CMSC 25400: HW1
Sohini Upadhyay

Exercise 1. Average square distance distortion

$$J_{avg^2} = \sum_{j=1}^{k} \sum_{x \in C_j} d(x, m_j)^2 \tag{1}$$

Intra-cluster sum of squared distances,

$$J_{IC} = \sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{x \in C_j} \sum_{x' \in C_j} d(x, x')^2$$

(a) *Given that in k-means,* $m_j = \frac{1}{|C_j|} \sum_{x \in C_{[j}} x$, *show that* $J_{IC} = 2J_{avg^2}$
*See next page***

Solution.

$$
\begin{aligned}
J_{IC} &= \sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{x \in C_j} \sum_{x' \in C_j} d(x, x')^2 \\
&= \sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{x \in C_j} \sum_{x' \in C_j} \|x - x'\|^2 \\
&= \sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{x \in C_j} \sum_{x' \in C_j} (\|x\|^2 - 2\|x\|\|x'\| + \|x'\|^2) \\
&= \sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{x \in C_j} (\sum_{x' \in C_j} \|x\|^2 - 2 \sum_{x' \in C_j} \|x\|\|x'\| + \sum_{x' \in C_j} \|x'\|^2) \\
&= \sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{x \in C_j} (C_j \|x\|^2 - 2 \sum_{x' \in C_j} \|x\|\|x'\| + \sum_{x' \in C_j} \|x'\|^2) \\
&= \sum_{j=1}^{k} \frac{1}{|C_j|} (C_j \sum_{x \in C_j} \|x\|^2 - 2 \sum_{x \in C_j} \sum_{x' \in C_j} \|x\|\|x'\| + \sum_{x \in C_j} \sum_{x' \in C_j} \|x'\|^2) \\
&= \sum_{j=1}^{k} \frac{1}{|C_j|} (C_j \sum_{x \in C_j} \|x\|^2 - 2 \sum_{x \in C_j} \sum_{x' \in C_j} \|x\|\|x'\| + C_j \sum_{x' \in C_j} \|x'\|^2) \\
&= \sum_{j=1}^{k} \frac{1}{|C_j|} (2C_j \sum_{x \in C_j} \|x\|^2 - 2 \sum_{x \in C_j} \sum_{x' \in C_j} \|x\|\|x'\|) \\
&= 2 \sum_{j=1}^{k} \frac{1}{|C_j|} (C_j \sum_{x \in C_j} \|x\|^2 - \sum_{x \in C_j} \sum_{x' \in C_j} \|x\|\|x'\|) \\
&= 2 \sum_{j=1}^{k} (\sum_{x \in C_j} \|x\|^2 - \frac{1}{|C_j|} \sum_{x \in C_j} \frac{1}{|C_j|} \sum_{x' \in C_j} \|x\|\|x'\|) \\
&= 2 \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \frac{1}{C_j} \sum_{x' \in C_j} x'\|^2 \\
&= 2 \sum_{j=1}^{k} \sum_{x \in C_j} \|x - m_j\|^2 \\
&= 2 J_{avg^2}
\end{aligned}
$$

(b) *Let $\gamma_i \in \{1, 2, \ldots, k\}$ be the cluster that the ith datapoint is assigned to, and assume that there are n points in total, $x_1, x_2, \ldots, x_n$. Then (1) can be written as*

$$J_{avg^2}(\gamma_1, \ldots, \gamma_n, m_1, \ldots, m_k) = \sum_{i=1}^{n} d(x_i, m_{\gamma_i})^2 \tag{2}$$

Recall that k-means clustering alternates the following two steps:
1. Update the cluster assignments:

$$\gamma_i \leftarrow \arg\min_{j \in \{1,2,\ldots,k\}} d(x_i, m_j) \text{ for } i = 1, 2, \ldots, n.$$

2. Update the centroids:

$$m_j \leftarrow \frac{1}{|C_j|} \sum_{i:\gamma_i = j} x_i \text{ for } j = 1, 2, \ldots, k.$$

Show that the first of these steps minimizes (2) as a function of $\gamma_1, \ldots, \gamma_n$, while holding $m_1, \ldots, m_k$ constant, while the second step minimizes it as a function of $m_1, \ldots, m_k$, while holding $\gamma_1, \ldots, \gamma_n$ constant.

Solution. Consider some x' in the dataset.

$$\begin{aligned}\arg\min_{j \in \{1,2,\ldots,k\}} d(x', m_j) &= min(d(x', m_{\gamma_1}), \ldots, d(x', m_{\gamma_n})) \\ &= d(x_i, m_{\gamma_l}) \text{ for some } \gamma_l \in \{\gamma_1, \ldots, \gamma_n\}\end{aligned}$$

Thus step 1 is entirely contingent on γ_l, minimizing (2) as a function of $\gamma_1, \ldots, \gamma_n$, while holding $m_1, \ldots, m_k$ constant.

Step 2 takes the average of the x_i in each cluster γ_i and assigns it too m_i. Whereas γ_i remains constant through this process, m_i changes. Thus the second step minimizes (2) as a function of $m_1, \ldots, m_k$, while holding $\gamma_1, \ldots, \gamma_n$ constant.

(c) *Prove that as k-means progresses, the distortion decreases monotonically iteration by iteration.*

Solution. By part (b), one step of k-means repeatedly minimizes distortion with respect to γ while holding m constant, and the other step minimizes distortion with respect to m while holding γ fixed. This means that distortion must monotonically decrease each time these steps are performed which occurs iteration by iteration.

(d) *Give an upper bound on the maximum number of iterations required for full convergence of the algorithm.*

Solution. No configuration of clusters can be repeated until the algorithm has converged. Thus for a dataset containing n points, the upper bound on the maximum number of iterations required for full convergence is the number of ways to group n points into k clusters. This is equal to the number of ways to group n into k clusters of any size minus the number of ways to have groups with 0 elements divided by the number of ways to order k groups $= \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$.

Exercise 2. See README.txt and committed code.

(b) *Comments on the plot: After a certain number of iterations, around 5, the distortion converges.*

(c) *Runs slow so I include my output image here for reference*

Exercise 3. Recall the "Mixture of k Gaussians" model used in clustering

$$p(x, z) = \pi_z \mathscr{N}(x; \mu_z, \Sigma_z),$$

where $x \in \mathbb{R}^d$, $z \in \{1, 2, \dots, k\}$ is its cluster assignment and $\mathscr{N}(x; \mu_z, \Sigma_z)$ is the density

$$\mathscr{N}(x; \mu_z, \Sigma_z) = (2\pi)^{d/2}|\Sigma_z|^{-1/2} exp(-(-x - \mu_z)^T \Sigma_z^{-1}(x - \mu_z)/2).$$

In this question we are going to derive the EM update rules for this model under the simplifying assumption that the covariance parameter of each cluster is the identity, I. The parameters of this restricted model are $\theta = (\pi_1, ..., \pi_k, \mu_1, ..., \mu_k)$.

(a) *Let* $\{(x_1, z_1), (x_2, z_2), \dots, (x_n, z_n)\}$ *be an* n *point sample from this model. Write down the corresponding log-likelihood* $l(\theta)$*.*

Solution. For a single point,

$$\begin{aligned} l(\theta) &= log(\pi_z) - \frac{d}{2}log(2\pi) - \frac{1}{2}log(|\Sigma_z|) - \frac{1}{2}(x - \mu_z)^T \Sigma_z^{-1}(x - \mu_z) \\ &= log(\pi_z) - \frac{1}{2}log(|\Sigma_z|) - \frac{1}{2}(x - \mu_z)^T \Sigma_z^{-1}(x - \mu_z) + \text{ constant.} \end{aligned}$$

Then for an n-point sample

$$
\begin{aligned}
l(\theta) &= \sum_{i=1}^{n} log(\pi_{z_i}) - \frac{n}{2}log(|\Sigma_{z_i}|) - \sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{2}(x_j-\mu_{z_i})^T\Sigma_{z_i}^{-1}(x_j-\mu_{z_i}) + \text{ constant.}\\
&= \sum_{i=1}^{n} log(\pi_{z_i}) - \frac{n}{2}log(|\Sigma_{z_i}|) - \sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{2}(x_j-\mu_{z_i})^T(x_j-\mu_{z_i}) + \text{ constant because } \Sigma_z = \mathbb{I}\\
&= \sum_{i=1}^{n} log(\pi_{z_i}) - \frac{n}{2}log(1) - \sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{2}(x_j-\mu_{z_i})^T(x_j-\mu_{z_i}) + \text{ constant}\\
&= \sum_{i=1}^{n} log(\pi_{z_i}) - \sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{2}(x_j-\mu_{z_i})^T(x_j-\mu_{z_i}) + \text{ constant}
\end{aligned}
$$

(b) *Let $p_{i,j}$ be the probability $p(z_i = j \mid x_i)$ of the ith data point coming from the jth cluster (given that its position is x_i). Derive an expression for this probability.*

Solution.

$$
\begin{aligned}
p_{i,j} = p(z_i = j \mid x_i) &= \frac{p(z_i = j \mid x_i)}{\sum_j p(z_i = j \mid x_i)}\\
&= \frac{\pi_j \mathscr{N}(x_i;\mu_j,\Sigma_j)}{\sum_{j'} \pi_{j'}\mathscr{N}(x_i;\mu_{j'},\Sigma_{j'})}
\end{aligned}
$$

(c) *The expectation maximization algorithm updates the parameters of th mixture by maximizing $l_{\theta_{old}}(\theta)$ in terms of these $p_{i,j}$ conditional probabilities. Here the expectation is taken over the values of the hidden variables $(z_i,\dots,z_n)$, and the subscript θ_{old} signifies that the $p_{i,j}$s are computed with respect to the old values of the parameters, whereas l itself is a function of the new values of the parameters.*

Solution. $E(l(\theta)) = \sum_\theta l(\theta)P(\theta \mid x_i)$, the products of parts (a) and (b) summed over the parameters of θ. Thus

$$
l_{\theta_{old}}(\theta) = \sum_j [\sum_i log(\pi_{z_i}) - \sum_j\sum_i \frac{1}{2}(x_j-\mu_{z_i})^T(x_j-\mu_{z_i})]p(z_i \mid x_j) \quad = \sum_j [\sum_i log(\pi_{z_i}) - \sum_j\sum_i \frac{1}{2}\|x_j - \mu
$$

(d) *Derive the update rule for π_j using the constraint that $\sum_{i=1}^{n}\pi_{z_i} = 1$.*

Solution. Using this constraint, the Lagrangian is

$$
\mathscr{L} = l_{\theta_{old}}(\theta) - \lambda(\sum_{i=1}^{n}\pi_{z_i} - 1)
$$

Plugging in $l_{\theta_{old}}$ from (c), and differentiating $\mathscr{L}$ with respect to π_{z_i}

$$\mathscr{L}' = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{z_i}} p_{i,j} - \lambda$$

Setting this to 0 reveals

$$\begin{aligned}
\lambda &= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{z_i}} p_{i,j} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{z_i}} \frac{\pi_{z_i} \mathscr{N}(x_j; \mu_{z_i}, \Sigma_{z_i})}{\Sigma_i \pi_{z_i} \mathscr{N}(x_j; \mu_{z_i}, \Sigma_{z_i})} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{z_i}} \cdot \frac{\sum_i \pi_{z_i} \mathscr{N}(x_j; \mu_{z_i}, \Sigma_{z_i})}{\sum_i \pi_{z_i} \mathscr{N}(x_j; \mu_{z_i}, \Sigma_{z_i})} \text{ by distributing the summations} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{z_i}}
\end{aligned}$$

Because $\sum_{i=1}^{n} \pi_{z_i} = 1$, $\sum_{i=1}^{n} \frac{1}{\pi_{z_i}} = \frac{1}{\sum_{i=1}^{n} \pi_{z_i}} = \frac{1}{1} = 1$. This implies that

$$\begin{aligned}
\lambda &= \sum_{j=1}^{n} 1 \\
&= n
\end{aligned}$$

Then the derivative of $\mathscr{L}$ with respect to π_{z_i} is

$$\mathscr{L}' = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{z_i}} p_{i,j} - n.$$

Setting this equal to zero yields

$$n = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{z_i}} p_{i,j}$$

which rearranges to the desired result, $\pi_{z_i} \leftarrow \frac{1}{n} \sum_{i=1}^{n} p_{i,j}$.

(e) *Derive the update rule for the $\mu_1, \mu_2, \ldots, \mu_k$ location parameters.*

Solution. The derivative of $l_{\theta_{old}}(\theta)$ with respect to μ_{z_i} is

$$l'_{\theta_{old}} = \sum_{i} \sum_{j} 2\|x_j - \mu_{z_i}\| p_{i,j}$$

Setting this equal to 0 yields,

$$0 = \sum_i \sum_j 2\|x_j - \mu_{z_i}\| p_{i,j}$$
$$= \|sum_i \sum_j x_j p_{i,j} - sum_i \sum_j \mu_{z_i} p_{i,j}\|.$$

So the derivative is 0 only when

$$sum_i \sum_j x_j p_{i,j} = sum_i \sum_j \mu_{z_i} p_{i,j}$$

Rearranging yields,

$$\mu_{z_i} \leftarrow \frac{\sum_i x_j p_{i,j}}{\sum_i p_{i,j}}$$

(f) *Compare these update rules to the k?means update rules derived in Question 1.*

Solution. These update rules parallel the second step in k-means where the centroid is set to the mean of the points in its cluster.