CS-C3160 Data Science
Hollmén
Exam, 17.12.2018

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) In Data Science, data must always be in matrix format.
b) Filtering a signal can only be done in time domain, with convolution of the signal and the filter.
c) Principal Component Analysis (PCA) performs a rotation of the original coordinate axis of the data
d) Principle of maximum likelihood states that the parameter estimates should maximize the likelihood of the observed data.
e) Maximum likelihood estimation is equivalent to Bayesian estimation using an uniform prior.
f) If two random variables X and Y are uncorrelated, the entries in the covariance matrix Cov(X, Y) have non-zero values.
g) Learning classifiers is form of supervised learning.
h) Clustering algorithms divide the data set into clusters of equal size.
i) k-means clustering algorithm is initialized by defining the boundaries of the clusters.
j) Apriori algorithm is used to find order within transactions that occur frequently.
k) The number of possible frequent itemsets for d-dimensional data is 3^d.
l) Hubs and authorities algorithm represents the relevance of the network nodes with two separate sets of weights.

2. Assume that d-dimensional data vectors are uniformly distributed in a hypercube with side length 1. Let us define as inner points those whose distance from the surface of the hypercube is at least $\epsilon > 0$. Show that the relative volume of the set of inner points to all points within the hypercube 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.

3. 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:

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

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 the parameter λ.

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.