STOCKHOLM UNIVERSITY
DEPT. OF MATHEMATICS
Div. of Mathematical statistics

MT 7039
EXAMINATION
25 Apr 2022

Re-Exam in Unsupervised Learning
25 Apr 2022, time 08:00-13:00

Examinator: Chun-Biu Li, cbli@math.su.se.
Permitted aids: When writing the home exam, you may use any literature. Electronic devices are NOT allowed

---

NOTE: The exam consists of 4 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.

NOTE: Your answers and explanations must be to the point, **redundant writing irrelevant to the solution will result in point deduction**.

Problem 1 (Basics of unsupervised learning, total 34p)

a) This part follows notations in the course book "Nonlinear dimensionality reduction". Show with clear steps that $W^T = I_{p\times D}V^T$ minimize the reconstruction error $E_y\{\| y - WW^T y \|_2^2\}$ **(10p)**

b) Consider the Gaussian mixture model (GMM) in 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: Eq. 9.41 has a small typo!). Under this setting, show that, in the limit $\epsilon \to 0$, maximizing the GMM log-likelihood Eq. 9.14 equals to minimizing the K-means objective function Eq. 9.1. **(15p)**

c) For each of the following three cases, draw and explain schematically the dendrograms resulted from applying hierarchical clustering to 1) data points generated from uniform distribution **(3p)**, 2) data points generated from a single multivariate Gaussian **(3p)**, 3) data points generated from two separated multivariate Gaussian with small overlapping **(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}, $c_{ij} = \mathrm{vol}(G)\sum_{\alpha=2}^{N}\frac{1}{\lambda_\alpha}\left(\frac{v_{\alpha i}}{\sqrt{d_i}} - \frac{v_{\alpha j}}{\sqrt{d_j}}\right)^2$, has the form of squared Euclidean distance, where $\mathrm{vol}(G)$ 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 $x_{\alpha i} = v_{\alpha i}\sqrt{\frac{\mathrm{vol}(G)}{\lambda_\alpha d_i}}$, 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 $P(i) = d_i/\mathrm{vol}(G)$. **(10p)**

b) With the same weights in part a, find the covariance matrix $E(x_\alpha x_{\alpha'})$ for $\alpha, \alpha' > 1$. **(10p)**

c) **In words**, compare the outcomes from the following two scenarios with explanations: 1) One applies PCA to the CTD embedding coordinates $x_{\alpha i}$; 2) One applies classical metric MDS to c_{ij}. **(5p)**

d) Draw one example where the mutual kNN graph construction with single connected component may end up with very large k **(2p)**, then propose a solution for it **(3p)**.

Problem 3 (Local linear embedding, total 25p)

This problem follows the notation in the paper "Nonlinear dimensionality reduction by locally linear embedding".

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

b) Consider the constrained least squares problem in solving the weights for a given data point $\vec{X}$, one minimizes $\epsilon(W) = \left|\vec{X} - \sum_j W_j \vec{\eta}_j\right|^2$ subject to $\sum_j W_j = 1$ where $\vec{\eta}_j$ are neighbors of $\vec{X}$. Show that the cost function $\epsilon(W)$ can be written as the quadratic form $\epsilon(W) = \sum_{j,k} W_j C_{jk} W_k$, where the scalar product matrix is defined by $C_{jk} = (\vec{X} - \vec{\eta}_j) \cdot (\vec{X} - \vec{\eta}_k)$. **(4p)** Furthermore, show that the solution of the constrained least squares problem is given by $W_j^{min} = \sum_p C_{jp}^{-1} / \sum_{l,m} C_{lm}^{-1}$. **(6p)**

d) When does the local Gram matrix C become singular? **(3p)** Suggest a way to deal with this problem so that the LLE can still be applied. **(3p)**

Problem 4 (Validation Methods, total 11p)

This problem refers to the lecture note on validation methods

a) Draw **(2p)** and explain **(3p)** schematically the plot of Silhouette coefficient versus the number clusters when one tries to cluster data points sampled from a uniform distribution.

b) Name one limitation of the Silhouette plot and index to validate clustering results and propose a solution for it. **(6p)**

Good Luck!