

Datasta Tietoon, Autumn 2010

EXAM

23. 8. 2011

(note: problems in Finnish on the reverse side)

1.

d dimensional data vectors are uniformly distributed in a hyperball with radius 1. Let us define as inner points those whose distance from the center point of the hypersphere is at most $1 - \epsilon < 1$. Show that the relative volume of the set of inner points tends to zero as $d \rightarrow \infty$, in other words, in very high dimensions almost all data points are on the surface of the hyperball. (Auxiliary result: The volume of a d -dimensional hyperball with radius r is $V_d(r) = C_d r^d$ where the constant C_d does not depend on the radius r .)

2.

We are given a sample $x(1), \dots, x(n)$ of a variable x known to be normally distributed:

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We have good reason to assume that the average value μ is close to zero. Let us code this assumption into a prior density

$$p(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2}.$$

Derive the Bayes MAP estimate for the value μ and interpret your result when the variance σ^2 changes from a small to a large value.

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 an input x in such a way that after the update the new weights will be in increasing order:

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

4.

(a) Define the frequent set of 0-1 data. Give an example of a small 0-1 data set and list its frequent sets using some suitable threshold value N .

(b) Describe the principle of the levelwise algorithm for finding frequent sets.

5.

Answer one of the following essay questions that are associated with the Matlab exercise:

A) "Eigenfaces" and the use of eigenvalues for clustering face images

B) k-nearest neighbor classifier.