CS-C3160 Data Science
Hollmén
Exam, 18.12.2017

Information for students: the questions are available only in English, but you may answer in Finnish, Swedish, or in English.

1. Please indicate whether the following statements are TRUE or FALSE.

a) When mining for frequent itemsets, all candidate sets are frequent
b) Only some of the subsets of frequent itemsets are frequent
c) PageRank only use one type of a weight to decribe the relevance of a network node
d) The number of possible frequent itemsets for a d-dimensional 0-1 data is 2^d
e) c-means clustering algorithm represents cluster centers as vectors in the data space
f) Hierarchical clustering will determine the optimal number of clusters for a data set
g) In a k nearest neighbor classifier, you always select k to be odd, that is, 1,3,5, ...
h) In k nearest neighbors classifier: smaller the k, better the results
i) Principle of maximum likelihood states that the parameter estimates should maximize the likelihood of the observed data
j) Hubs and authorities algorithm represents the relevance of the network nodes with two separate sets of weights
k) The prior distribution always depends on data
l) Linear filtering can be realized with convolution

2. Assume that 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 to all points within the hypershepere tends to zero as $d \to \infty$, 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.)

3. Derive the maximum likelihood estimate for the parameter λ of the exponential probability density

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

when there is available a sample $x(1), ..., x(n)$ of the variable x.

4. Let's assume a set of five data vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_5$, with readily computed pairwise distances $d(\mathbf{x}_i, \mathbf{x}_j)$ given in the following matrix:

$$D = \begin{bmatrix} 0 & 4 & 9 & 6 & 5 \\ 4 & 0 & 1 & 8 & 7 \\ 9 & 1 & 0 & 2 & 3 \\ 6 & 8 & 2 & 0 & 1 \\ 5 & 7 & 3 & 1 & 0 \end{bmatrix}$$

Describe a hierarchical clustering algorithm and apply it on the given data set. Assume that the distance between two groups is calculated based on the single linkage distance, using the shortest distance between the data points in the groups. Draw a dendrogram and use the clustering solution to partition (divide) the data to three groups.