaan Manhattan-etäisyys $d(x, z) = \sum_{j=1}^{d} |x_j - z_j|$.

   (Vihje: K-means algoritmin todistuksen peruslinja soveltuu myös tähän)

4. Eräs ryppäiden määrän valinnassa usein käytetty indeksi määritellään

$$I_{CV}(K) = \frac{1}{K} \sum_{k=1}^{K} \max_{l \neq k} \frac{s_k + s_l}{d_{k,l}}$$

   missä

   - $x_1, ..., x_n$ datapisteet
   - K klusterien lukumäärä
   - $C_k, k = 1, ..., K$, klusteriin k kuuluvien pisteiden joukko
   - $\mu_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$
   - $s_k = \frac{1}{|C_k|} \sum_{x \in C_k} ||x - \mu_k||$
   - $d_{k,l} = ||\mu_k - \mu_l||$

   Ryppäiden määräksi valitaan se K joka antaa pienimmän indeksin arvon, eli

$$K = \text{argmin}_K I_{CV}(K)$$

   Selitä mitä indeksi mittaa ja miksi se on (ainakin jollain tapaa) järkevä ryppäiden määrän valintaperiaate.

5. Testaa indeksin I_{CV} käyttäytymistä tiedostosta `cv-index-test-data.mat` löytyvällä datalla ja laskemalla ryvästystuloksia K-means-algoritmilla. Laske ensin mahdollisimman hyvät ryvästystulokset K:n arvoille $K = 2, ..., 5$ ajamalla K-means algoritmia 10 kertaa ja valitsemalla se ryvästys joka minimoi ryvästyksen kustannuksen.

   Laske sen jälkeen indeksi $I_{CV}(K)$ kullekin K:n arvolle valitusta ryvästyksestä. Tulosta indeksi I_{CV} klusterien määrän K funktiona. Näyttääkö indeksi käyttö valintakriteerinä tuottavan mielekkään klusterien määrän?

   Vertaa indeksin käyttöä valintakriteerinä kyynärpääkriteerin käyttöön.

Ohje tehtävissä 4-5 tarvittavien funktioiden käyttöön

Tarvittavat funktiot ja data löytyvät osoitteesta `http://www.cs.helsinki.fi/group/joko/clustering.tar.gz`

Matlabin SOM-paketin funktio `[clusterMeans,clusterIndices,objective] = som_kmeans(batch, X, K)`; implementoi K-means-algoritmin. Funktion kutsussa ensimmäinen argumentti on näissä tehtävissä aina *batch*, toinen argumentti X on datamatriisi ja kolmas argumentti K on etsittyjen klustereiden määrä. Paluuarvoina funktio palauttaa klusterien keskipisteet $clusterMeans$, datapisteiden klusterinumerot $clusterIndices$ (kokonaislukuja välillä $1, ..., K$) sekä K-means kustannusfunktion arvon *objective*.

Indeksin I_CV implementoi funktio `cvInd = cvIndex (X, clusterIndices)` missä ensimmäinen argumentti on datamatriisi ja toinen funktion som_kmeans palauttama datapisteiden ryvästys. Testidata on kaksiulotteista; ryvästystuloksia voit halutessasi visualisoida funktiolla `plotClusteringResult (X, clusterIndices)`