STOCKHOLM UNIVERSITY
DEPT. OF MATHEMATICS
Div. of Mathematical statistics

MT 7039
EXAMINATION
18 Mar 2021

Exam in Unsupervised Learning
18 Mar 2021, time 13:00-18:30

Examinator: Chun-Biu Li, cbli@math.su.se.
Permitted aids: When writing the home exam, you may use any literature.
Return of the exam: To be announced later.

NOTE: The exam consists of 3 problems with 100 points in total. Logical explanation and steps leading to the final solution must be clearly shown in order to receive full marks. Minimum points to receive a given grade are as follows:

A	B	C	D	E
90	80	70	60	50

NOTE: For those parts require explanation in words, your writing must be to the point, **redundant writing irrelevant to the solution will result in point deduction**.

Problem 1 (Basics of unsupervised learning, total 30p)

a) Consider the Gaussian mixture model (GMM) discussed in chapter 9 of the book "Pattern recognition and machine leanring", and suppose that the covariance matrices of all mixture components are given by $\epsilon\mathbf{I}$ such that the probability distribution function of the k-th Gaussian component is given by Eq. 9.41 in the book (Note: there is a small typo in Eq. 9.41!). Under this setting, show that, in the limit $\epsilon \to 0$, the maximization of the GMM log-likelihood Eq. 9.14 is the same as the minimization of the K-means objective function Eq. 9.1. **(18p)**

b) Consider the intergroup dissimilarities in hierarchical clustering discussed in chapter 14 of "Element of statistical learning". With the help of drawing illustrative diagrams, give an example showing the drawback of using the single linkage dissimilarity, $d_{SL}(G, H) = \min_{i\in G, i'\in H} d_{ii'}$, to merge groups **(3p)**.

   Similarly, give an example showing the drawback of complete linkage dissimilarity, $d_{CL}(G, H) = \max_{i\in G, i'\in H} d_{ii'}$, to merge groups. **(3p)** In both examples, the drawbacks should be explained concisely.

c) Explain **in words** why K-means is replaced by K-medoids for clustering when the dissimilarity measure is not Euclidean. **(3p)**

   For a dataset with finite samples, suppose that the dissimilarity measure is Euclidean, do K-medoids and K-means result in the same cluster centers for a fixed K? Please justify your answer. **(3p)**

Problem 2 (Graph based methods, total 30p)

For graphs with a single connected componet, the commute time distances (CTD), c_{ij}, expressed in terms of the eigen-values λ_α and -vectors $v_{\alpha i}$ of the normalized graph Laplacian L_{sym}, $c_{ij} = \text{vol}(G) \sum_{\alpha=2}^{N} \frac{1}{\lambda_\alpha} \left(\frac{v_{\alpha i}}{\sqrt{d_i}} - \frac{v_{\alpha j}}{\sqrt{d_j}} \right)^2$, has the form of squared Euclidean distance, where vol(G) is volume of the graph, d_i is the degree of the i-th node, with $i = 1, \cdots, N$ and $\alpha = 2, \cdots, N$. This suggests that one can embed the data points in a Euclidean space with the Cartesian coordinates $x_{\alpha i} = v_{\alpha i} \sqrt{\frac{\text{vol}(G)}{\lambda_\alpha d_i}}$, called the CTD embedding. Here α labels the directions and i labels the data point.

a) Show that $E(x_\alpha) = 0$ for $\alpha > 1$ with the weight of each data given by $P(i) = d_i/\text{vol}(G)$. **(10p)**

b) With the same weights in part a, find the covariance matrix $E(x_\alpha x_{\alpha'})$ for $\alpha, \alpha' > 1$. **(15p)**

c) **In words**, compare the outcomes from the following two scenarios with explanations: A) One applies PCA to the CTD embedding coordinates $x_{\alpha i}$; B) One applies classical MDS to c_{ij}. **(5p)**

Problem 3 (Modern nonlinear dimensionality reduction, total 40p)

This problem considers and follows the notations in the paper "Nonlinear dimensionality reduction by locally linear embedding" in the course literatures.

a) Show that the weights W_{ij}^{min} that minimize the cost function $\epsilon(W) = \sum_i \left| \overrightarrow{X}_i - \sum_j W_{ij} \overrightarrow{X}_j \right|^2$ (i.e., Eq. 1 in the paper) subject to the constraints $\sum_j W_{ij} = 1$ are invariant under translation **(3p)**, rotation **(3p)** and rescaling **(3p)** of the data coordinates $\overrightarrow{X}_i$, $i = 1, \cdots, N$.

b) Consider the constrained least square problem in solving the weights for a given data point $\overrightarrow{X}$, one minimizes $\epsilon(W) = \left| \overrightarrow{X} - \sum_j W_j \overrightarrow{\eta}_j \right|^2$ subject to $\sum_j W_j = 1$ where $\overrightarrow{\eta}_j$ are neighbors of $\overrightarrow{X}$. Show that the solution of the constrained least square problem is given by $W_j^{min} = \sum_k C_{jk}^{-1} / \sum_{l,m} C_{lm}^{-1}$ where the scalar product matrix is defined by $C_{jk} = (\overrightarrow{X} - \overrightarrow{\eta}_j) \cdot (\overrightarrow{X} - \overrightarrow{\eta}_k)$. **(15p)**

c) Related to the choice of the number of nearest neighbors, discuss in what situation the matrix C_{jk} is singular such that W_j^{min} does not exist. **(5p)** Suppose one can choose the number of nearest neighbors freely, propose **in words** a simple method to fix this problem of singular matrix. **(5p)**

d) If one does not have the feature vectors but only has the dissimilarity measures between any two data points $d(\overrightarrow{X_i}, \overrightarrow{X_j})$, discuss **in words** how the locally linear embedding method can still be applied. **(6p)**

Good Luck!