STOCKHOLM UNIVERSITY	MT 7050
DEPT. OF MATHEMATICS	EXAMINATION
Div. of Mathematical statistics	4 Jan 2024

Exam in Unsupervised Learning
4 Jan 2024, time 08:00-13:00

Examinator: Chun-Biu Li, cbli@math.su.se.
Permitted aids: When writing the 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 29p)

a) 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. **(18p)**

b) Show that the principal coordinates $\hat{X}_{\text{MDS}} = I_{p\times N}\Lambda_{\text{MDS}}^{1/2}U^\top$ is centered. **(5p)**

c) Consider 3 cases, I) data points generated from uniform distribution, II) data points generated from a single multivariate Gaussian, and III) data points generated from two separated multivariate Gaussian with small overlapping. Draw schematically the binary trees resulted from applying hierarchical clustering to the 3 cases. Explain and highlight the most distinct features in the trees for the 3 cases. **(6p)**

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 $P(i) = d_i/\mathrm{vol}(G)$. **(10p)**

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

c) Show that the CTD defined above does not change if all graph weights are rescaled by the same constant, i.e., $w_{ij} \to a \cdot w_{ij}$ with $a > 0$. **(5p)**

d) Since CTD depends on the volume of the graph $\mathrm{vol}(G)$, bigger graphs with more nodes and connections have larger values of CTD. This may be a problem if one wants to compare two graphs with the same statistical properties (e.g. both of them are random graphs) but with different number of nodes and connections. Propose a way to modify the graph distance in terms of CTD that is not sensitive to the graph and at the same time keeping the graph distance invariant under rescaling of graph weights as in Past c. Justify your answer. **(5p)**.

Problem 3 (Local linear embedding, total 27p)

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 orthogonal transformation **(3p)** and rescaling **(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 $C_{jk} = (\vec{X} - \vec{\eta}_j) \cdot (\vec{X} - \vec{\eta}_k)$. **(5p)**

c) Discuss in what situation that the matrix C^{-1} in part b does not exist (i.e., when C is a singular matrix) and what are the implications in terms of the intrinsic dimension of the data structure locally around $\vec{X}$. **(6p)**

d) Consider the eigenvector problem where the $N \times N$ weight matrix W is given, one minimizes $\phi(Y) = \sum_i \left|\vec{Y}_i - \sum_j W_{ij}\vec{Y}_j\right|^2$ subject to the constraints $\sum_i \vec{Y}_i = 0$ and $\sum_i \vec{Y}_i \vec{Y}_i^T = NI$. Show that the cost function $\phi(Y)$ can be written as $\phi(Y) = \mathrm{Tr}(Y^T M Y)$ where $M = (I - W)^T (I - W)$ and Y is the $N \times d$ data matrix in the lower dimensional space. **(10p)**

Problem 4 (Validation Methods, total 14p)

This problem refers to the lecture note on validation methods

a) Name TWO limitations of the Silhouette plot and index to validate clustering results and propose solutions for each of them. **(6p)**

b) Referring to P.21 of the lecture note on co-rank matrix, show that $Q_{NX}(K) = \frac{1}{Kn}\sum_{i=1}^{n}|\Psi_K(i)\bigcap\Psi'_K(i)|$. **(5p)** What is the range of $Q_{NX}(K)$? **(3p)**

Good Luck!