CS-C3160 Data Science
Hollmén
Exam, 21.2.2019

Information for students: the questions are available only in English, but you may answer in Finnish, Swedish, or in English. You are allowed the use a calculator in the exam.

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

a) Size of the data matrix depends on the dimensionality of the feature vector
b) Convolution of two signals in temporal domain can be realized with a sum of their transforms in frequency domain
c) Linear filtering can be realized with convolution
d) Principal Component Analysis (PCA) performs a rotation of the original coordinate axis of the data
e) Principle of maximum likelihood states that the parameter estimates should maximize the likelihood of the observed data.
f) Maximum likelihood estimation is equivalent to Bayesian estimation using an uniform prior.
g) Prior distribution describes the parameter distribution after the measurements
h) Learning classifiers is form of supervised learning.
i) In a k nearest neighbors classifier, you always select k to be even, that is, 2,4,6, ...
j) K nearest neighbors classification gives the same results no matter what distance measure is used
k) Clustering algorithms require class labels for the data vectors
l) k-means clustering algorithm represents cluster centers as vectors in the data space
m) Hierarchical clustering gives all solutions from 1 to N clusters, where N is the number of data points
n) Apriori algorithm is used to find order within transactions that occur frequently
o) The number of possible frequent itemsets for a d-dimensional 0-1 data is 2^d
p) Only some of the subsets of frequent itemsets are frequent
q) Hubs and authorities algorithm represents the relevance of the network nodes with two separate sets of weights
r) PageRank algorithm has been the original basis of the Google search engine

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 hypershpere 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.