Problem 1 Kmeans Clustering **(32 points)**

In Lecture 15 (slide 5), we describes the loss function of k-means as follows -

$$F(\{\gamma_{nk}\}, \{\mu_l\}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} ||x_n - \mu_k||_2^2$$

Refer to slide 10 of lecture 15 for the update rule of k-means.

1.1 Show that the loss function decreases in each iteration of K-means. **(8 points)**

Initialize centers as $\left\{\mu_k^{(1)}\right\}\ \forall\ k\ \in\ [K]$
Given centers $\left\{\mu_k^{(t)}\right\}$ *compute* $\left\{\gamma_{nk}^{t+1}\right\}$

let γ_{nk}^{t} *be the computed assignments for each data point* x_n *in the previous iteration and*

γ_{nk}^{t+1} *be the new assignment obtained from* $\gamma_{nk}\ =\ I\Big[k\ ==\ argmin_c ||x_n - \mu_c||_2^2\Big]$.

assign each x_n *to the closet* μ_k *by* $\gamma_{nk}\ =\ I\Big[k\ ==\ argmin_c ||x_n - \mu_c||_2^2\Big]$.

the change in the loss function after iteration t will be

$$\nabla_{\gamma_{n,k}} = F\Big(\left\{\gamma_{nk}^{t+1}\right\},\ \{\mu_k\}\Big)\ -\ F\Big(\left\{\gamma_{nk}^{t}\right\},\ \{\mu_k\}\Big)\ \ =$$
$$\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}^{t+1} ||x_n - \mu_k^t||_2^2\ -\ \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}^{t} ||x_n - \mu_k^t||_2^2\ \leq\ 0$$

The inequality above holds because x_n will either belong to the same cluster as in the previous iteration, implying that $F\Big(\left\{\gamma_{nk}^{t+1}\right\},\ \{\mu_k\}\Big)\ -\ F\Big(\left\{\gamma_{nk}^{t}\right\},\ \{\mu_k\}\Big)\ = 0$ or x_n will belong to a different cluster, implying that $F\Big(\left\{\gamma_{nk}^{t+1}\right\},\ \{\mu_k\}\Big)\ -\ F\Big(\left\{\gamma_{nk}^{t}\right\},\ \{\mu_k\}\Big)\ \leq 0$. Since in the new cluster the Euclidean distance is minimized or remains the same and the loss function must decrease.

Similarly,

$$\nabla_{\mu_k} = F\Big(\{\gamma_{nk}\},\ \{\mu_k^{t+1}\}\Big) - F\Big(\{\gamma_{nk}\},\ \{\mu_k^{t}\}\Big) =$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}||x_n - \mu_k^{t+1}||_2^2 - \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}||x_n - \mu_k^{t}||_2^2 \leq 0$$

This hold because the update of $\mu_k^{t+1} = \dfrac{\sum_n \gamma_{n,k} x_n}{\sum_n \gamma_{n,k}}$. μ_k^{t+1} is the new center and the loss function still must remain the same or the loss function must decrease.■

1.2 Argue that k-means converges in a finite number of iterations. **(8 points)**

Since the loss function is being minimized at every iteration every iteration is upper bounded by the previous iteration (see result of 1.1). There will only exist an infinite amount of iterations if the loss function can increase from iteration to some other iteration. However, since this is impossible it implies that the loss function will converge to some minimum value in a finite number of iterations.The number of possible assignments for the data points is K^N, which is large but finite.

1.3 We update our loss function to use l_1 norm instead of l_2 norm such that

$$F(\{\gamma_{nk}\},\{\mu_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}|x_n - \mu_k|$$

How does the performance of this compare to the original loss function for the case of data consisting of outliers? **(4 points)**

The performance of the loss function using L1 norm in the case of data consisting of outliers performs better than if the L2 norm were to be used. The L1 case uses the median as the choice of the center and the L2 uses the average. Since outliers would affect the average of L2 the L1 would be more robust against outliers since by definition the median is more robust to outliers in a dataset.

1.4 Show that cluster centers are medians of the points in the cluster. **(12 points)**

$$(1)\ \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}|x_n-\mu_k|$$

Normally we would take the derivitive of eqn (1) but since the derivative doesnt exist for |x| (L1 norm) at zero we must come up with another method.

Another method to find $\nabla\mu_k$ is for a single cluster $k \in [K]$, lets call this cluster i, with fixed $\gamma_{n,k}$, and include only the data points that correspond to a single cluster.
thus, we obtain the expression that model this siutation as

$$\mu_k = argmin_{\mu_k}\sum_{n=1}^{N}\gamma_{nk=i}|x_n-\mu_{k=i}|$$

To elimate the expression for the L1 norm ($|x_n-\mu_{k=i}|$) we are going to take the seperate the data points in one specific cluster that are smaller and larger than some $\mu_{k=i}$.

We can write an expression for the points in cluster i as

$$\left[\sum_{n=1}^{N}\gamma_{nk=i}|x_n-\mu_{k=i}|\ \ s.t.\ x_n \le \mu_{k=i}\right]+\left[\sum_{n=1}^{N}\gamma_{nk=i}|x_n-\mu_{k=i}|\ \ s.t.\ x_n \ge \mu_{k=i}\right]$$

we can now drop the absolute value symbols if we introduce new symbols to seperate the data properly ($x_n \le \mu_{k=i} \le x_n$)

$$\text{thus, } (2)\ \left[\sum_{n=1}^{N}\gamma_{nk=i}\,(x_n-\mu_{k=i})\ \ s.t.\ x_n \le \mu_{k=i}\right]+\left[\sum_{n=1}^{N}\gamma_{nk=i}(x_n-\mu_{k=i})\ s.t.\ x_n \ge \mu_{k=i}\right]$$

The derivative of this expression is simply +1 and -1 depending on if the data point is less than or greater than $\mu_{k=i}$

The summation symbols will count the number of data points in each bracket when we take the derivitive.

Since when we take the derivative we must set the expression (2) equal to 0, the derivative of the L1 norm will only be computable when the counts in each bracket have the same number of data points. In other words this is the definition of a median. Thus the center of the cluster μ_k is a median. ■

Problem 2 Gaussian Mixture Model **(32 points)**

Let $X_1, ..., X_n \in \mathbb{R}^d$ be independent, identically distributed points sampled from a mixture of two normal (Gaussian) distributions. They are drawn independently from the probability distribution function (PDF) $p(x) = \theta N_1(x) + (1-\theta)N_2(x)$, where $N_1(x) = \frac{1}{(\sqrt{2\pi})^d} e^{-\|x-\mu_1\|^2/2}$, and $N_2(x) = \frac{1}{(\sqrt{2\pi})^d} e^{-\|x-\mu_2\|^2/2}$ are the PDFs for the isotropic multivariate normal distributions $\mathcal{N}(\mu_1, 1)$ and $\mathcal{N}(\mu_2, 1)$, respectively. The parameter $\theta \in (0,1)$ is called the mixture proportion. In essence, we flip a biased coin to decide whether to draw a point from the first Gaussian (with probability θ) or the second (with probability $1-\theta$).

Each data point is generated as follows. First draw a random Z_i, which has value 1 with probability θ, and has value 2 with probability $1-\theta$. Then, draw $X_i \sim \mathcal{N}(\mu_{Z_i}, 1)$. Our learning algorithm gets X_i as an input, but does not know Z_i .

Our goal is to find the maximum likelihood estimates of the three unknown distribution parameters $\theta \in (0,1), \mu_1 \in R^d$, and $\mu_2 \in R^d$ from the sample points $X_1, ..., X_n$. Unlike MLE for one Gaussian, it is not possible to give explicit analytic formulas for these estimates. Instead, we develop a variant of k-means clustering which (often) converges to the correct maximum likelihood estimates of $\theta, \mu_1, and \mu_2$. This variant doesn't assign each point entirely to one cluster; rather, each point is assigned an estimated posterior probability of coming from normal distribution 1.

2.1 Let $\tau_i = P(Z_i = 1|X_i)$. That is, τ_i is the posterior probability that point X_i has $Z_i = 1$. Use Bayes' Theorem to express τ_i in terms of $X_i, \theta, \mu_1, \mu_2$, and the Gaussian PDFs $\mathcal{N}_1(x)$ and $\mathcal{N}_2(x)$. To help you with question 2.3, also write down a similar formula for $1-\tau_i$, which is the posterior probability that $Z_i = 2$. **(6 points)**

$$\tau_i = P(Z_i = 1 \mid X_i) = \frac{P(z_i = 1)(P(X_i \mid Z_i = 1))}{p(x)} = \frac{\omega_k N(\mu_{Z_1}, 1)}{p(x)} = \frac{\theta N(\mu_{Z_1}, 1)}{\theta N_1(x) + (1-\theta)N_2(x)}$$

$$= \frac{\theta N(\mu_{Z_1}, 1)}{\theta\left(\frac{1}{\sqrt{2\pi}^d}\right)e^{\frac{-||x-\mu_1||^2}{2}} + (1-\theta)\left(\frac{1}{\sqrt{2\pi}^d}\right)e^{\frac{-||x-\mu_2||^2}{2}}}$$

$$= \frac{\theta N_1(X_i)}{\theta N_1(x) + (1-\theta)N_2(x)}$$

$$1 - \tau_i = P(Z_i = 2 \mid X_i) = \frac{P(z_i = 2)(P(X_i \mid Z_i = 2))}{p(x)} = \frac{\omega_k N(\mu_{Z_2}, 1)}{p(x)}$$

$$= \frac{(1-\theta)N(\mu_{Z_2}, 1)}{\theta N_1(x) + (1-\theta)N_2(x)} = \frac{(1-\theta)N_2(X_i)}{\theta\left(\frac{1}{\sqrt{2\pi}^d}\right)e^{\frac{-||x-\mu_1||^2}{2}} + (1-\theta)\left(\frac{1}{\sqrt{2\pi}^d}\right)e^{\frac{-||x-\mu_2||^2}{2}}}$$

$$= \frac{(1-\theta)N_2(X_i)}{\theta N_1(x) + (1-\theta)N_2(x)}$$

2.2 Write down the log-likelihood function, $l(\theta, \mu_1, \mu_2; X_1, ..., X_n) = \ln p(X_1, ..., X_n)$, as a summation. Note: it doesn't simplify much.

(5 points)

$$l(\theta,\ \mu_1,\ \mu_2\ ;\ X_1,....,X_n)\ =\ ln\ p(X_1,....\ X_n)\ =\ ln\ \prod_{i=1}^{N} P(x_i)$$

$$=\ \sum_{i=1}^{N} ln[\theta N_1(\ X_i)\ +\ (1-\theta)N_2(\ X_i)]$$

2.3 Express $\frac{\partial l}{\partial \theta}$ in terms of θ and $\tau_i, i \in \{1, ..., n\}$ and simplify as much as possible. There should be no normal PDFs explicitly in your solution, though the τ_i's may implicitly use them. Hint: Recall that $(\ln f(x)) = \frac{f(x)}{f(x)}$).

(5 points)

$$L\ = ln[\theta N_1(\ X_i)\ +\ (1-\theta)N_2(\ X_i)]\ ;\ \ \tau_i\ =\ \frac{\theta N_1(\ X_i)}{\theta N_1(\ X_i)\ +\ (1-\theta)N_2(\ X_i)}$$

$$1-\tau_i\ =\ \frac{(1-\theta)N_2(\ X_i)}{\theta N_1(\ X_i)\ +\ (1-\theta)N_2(\ X_i)}$$

$$\frac{\partial L}{\partial \theta}\ =\ \sum_{i=1}^{N} \frac{N_1(\ X_i) - N_2(\ X_i)}{\theta N_1(\ X_i)\ +\ (1-\theta)N_2(\ X_i)}$$

$$=\ \sum_{i=1}^{N} \frac{N_1(\ X_i)}{\theta N_1(\ X_i)\ +\ (1-\theta)N_2(\ X_i)}\ -\ \frac{N_2(\ X_i)}{\theta N_1(\ X_i)\ +\ (1-\theta)N_2(\ X_i)}$$

$$=\ \sum_{i=1}^{N} \frac{\tau_i}{\theta} - \frac{1-\tau_i}{(1-\theta)}\ \ =\ \sum_{i=1}^{N} \frac{\tau_i(1-\theta)}{\theta-\theta^2} - \frac{1-\tau_i(\theta)}{\theta-\theta^2}\ =\ \sum_{i=1}^{N} \frac{\tau_i-\theta\tau_i}{\theta-\theta^2} - \frac{\theta-\theta\tau_i}{\theta-\theta^2}$$

$$=\ \frac{1}{\theta-\theta^2}\sum_{i=1}^{N} \tau_i - \ \theta\ =\ \frac{\sum_{i=1}^{N}(\tau_i - \ \theta)}{\theta-\theta^2}$$

2.4 Express $\nabla_{\mu_1} l$ in terms of μ_1 and $\tau_i, X_i, i \in \{1, ..., n\}$. Do the same for $\nabla_{\mu_2} l$ (but in terms of μ_2 rather than μ_1). Again, there should be no normal PDFs explicitly in your solution, though the τ_i's may implicitly use them. Hint: It will help (and get you part marks) to first write $\nabla_{\mu_1} N_1(x)$ as a function of $N_1(x), x$, and μ_1.

(7 points)

$$recall: \; N_1(X_i) = \left(\frac{1}{\sqrt{2\pi}^d}\right) e^{\frac{-||x-\mu_1||^2}{2}}$$

$$Hint: \; \nabla\mu_1 N_1(X_i) = \left(\frac{1}{\sqrt{2\pi}^d}\right) e^{\frac{-||x-\mu_1||^2}{2}} * (x_i - \mu_1)$$

$$\nabla_{\mu 1} l = \sum_{i=1}^{N} \frac{\theta \nabla_{\mu_1} N_1(X_i)}{\theta N_1(X_i) + (1-\theta) N_2(X_i)} = \sum_{i=1}^{N} \frac{\theta N_1(X_i)}{\theta N_1(X_i) + (1-\theta) N_2(X_i)} (x_i - \mu_1)$$

$$= \sum_{i=1}^{N} \tau_i (x_i - \mu_1)$$

$$\nabla_{\mu 2} l = \sum_{i=1}^{N} \frac{(1-\theta) \nabla_{\mu_2} N_2(X_i)}{\theta N_1(X_i) + (1-\theta) N_2(X_i)} = \sum_{i=1}^{N} \frac{(1-\theta) N_2(X_i)}{\theta N_1(X_i) + (1-\theta) N_2(X_i)} (x_i - \mu_2)$$

$$= \sum_{i=1}^{N} (1-\tau_i)(x_i - \mu_2)$$

2.5 We conclude: if we know μ_1, μ_2, and θ, we can compute the posteriors τ_i. On the other hand, if we know the τ_i's, we can estimate μ_1, μ_2, and θ by using the derivatives in 2.3 and 2.4 to find the maximum likelihood estimates. This leads to the following k-means-like algorithm.

- Initialize $\tau_1, \tau_2, ..., \tau_n$ to arbitrary values in the range $[0, 1]$.
- Repeat the following two steps.
  1. Update the Gaussian cluster parameters: for fixed values of $\tau_1, \tau_2, ..., \tau_n$, update μ_1, μ_2, and θ.
  2. Update the posterior probabilities: for fixed values of μ_1, μ_2, and θ, update $\tau_1, \tau_2, ..., \tau_n$.

In part 2.1, you wrote the update rule for step 2. Using your results from parts 2.3 and 2.4, write down the explicit update formulas for step 1.

(9 points)

$$set \; \nabla_{\mu_1} l = 0 \quad to \; get \; the \; update \; for \; the \; \mu_1 \; variable \; and \; similarly \; for \; \mu_2 \; and \; \theta$$

(*result from* 2.4)

$$\nabla_{\mu_1} l = \sum_{i=1}^{N} \tau_i (x_i - \mu_1) = 0 \implies \sum_{i=1}^{N} \tau_i x_i - \sum_{i=1}^{N} \tau_i \mu_1 = 0 \implies \mu_1 = \frac{\sum_{i=1}^{N} \tau_i x_i}{\sum_{i=1}^{N} \tau_i}$$

(*result from* 2.4)

$$\nabla_{\mu_2} l = \sum_{i=1}^{N} (1-\tau_i)(x_i - \mu_2) = 0 \implies \sum_{i=1}^{N} (1-\tau_i) x_i - \sum_{i=1}^{N} (1-\tau_i) \mu_2 = 0$$

$$\implies \mu_2 = \frac{\sum_{i=1}^{N} (1-\tau_i) x_i}{\sum_{i=1}^{N} (1-\tau_i)}$$

(*result from* 2.3)

$$\nabla_{\theta} l = \frac{\sum_{i=1}^{N} (\tau_i - \theta)}{\theta - \theta^2} = 0 \implies \sum_{i=1}^{N} \tau_i - \sum_{i=1}^{N} \theta = 0 \implies \theta = \frac{\sum_{i=1}^{N} \tau_i}{N}$$