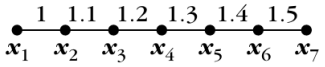
EE708: Fundamentals of Data Science and Machine Intelligence

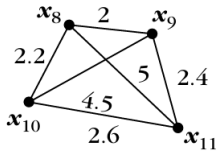
Assignment 3

Based on Module 4: Clustering and Module 5: Gaussian Mixture Modeling

1. Prove whether $d(x, y) = |x - y|^2$ satisfies the properties of a valid distance metric.
2. In many clustering schemes, a vector $\boldsymbol{x}$ is assigned to a cluster C, considering the proximity between x and C, $\mathrm{D}(x, C)$, which can be defined as:
   a. $D_{\min}(x, C) = \min_{v \in C}\{\delta(x, v)\}$ single-linkage clustering
   b. $D_{\text{avg}}(x, C) = \langle \delta(x, v) \rangle_{v \in C}$ average-linkage clustering
   c. $D_{\max}(x, C) = \max_{v \in C}\{\delta(x, v)\}$ complete-linkage clustering

   Let $C = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$, where
   $x_1 = [1.5,1.5]^T$ $x_4 = [1.5,2]^T$ $x_7 = [2,3]^T$
   $x_2 = [2,1]^T$ $x_5 = [3,2]^T$ $x_8 = [3.5,3]^T$
   $x_3 = [2.5,1.75]^T$ $x_6 = [1,3.5]^T$
   and let $x = [6,4]^T$. Assume that the Euclidean distance measures the dissimilarity between two points. Then find $D_{\min}(x, C)$, $D_{\max}(x, C)$, and $D_{\text{avg}}(x, C)$.
3. Consider the data set shown in the figure. The first seven points form an elongated cluster, while the remaining four form a rather compact cluster. The numbers on top of the edges connecting the points correspond to the respective (Euclidean) distances between vectors. These distances are also taken to measure the distance between two initial point clusters. Distances that are not shown are assumed to have very large values. Draw the corresponding dendrograms based on dissimilarity.





4. For a dataset $\mathcal{C}$ with 5 samples, consider the dissimilarity matrix

$$P = \begin{bmatrix} 0 & 4 & 9 & 6 & 5 \\ 4 & 0 & 3 & 8 & 7 \\ 9 & 3 & 0 & 3 & 2 \\ 6 & 8 & 3 & 0 & 1 \\ 5 & 7 & 2 & 1 & 0 \end{bmatrix}$$

   Where $P_{i,j} = \delta(x_i, x_j)$. Determine all possible dendrograms resulting from applying the single and the complete link algorithms to P and comment on the results.
5. Consider a 2-dimensional feature space with a dataset of $N = 10$ points. A vector quantization (VQ) system maps these points into $K = 3$ clusters using a codebook. The distortion function is the squared Euclidean distance between the original points and their assigned cluster centroids. Given the following initial cluster centroids:

$$C_1 = (2,3), \quad C_2 = (5,8), \quad C_3 = (9,4)$$

   Assign the following data points to their closest centroid using squared Euclidean distance:

$$(1,2), \quad (3,4), \quad (6,7), \quad (8,3), \quad (5,5)$$

   a. Compute the new centroids after one iteration of vector quantization.
   b. Show whether the distortion decreases after this iteration.
6. Show that if we maximize the first equation with respect to Σ_k and π_k while keeping the responsibilities $\gamma(z_{nk})$ fixed, we obtain the closed-form solutions given by the following equations:

$$E_Z[\ln p\,(X, Z|\mu, \Sigma, \pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\big(\ln \pi_k + \ln \mathcal{N}\,(x_n|\mu_k, \Sigma_k)\big)$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^{\mathrm{T}}$$

$$\pi_k = \frac{N_k}{N}$$

7. Consider a density model given by a mixture distribution

$$p(x) = \sum_{k=1}^{K}\pi_k p(\,x \mid k\,)$$

and suppose that we partition the vector **x** into two parts so that **x** = (**x**$_a$, **x**$_b$). Show that the conditional density $p(\,x_b \mid x_a\,)$ is itself a mixture distribution and find expressions for the mixing coefficients and component densities.

8. Consider a mixture of Gaussian distributions given by

$$p(x|\Theta) = \sum_{k=1}^{K}\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where:
K: number of Gaussian components
π_k: mixing coefficients such that $\sum_{k=1}^{K}\pi_k = 1$ and $\pi_k > 0$
$\mathcal{N}(x|\mu_k, \Sigma_k)$: Gaussian density with mean μ_k and covariance Σ_k
$\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{\mathrm{k=1}}^{\mathrm{K}}$ represents the parameters of the model.

   a. Write down the complete log-likelihood function for a dataset $\{x_1, x_2, \ldots, x_N\}$ assuming that the data points are drawn independently from the mixture model.
   b. Derive the Maximum Likelihood Estimation (MLE) update rules for $\pi_k, \mu_k$and Σ_k assuming that the component that generated each data point is known.

Programming Questions:

9. K-means clustering: Using the dataset in *A3_P1.csv*, implement K-means clustering and determine the number of clusters using the Elbow method.
   a. Plot the inertia (Within-Cluster Sum of Squares - WCSS) for number of cluster ranging from 1 to 15.
   b. Find the optimal number of clusters using the elbow method.
   c. Perform clustering using the optimal number of clusters, plot the clustering results with each cluster data in a different colour, and highlight the cluster centres.

   Hint: Use *KMeans* from the Python package *sklearn*.

10. Hierarchical Clustering: Using the dataset in *A3_P2.csv*, implement bottom-up hierarchical clustering from scratch using Euclidean distance as the distance metric. Compute the distance between two clusters using the following methods:
    a. $D_{min}(A, B) = \min_{u\in A,\, v\in B}\{\delta(u, v)\}$ single-linkage clustering
    b. $D_{avg}(A, B) = \langle\delta(u, v)\rangle_{u\in A,\, v\in B}$ average-linkage clustering
    c. $D_{max}(A, B) = \max_{u\in A,\, v\in B}\{\delta(u, v)\}$ complete-linkage clustering.

    Plot dendrograms for each clustering method to visualize the hierarchical clustering process.