

Datasta Tietoon, syksy 2007

HARJOITUSTEHTÄVÄT 4

[pe 30.11.2007, ke 5.12.2007]

H4 / 1. (Ryhmittelyanalyysi)

On annettuna n vektoria. Monellako tapaa ne voi jakaa kahteen ryhmään? Ratkaise ainakin tapaukset $n = 2, 3, 4, 5$.

H4 / 2. (Ryhmittelyanalyysi)

On annettuna seuraava datamatriisi:

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & 2.5 & 3 & 3 & 5 \\ 0 & 1 & 2.5 & 2 & 4 & 3 \end{bmatrix}$$

a) Piirrä datavektorit tasoon.

b) Tee vektoreille hierarkinen ryhmittely kuvan avulla. Käytä ryhmien etäisyytenä pienintä niihin kuuluvien vektoreiden etäisyyttä. Piirrä ryhmittelypuu. Mikä on paras ryhmittely kolmeen ryhmään?

H4 / 3. (Ryhmittelyanalyysi)

On annettuna kolme vektoria $\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2$. Aluksi $C_1 = \{\mathbf{x}\}$, $C_2 = \{\mathbf{z}_1, \mathbf{z}_2\}$.

a) Laske ryhmien C_1, C_2 keskipisteet $\mathbf{m}_1, \mathbf{m}_2$.

b) Ilmenee että $\|\mathbf{z}_1 - \mathbf{m}_1\| < \|\mathbf{z}_1 - \mathbf{m}_2\|$ ja siten c-means-algoritmissa vektori \mathbf{z}_1 siirtyy ryhmästä C_2 ryhmään C_1 . Merkitään uusia ryhmiä $C'_1 = \{\mathbf{x}, \mathbf{z}_1\}$, $C'_2 = \{\mathbf{z}_2\}$. Laske uudet keskipisteet $\mathbf{m}'_1, \mathbf{m}'_2$.

c) Todista että

$$\sum_{\mathbf{x} \in C_1} \|\mathbf{x} - \mathbf{m}_1\|^2 + \sum_{\mathbf{x} \in C_2} \|\mathbf{x} - \mathbf{m}_2\|^2 > \sum_{\mathbf{x} \in C'_1} \|\mathbf{x} - \mathbf{m}'_1\|^2 + \sum_{\mathbf{x} \in C'_2} \|\mathbf{x} - \mathbf{m}'_2\|^2$$

eli c-means-ryhmittelyn kriteeri J pienenee.

H4 / 4. (SOM)

Tarkastellaan SOM-algoritmin laskennallista vaativuutta. Olkoon kartan koko $N \times N$ yksikköä (neuronia), ja syöte- ja painovektoreiden dimensio olkoon d . Montako kerto- ja yhteenlaskua tarvitaan, kun syötevektorille \mathbf{x} etsitään voittajaneuroni käyttäen euklidista etäisyyttä painovektoriin?

H4 / 5. (SOM)

Oletetaan tässä, että SOM-kartan painovektorit \mathbf{m}_i ja syötet \mathbf{x} sijaitsevat yksikköympyrällä (ovat 2-dimensioisia yksikkövektoreita). Kartta on 1-ulotteinen 5 yksikön kartta, jonka painovektorit alkutilanteessa on näytetty alla olevassa kuvassa. Naapurusto määritellään nyt syklistesti niin, että yksiköiden $b = 2, 3, 4$ naapurit ovat $b-1, b+1$, yksikön 5 naapurit ovat 4 ja 1 sekä yksikön 1 naapurit ovat 5 ja 2.

Opetuksessa kerroin $\alpha = 0.5$, eli kullakin askeleella voittajayksikön ja sen naapureiden painovektorit siirtyvät ympyränkehää pitkin puoleenväliin kohti pistettä \mathbf{x} . Syötevektorit voi valita vapaasti yksikköympyrältä. Valitse jono syötevektoreita niin, että painovektorit tulevat järjestykseen.

