Datasta Tietoon, syksy 2011

TENTTI 17. 12. 2011 (note: problems in English on the reverse side)

1.

d-ulotteiset datavektorit ovat tasaisesti jakautuneita hyperkuutioon, jonka sivun pituus on s. Määritellään sisäpisteiksi ne, joiden etäisyys hyperkuution pinnalta on vähintään $\epsilon > 0$. Osoita että sisäpisteiden joukon kokonaistodennäköisyys (tasainen tiheysjakauma integroituna sisäpisteiden joukon yli) menee nollaan kun $d \to \infty$, toisin sanoen hyvin suurissa dimensioissa melkein kaikki datapisteet ovat hyperkuution pinnalla.

2.

Palvelukeskukseen saapuu keskimäärin λ puhelua minuutissa satunnaisina hetkinä. Voidaan osoittaa että tällöin todennäköisyys sille, että puheluja tulee minuutissa k kpl, noudattaa Poisson-jakaumaa:

$$P(\text{puheluja } k \text{ kpl}) = p(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{1}$$

Mitataan puhelujen määrää n:n minuutin mittaisen ajanjakson aikana ja saapuvat määrät minuutissa ovat $k_1, k_2, ..., k_n$. Laske suurimman uskottavuuden estimaatti odotusarvolle λ.

3.

Tarkastellaaan 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

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

Valitse jono syötteitä x niin, että päivitysten jälkeen uudet painot ovat suuruusjärjestyksessä:

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

4.

Mikä on PageRank-algoritmi ja mihin sitä käytetään? Määrittele PageRank-algoritmin tarvitsema syöte (eli mitä tarvitaan PageRankin laskemiseksi), kerro miten algoritmi toimii (sanallisesti ja pseudokoodilla sellaisella tarkkuudella, että sen perusteella tietotekniikan DI joka ei ole PageRankista kuullut osaisi algoritmin implementoida - luentokalvojen tarkkuus riittää) ja kerro miksi PageRank on niinkin hyödyllinen.

5.

a) (3 p) Pääkomponenttianalyysi (PCA): kerro lyhyesti, mitä sillä tarkoitetaan ja mitä laskennan vaiheita siinä on.

b) (3 p) Automaatiojärjestelmästä on luettu viiden reaalimuuttujan arvoja datamatriisiin $\mathbf{X}$. PCA-menetelmän ominaisarvoiksi on saatu

$$\lambda_1 \approx 0.16,\ \lambda_2 \approx 0.63,\ \lambda_3 \approx 0.70,\ \lambda_4 \approx 1.4,\ \lambda_5 \approx 2.1$$

ja vastaaviksi ominaisvektoreiksi

$$\mathbf{e}_1 \approx \begin{bmatrix} -0.72 \\ 0.06 \\ 0.69 \\ -0.01 \\ 0.02 \end{bmatrix},\quad \mathbf{e}_2 \approx \begin{bmatrix} -0.10 \\ 0.30 \\ -0.15 \\ -0.67 \\ 0.66 \end{bmatrix},\quad \mathbf{e}_3 \approx \begin{bmatrix} -0.26 \\ 0.84 \\ -0.33 \\ 0.23 \\ -0.26 \end{bmatrix},\quad \mathbf{e}_4 \approx \begin{bmatrix} 0.04 \\ 0.00 \\ 0.05 \\ -0.71 \\ -0.70 \end{bmatrix},\quad \mathbf{e}_5 \approx \begin{bmatrix} -0.63 \\ -0.46 \\ -0.62 \\ -0.02 \\ -0.06 \end{bmatrix}$$

Laske datapisteen $\mathbf{x} \in \mathbb{R}^5$:

$$\mathbf{x} = \begin{bmatrix} 0 & -0.5 & 1.0 & -1.0 & -0.5 \end{bmatrix}^T$$

kaksiulotteinen projektiopiste $\mathbf{y} \in \mathbb{R}^2$, kun datan vaihtelusta (energiasta) halutaan säilyttää mahdollisimman paljon.

HUOM! Annathan palautetta käymistäsi kursseista osoitteessa:
<http://www.cs.hut.fi/Opinnot/Palaute/kurssipalaute.html>
OBS! Var så god och ge kursfeedback i följande address:
<http://www.cs.hut.fi/Opinnot/Palaute/kurssipalaute.html>