STOCKHOLM UNIVERSITY
DEPT. OF MATHEMATICS
Div. of Mathematical statistics

MT 7050
EXAMINATION
13 Dec 2024

Re-Exam in Unsupervised Learning
13 Dec 2024, time 08:00-11:00

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

---

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.

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

Problem 1 (Basics, total 40p)

a) The single and complete linkage were introduced as group-to-group distances for bottom-up hierarchical clustering. Two artifacts, chaining for single linkage and failure in handling clusters with long shape for complete linkage, were mentioned in the class. Give more examples, one for the single linkage **(5p)** and one for the complete linkage **(5p)** that they may fail. Justify your answer and draw schematic figures to illustrate your examples.

b) Consider the Gaussian mixture model (GMM) in the PRML book, 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)**. Note: You should start from Eq. 9.14 and show that it is related to Eq. 9.1 by a multiplication/addition constant as $\epsilon \to 0$.

c) Show that $W^\top = I_{p\times D}V^\top$ minimize the reconstruction error Eq. 2.12 in NDR. **(10p)**

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

Problem 2 (Modern Methods, total 35p)

For graphs with a single connected component, 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 $\text{vol}(G)$ 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 $P(i) = d_i/\text{vol}(G)$. **(5p)**

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) The transition probability of a graph is given w_{ij}/d_i in the random walk Laplacian, show that the stationary distribution is given by $\pi_i = d_i/\text{vol}(G)$, where d_i and $\text{vol}(G)$ are the degree of the i-th node and the volume of the graph, respectively **(6p)**.

e) At the stationary state mentioned in part d and consider that the vertices of the graph are partitioned into only two non-overlapping subsets, A and $\bar{A}$ (i.e., the complement of A). Show that $P(A|\bar{A})P(\bar{A}) = P(\bar{A}|A)P(A)$ **(6p)** and discuss the meaning of the equality **(3p)**.

Problem 3 (DESCAN and Validation Methods, total 25p)

a) Propose TWO improvements for DBSCAN. Show the details and justify your suggestions. **(6p)**

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

c) Under-fitting and over-fitting in supervised learning are usually explained in terms of the training/validation curves and their relation with the bias-variance tradeoff, provide a similar explanation with schematic pictures for the under-clustering and over-clustering in terms of the cohesion-separation tradeoff for clustering analysis. **(8p)**

d) For the density based clustering method proposed in the 2014 paper by Rodriguez and Laio, explain why one CANNOT simply rely on the decision graph to validate the clustering results. **(5p)**

Good Luck!