Datasta Tietoon, Autumn 2011

EXAM 17. 12. 2011 (note: problems in Finnish on the reverse side)

1.

d dimensional data vectors are uniformly distributed in a hypercube with side length s. Let us define as inner points those whose distance from the surface of the hypercube is at least $\epsilon > 0$. Show that the total probability of the set of inner points (the uniform density integrated over the set of inner points) tends to zero as $d \to \infty$, in other words, in very high dimensions almost all data points are on the surface of the hypercube.

2.

A service center receives on average λ phone calls per minute, at random moments. It can be shown that the probability of receiving k calls within one minute follows the Poisson distribution:

$$P(k \text{ calls}) = p(k|\lambda) = \frac{\lambda^k exp(-\lambda)}{k!} \tag{1}$$

Let us measure the number of phone calls during n one-minute intervals and the counts are $k_1, k_2, ..., k_n$. Derive the maximum likelihood estimate for λ.

3.

Let us consider a 1-dimensional SOM map with three units, whose weights and inputs are scalars on the interval [0,1]. The neighbor of unit 1 is 2, the neighbor of unit 3 is 2, and the neighbors of unit 2 are 1 and 3. Initially, the weights are $m_1 = 0.5$, $m_2 = 0.25$ ja $m_3 = 0.75$. Once a new input x has been chosen, the nearest unit is found and the weights of itself and its neighbors are updated according to

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

Choose a sequence of inputs x in such a way that after the updates the new weights will be *in increasing* order:

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

4.

What is the PageRank algorithm and what is it used for? Define the input required by the PageRank algorithm (i.e., what is needed to compute the PageRank), describe how the algorithm works (with words and pseudocode so that a CS graduate who has not heard of the PageRank could implement it - the information presented on the lecture slides is sufficient), and explain why PageRank is so useful.

5.

a) (3 p) Principal component analysis (PCA): explain briefly, what it is and how to compute it.

b) (3 p) There are data samples from an automation system, which are read into a five-variable data matrix $\mathbf{X}$. The eigenvalues of PCA method are

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

and the corresponding eigenvectors

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

Compute for a new data $\mathbf{x} \in \mathbb{R}^5$:

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

a two-dimensional projection point $\mathbf{y} \in \mathbb{R}^2$, when as much variation (energy) as possible is saved.

NOTE! Please give feedback on the courses at the Web address
<http://www.cs.hut.fi/Opinnot/Palaute/kurssipalaute-en.html>.