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

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$x_3$</td>
<td>2.5</td>
<td>1</td>
</tr>
<tr>
<td>$x_4$</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>$x_5$</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

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 $P(C_j|x_i) = 0.5$.

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

$$
Pr(y | \theta) = \frac{n!}{y_1!y_2!y_3!y_4!} \theta^{y_1} \theta^{y_2} \theta^{y_3} \theta^{y_4}.
$$

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

$$
p_y = \left[ \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right]^T.
$$

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

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

Solution.

a) Given, $P(C_1|x_j) = 0.5$ and $P(C_2|x_j) = 0.5$, for all points $x_j$. We start with M step:

$$
\mu = \frac{\sum_{j = 1}^{5} 0.5 \cdot x_j}{\sum_{j = 1}^{5} 0.5} = (2.5, 1)
$$

The variance for cluster $i$ in dimension $a$ is $(\sigma_{ia}^2) = \frac{\sum_{j=1}^{5} w_{ij}(x_{ia} - \mu_{ia})^2}{\sum_{j=1}^{5} w_{ij}}$

We have two dimensional data, substituting in the formula above for first cluster we get:

$$
(\sigma_{11}^2) = \frac{\sum_{j=1}^{5} 0.5(x_{j1} - \mu_{11})^2}{\sum_{j=1}^{5} 0.5}
$$

$$
= \frac{\sum_{j=1}^{5} 0.5(x_{j1} - \mu_{11})^2}{\sum_{j=1}^{5} 0.5}
$$

$$
= 0.5((x_{11} - \mu_{11})^2 + (x_{21} - \mu_{11})^2 + (x_{31} - \mu_{11})^2 + (x_{41} - \mu_{11})^2 + (x_{51} - \mu_{11})^2)
$$

$$
\text{1 of 7}
$$
\[
= \frac{(0 - 2.5)^2 + (5 - 2.5)^2 + (5 - 2.5)^2}{5} = 5
\]

Similarly,
\[
(\sigma_{22}^1)^2 = \frac{\sum_{j=1}^{5} 0.5(x_{j2} - \mu_{12})^2}{\sum_{j=1}^{5} 0.5} = \frac{(2 - 1)^2 + (0 - 1)^2 + (-1)^2 + (0 - 1)^2 + (2 - 1)^2}{5} = 0.8
\]

For the second cluster, we get \((\sigma_{11}^2)^2 = 5\) and \((\sigma_{22}^2)^2 = 0.8\)

Therefore, the variance matrix for \(C_1\) is \(\Sigma_1 = \begin{bmatrix} 5 & 0 \\ 0 & 0.8 \end{bmatrix}\)

and the variance matrix for \(C_2\) is \(\Sigma_2 = \begin{bmatrix} 5 & 0 \\ 0 & 0.8 \end{bmatrix}\)

\[P(C_1) = \sum_{i=1}^{5} 0.5 = 0.5\]

Similarly, \(P(C_2) = 0.5\)

Iteration 2: E step:
To compute \(P(C_i|x_j)\), we first compute
\[
|\Sigma_1| = 4 \quad \text{and} \quad \Sigma_1^{-1} = \begin{pmatrix} 0.2 & 0 \\ 0 & 1.25 \end{pmatrix}
\]

Let us compute \(f_1(x_1)\)
\[
= (2\pi^{-\frac{1}{2}})(0.5) \exp \left\{ -\frac{1}{2} (2.5)^2 \right\} = 0.0202
\]

where, \(\left( \begin{array}{c} 2.5 \\ 1 \end{array} \right)^T \begin{pmatrix} 0.2 & 0 \\ 0 & 1.25 \end{pmatrix} \left( \begin{array}{c} 2.5 \\ 1 \end{array} \right) = 2.5\) and \(1/\sqrt{|\Sigma_1|} = 0.5\).

Similarly, we compute \(f_2(x_1)\), which gives the same value as \(f_1(x_1)\).

<table>
<thead>
<tr>
<th>(x_j)</th>
<th>(f_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_1)</td>
<td>0.26</td>
</tr>
<tr>
<td>(x_2)</td>
<td>0.26</td>
</tr>
<tr>
<td>(x_3)</td>
<td>0.32</td>
</tr>
<tr>
<td>(x_4)</td>
<td>0.026</td>
</tr>
<tr>
<td>(x_5)</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Table 1: Result of \(f\) after one iteration

The pdf \(x_1\) is a mixture model of the 2 cluster Gaussians: \(f(x_1) = \sum_{i=1}^{2} f_i(x)P(C_i = 0.5)\)

Similarly, for different points \((j=1 \text{ to } 5)\), Table 2 lists the values.

| \(x_j\) | \(P(C_1|x_j)\) | \(P(C_2|x_j)\) |
|-------|----------------|----------------|
| \(x_1\) | 0.5            | 0.5            |
| \(x_2\) | 0.5            | 0.5            |
| \(x_3\) | 0.5            | 0.5            |
| \(x_4\) | 0.5            | 0.5            |
| \(x_5\) | 0.5            | 0.5            |

Table 2: Result of EM algorithm after one iteration
Note that this question highlights a potential drawback of EM algorithm. EM can get stuck when there are uniform posterior probability. Thus, EM cannot improve with iterations. If we had provided different initial posterior probabilities like 0.6 and 0.4, we would have observed improvements with iterations.

b) Let $D(x)$ be the shortest distance from $x$ to the already-selected centroid $x_2$. Choose next centroid to be $x'$ with probability

$$
\frac{D(x')^2}{\sum_{x \in X} D(x)^2}
$$

<table>
<thead>
<tr>
<th>dist($x_2$)</th>
<th>$\frac{D(x')^2}{\sum_{x \in X} D(x)^2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>2</td>
</tr>
<tr>
<td>$x_3$</td>
<td>2.69</td>
</tr>
<tr>
<td>$x_4$</td>
<td>5</td>
</tr>
<tr>
<td>$x_5$</td>
<td>5.38</td>
</tr>
</tbody>
</table>

Table 3: Choose initial centroid using $k$-means ++

The maximum probability is associated with $x_5$. Thus, the second centroid is $x_5$

Compute the distance of each point from $\mu_1$ and $\mu_2$. Assign each cluster to the closest centroid, giving $C_1=[x_1, x_2, x_3]$ and $C_2=[x_4, x_5]$.

On performing $k$-means with this initialization, we find that the final clusters remain the same i.e. $C_1=[x_1, x_2, x_3]$ and $C_2=[x_4, x_5]$.

c) Since $X$ is assumed to follow a quite convenient assumption, we can write the observed data $Y$ in terms of a deterministic function of $X$, i.e.

$$
Y = f(X) = [X_1 + X_2, X_3, X_4, X_5].
$$

Then the likelihood of a realization $x$ of the data is

$$
\Pr(X = x \mid \theta) = \frac{n!}{\prod_{i \in [5]} x_i!} \left( \frac{1}{2} \right)^{x_1} \left( \theta \right)^{x_2+x_5} \left( \frac{1 - \theta}{2} \right)^{x_3+x_4}
$$

For EM, we must maximize the $Q$-function:

$$
\theta_{t+1} = \arg \max_{\theta \in (0,1)} Q(\theta \mid \theta^t)
= \arg \max_{\theta \in (0,1)} E_{X \mid y, \theta} [\log p(X \mid \theta)]
$$

To solve the above equation, we need the terms $p(X \mid \theta)$ that depend on $\theta$ and by taking the logarithm we get by linearity of expectation

$$
\theta_{t+1} = \arg \max_{\theta \in (0,1)} E_{X \mid y, \theta} [(X_2 + X_5) \log \theta + (X_3 + X_4) \log(1 - \theta)]
= \arg \max_{\theta \in (0,1)} \log \theta \left( E_{X \mid y, \theta} [X_2] + E_{X \mid y, \theta} [X_5] \right) + \log(1 - \theta) \left( E_{X \mid y, \theta} [X_3] + E_{X \mid y, \theta} [X_4] \right).
$$

Now, we need the expectation of $X$ conditioned on an already known histogram $y$. By assumption we know $y_1 = X_1 + X_2$, we can say that given $y_1$, the pair $X_1, X_2$ is binomially distributed

$$
\Pr(X = x \mid y, \theta) = \frac{y_1!}{x_1!x_2!} \left( \frac{2}{2 + \theta} \right)^{x_1} \left( \frac{\theta}{2 + \theta} \right)^{x_2} 1_{x_1+x_2=y_1} \prod_{i=3}^{5} 1_{x_i=y_{i-1}}.
$$
Note: It is not necessary to provide this equation. The crucial part is to recognize that it is binomially distributed. This is a binomial distribution for events $X_1$ and $X_2$, by taking the mean of the binomial distribution we get
\[
E_{X|y, \theta}[X] = \left[ \frac{2}{2+\theta} y_1, \frac{2}{2+\theta} y_1, y_2, y_3, y_4 \right].
\]
By substituting $E_{X|y, \theta}[X]$ in Eq. ??, we get the the EM-estimator for $\theta^{t+1}$, which is
\[
\theta^{t+1} = \arg \max_{\theta \in (0, 1)} \log \left( \frac{\theta y_1}{2 + \theta} + y_4 \right) + \log(1 - \theta) (y_2 + y_3)
\]
To find the maximizer, we take the derivative w.r.t. $\theta$
\[
\frac{d}{d\theta} \log \left( \frac{\theta y_1}{2 + \theta} + y_4 \right) + \log(1 - \theta) (y_2 + y_3) = \frac{\theta y_1}{2 + \theta} + y_4 - \frac{y_2 + y_3}{1 - \theta}
\]
Calculating the root of the above equation yields
\[
y_2 + y_3 + \frac{\theta y_1}{2 + \theta} + y_4 = \frac{\theta y_1}{2 + \theta} + y_4
\]
We randomly initialize $\theta^0 \in (0, 1)$ and iterating the algorithm above until convergence. The final estimator is called $\hat{\theta}_{EM}$.

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

Solution. For a cluster $C_j$, lets look at the derivative of the objective function w.r.t $\nu_j$.
\[
\frac{\partial}{\partial \nu_j} \sum_{x_i \in C_j} |\nu_j - x_i| = \sum_{x_i \in C_j} \frac{(\nu_j - x_i)}{\sqrt{(x_i - \nu_j)^2}}
\]
The derivative can take the following values:

$$\frac{\partial}{\partial \nu_j} \sum_{x_i \in C_j} |\nu_j - x_i| = \begin{cases} 
+1 & \text{if } \nu_j > x_i \\
-1 & \text{if } \nu_j < x_i \\
\text{undefined} & \text{if } \nu_j = x_i 
\end{cases}$$

The objective function is not differentiable when $\nu_j = x_i$ since the derivative becomes undefined. For all other values the derivative is either +1 (when $\nu_j > x_i$) or -1 when ($\nu_j < x_i$). Hence, the derivative indicates how many $x_i$’s are smaller than $\nu_j$.

To minimize the sum of absolute errors, we need to find the value of $\nu_j$ for which the derivative takes the value zero. It can do so if there are equal number of $x_i$’s that are smaller and larger than $\nu_j$ (for even number of $x_i$’s). If there is an odd number of $x_i$’s then the derivative is -1 left of the median value and +1 right of it, hence $\nu_j$ should be the median points in $C_j$. 

5 of 7
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 4. Compute the hierarchical clustering using the average link method. Show the distance thresholds.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Distance Matrix

Solution.

(a) and (b) (See figure)

(c) average linkage:

Initially, there are five clusters with one element each, namely, a, b, c, d, e. The next step is to find the clusters with closest distance pair. We can do this by looking up the table, ab (avg distance 2) and cd (avg distance 2). Now, we have three clusters ab, cd, e. The next step is to find the distance between \( \text{dist}_{avg}(ab, e) \) and \( \text{dist}_{avg}(ab, cd) \) and \( \text{dist}_{avg}(cd, e) \).

\[
\text{dist}_{avg}(ab, e) = \frac{\text{dist}(a, e) + \text{dist}(b, e)}{2} = \frac{8 + 6}{2} = 7
\]

\[
\text{dist}_{avg}(ab, cd) = \frac{\text{dist}(a, c) + \text{dist}(a, d) + \text{dist}(b, c) + \text{dist}(b, d)}{4} = \frac{6 + 4 + 6 + 4}{4} = 5
\]

\[
\text{dist}_{avg}(cd, e) = \frac{\text{dist}(c, e) + \text{dist}(d, e)}{2} = \frac{6 + 10}{2} = 8
\]

The minimum distance is 2.5 for ab, cd. Thus, ab and cd will be clustered next. Finally, we have abcd and e as the two clusters.

\[
\text{dist}_{avg}(abcd, e) = \frac{\text{dist}(a, e) + \text{dist}(b, e) + \text{dist}(c, e) + \text{dist}(d, e)}{4} = \frac{8 + 6 + 6 + 10}{4} = 7.5
\]
Problem 4 (Distance). Consider the data in Figure 2. Answer to the following questions assuming that we are using Euclidean distance, that $\varepsilon = 2$, and $\text{minpts} = 3$.

![Figure 2: Points in a space](image)

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.

Solution. Intuitively, if one point have more than $\text{minpts}$ ($\text{minpts} = 3$) neighbours (including itself) less than distance of $\varepsilon$ ($\varepsilon = 2$) from it, the point will be a core point. If the point is not a core point, and has one core point neighbour within $\varepsilon$ distance is called border point. Other points called noise points. A point (core point or border point) is directly density-reachable from a core point, if it is included within $\varepsilon$ distance from the core point (like $a$ and $d$). Likewise, density-reachable means the point (core point or border point) cannot be directly included by the core point, but can be reached after several passed (like $o$ and $i$). Meanwhile, density-connected means two border points are included by mutually density-reachable core points.

a) The core points are: $a$, $b$, $c$, $d$, $h$, $i$, $j$, $k$, $n$, $o$, $p$, $q$, $r$.

b) $r$ is directly density-reachable from $q$.

c) $f$ is not density-reachable from $a$. The chain is $a$, $c$, $d$ and breaks at $d$.

d) $g$ is density-connected to $r$. The intermediate point is $o$. 