CS-C3160 Data Science
Hollmén
Exam, 9.4.2019

Information for students: the questions are available only in English, but you may answer in Finnish, Swedish, or in English. Use of a calculator is allowed 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) Prinicipal Component Analysis (PCA) performs a rotation of the original coordinate axis of the data

e) Assuming a diagonal covariance matrix in the Gaussian distribution reduces the number of parameters

f) Principle of maximum likelihood states that the parameter estimates should maximize the likelihood of the observed data

g) In a k nearest neighbor classifier, you always select k to be even, that is, 2,4,6, ...

h) K nearest neighbors classification gives the same results no matter what distance measure is used

i) Prior distribution describes the parameter distribution after the measurements

j) Clustering algorithms require class labels for the data vectors

k) c-means clustering algorithm represents cluster centers as vectors in the data space

l) Hierarchical clustering needs the optimal number of clusters for a data set before running the algorithm

m) When one-dimensional Self-Organizing Map has been ordered in one-dimensional space, it can not be unordered.

n) When mining for frequent itemsets, all frequent sets have previously been candidate sets

o) Only some of the subsets of frequent itemsets are frequent

p) The number of possible frequent itemsets for a d-dimensional 0-1 data is d^2

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. Describe the k-means algorithm and write down the associated cost function J it attempts to minimize. Assume you have a data set $x(1), x(2), \ldots, x(n)$. Denote the cluster centers with $m_1, m_2, \ldots, m_k$ and the set of data vectors associated with the center m_i with C_i.

3. Derive the maximum likelihood estimate for the location parameter μ of the Gaussian distribution

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

$x_1^2 - 2x_1\mu + \mu^2$

when there is available a data sample $x(1), x(2), \ldots, x(n)$ of the variable x.

$\sum_{i=1} x_i^2 - 2\mu\left(\sum_{i=1}^{n} x_i\right) + n\mu^2$