Problem 1 (Expectation-Maximization and k-means++). Given the five points below:

\[
\begin{array}{c|cc}
  & 0 & 2 \\
 1 & 0 \\
 2 & 2.5 & 1 \\
 3 & 5 & 0 \\
 4 & 5 & 2 \\
 5 & 0 & 0 \\
\end{array}
\]

a) Apply the EM algorithm to the above data with \( k = 2 \). Show one complete application of the \( E \) and the \( M \) steps, starting from the \( M \) step. Start with the assumption that \( P(C_j|x_i) = 0.5 \). Use Gaussian distributions and you may assume that the dimensions are independent.

b) Use k-means++ assuming that \( x_2 \) is selected as the first centroid and for the second centroid you select the most probable one. Use \( L_2 \) and show the initialization step and final clustering.

c) Consider the following scenario: Each of \( n \) students has to choose one item out of four choices. Consider histogram \( y = [y_1, \ldots, y_4] \), where \( y_i \) is the number of students that chose item \( i \). The probability of observing a particular histogram is modelled as a multinomial distribution

\[
\Pr(y | \theta) = \frac{n!}{y_1!y_2!y_3!y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} .
\]

A recent study concluded that the probability of choosing each of the items is parameterized by a single hidden coefficient \( \theta \in (0, 1) \), such that the probability of observing a histogram \( y \) is

\[
p_y = \left[ \frac{1}{2}, \frac{1}{4} \theta, \frac{1}{4} (1 - \theta), \frac{1}{4} (1 - \theta) \right]^T.
\]

Estimate coefficient \( \theta \) using Expectation-Maximization under the assumption that data \( X = [X_1, \ldots, X_5] \) follows a multinomial distribution \( q_\theta \), that is

\[
q_\theta = \left[ \frac{1}{2}, \frac{1}{4} \theta, \frac{1}{4} (1 - \theta), \frac{1}{4} (1 - \theta) \right]^T.
\]

Problem 2 (k-means). In k-means algorithm we compute the cluster centroid (prototype) as the mean of the points in the cluster, as this minimizes the sum of squared errors. Consider a variation of k-means for one-dimensional data where we want to minimize the sum of absolute errors, that is, our goal is to find clusters \( C_1, C_2, \ldots, C_k \) that minimize

\[
\sum_{j=1}^{k} \sum_{x_i \in C_j} |\nu_j - x_i|,
\]

where \( \nu_j \) is the prototype of cluster \( C_j \). Prove that we should use the median of points in \( C_j \) as \( \nu_j \). Hint: compute the derivative of sum of absolute errors w.r.t. cluster prototypes.
Problem 3 (Hierarchical clustering). Consider the data set \{1, 1.5, 3, 4.5, 8, 9\} of data points \(x_1\) to \(x_6\). Perform hierarchical clustering and show your results by drawing a dendrogram for

(a) Single linkage

(b) Ward’s linkage

(c) Consider the distance matrix in Table 1. Compute the hierarchical clustering using the average link method. Show the distance thresholds.

\[
\begin{array}{ccccc}
  & a & b & c & d & e \\
 a & 0 & 2 & 6 & 4 & 8 \\
b & 0 & 6 & 4 & 6 & \\
c & 0 & 2 & 6 & \\
d & 0 & 10 & \\
e & 0 & & & \\
\end{array}
\]

Table 1: Distance Matrix

Problem 4 (Distance). Consider the data in Figure 1. Answer to the following questions assuming that we are using Euclidean distance, that \(\varepsilon = 2\), and \(\text{minpts} = 3\).

Figure 1: Points in a space

a) List all the core points.

b) Is \(r\) directly density-reachable from \(q\)?

c) Is \(f\) density-reachable from \(a\)? Show the complete chain or where it breaks.

d) Is \(g\) density-connected to \(r\)? Show the intermediate points that make them density-connected or that break the condition.