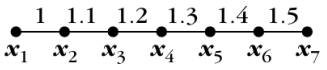
EE708: Fundamentals of Data Science and Machine Intelligence

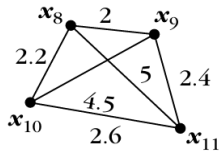
Assignment 3

Based on Module 4: Clustering and Decision Trees

1. Show that:
   a. If s is a metric similarity measure on a set X with $s(x,y) \geq 0,\ \forall x,y \in X,$ then $s(x,y) + a$ is also a metric similarity measure on $X,\ \forall a \geq 0.$
   b. If d is a metric dissimilarity measure on X, then $d + a$ is also a metric dissimilarity measure on $X,\ \forall a \geq 0.$
2. Prove that the Euclidean distance satisfies the triangular inequality. Hint: Use the Minkowski inequality, which states that for a positive integer p and two vectors $x = [x_1, \dots, x_l]^T$ and $y = [y_1, \dots, y_l]^T$ it holds that
$$\left(\sum_{i=1}^{l} |x_i + y_i|^p\right)^{1/p} = \left(\sum_{i=1}^{l} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{l} |y_i|^p\right)^{1/p}$$
3. Prove whether $d(x,y) = |x-y|^2$ satisfies the properties of a valid distance metric.
4. 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 $D_{\text{avg}}(x,C)$.
5. 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.





6. 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.
7. Having generated a dendrogram, can we "prune" it? If yes, how?
8. How can we make k-means robust to outliers?

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.