

Datasta Tietoon, syksy 2013

TENTTI, 16.12.2013

(note: problems in English on the reverse side)

1. Oletetaan että d -ulotteiset datavektorit ovat tasaisesti jakautuneita hyperpalloon, jonka säde on 1. Määritellään hyperpallon sisäpisteiksi ne, joiden etäisyys pallon keskipisteestä on korkeintaan $1 - \epsilon < 1$. Osoita että sisäpisteiden joukon suhteellinen tilavuus menee nolnaan kun $d \rightarrow \infty$, toisin sanoen hyvin suurissa dimensioissa melkein kaikki datapisteet ovat hyperpallon kuorella. (Aputulos: r -säteisen d -ulotteisen hyperpallon tilavuus on $V_d(r) = C_d r^d$ missä vakio C_d ei riipu säteestä r .)

2. Laske suurimman uskottavuuden estimaatti eksponentiaali-jakauman

$$p(x|\lambda) = \lambda e^{-\lambda x}$$

parametrille λ kun suureesta x on olemassa otos $x(1), \dots, x(n)$.

3. Tarkastellaan 1-ulotteista 3 yksikön SOM-karttaa, jonka painot ja syöte ovat skalaareja välillä $[0,1]$. Yksikön 1 naapuri on 2, yksikön 3 naapuri on 2, ja yksikön 2 naapurit ovat 1 ja 3. Alkutilanteessa painot ovat $m_1 = 0.5$, $m_2 = 0.25$ ja $m_3 = 0.75$. Kun uusi syöte x on valittu, etsitään ensin lähin yksikkö ja sitten sen ja sen naapureiden painoja päivitetään säännöllä

$$m_i^{\text{uusi}} = m_i + 0.5(x - m_i).$$

Valitse syöte x niin, että päivityksen jälkeen uudet painot ovat suuruusjärjestyksessä:

$$m_1^{\text{uusi}} < m_2^{\text{uusi}} < m_3^{\text{uusi}}.$$

4. (a) Määrittele 0-1 datan kattava joukko (frequent set). Anna esimerkki pienestä 0-1-datajoukosta ja luettele sen kattavat joukot jollakin sopivalla kynnyksarvolla N .

4. (b) Kuvaile kattavien joukkojen etsintään käytettävän tasoittaisen algoritmin toimintaperiaate.

5. Kirjoita essee aiheesta hierarkinen ryhmittely. Esitä ryhmittelyn periaate ja laskennan vaiheet. Voit käyttää esimerkkinä sekä 2-ulotteista dataa kuin myös harmaasävyisiä digitaalisia kasvokuvia.