CIS5200: Machine Learning Spring 2025

Homework 5

Release Date: April 15, 2025 *Due Date: April 29, 2025*

Name: Haoze Wu
PennKey: haozewu
Collaborators: None

Problem 1: k-means Clustering

1.1 Answer:
For the first term:

$$Z(C_1, \cdots, C_k) = \sum_{i=1}^{k} \frac{1}{2|C_l|} \sum_{i,j \in C_l} \|x_i - x_j\|_2^2 \tag{1}$$

, we can expand the 2-norm into:

$$\|x_i - x_j\|_2^2 = \|x_i - \mu_l + \mu_l - x_j\|_2^2 = \|x_i - \mu_l\|_2^2 + \|x_j - \mu_l\|_2^2 + 2(x_i - \mu_l)^T(\mu_l - x_j) \tag{2}$$

And by intergating it into the summation, we have:

$$\sum_{i,j \in C_l} \|x_i - x_j\|_2^2 = \sum_{i,j \in C_l} \|x_i - \mu_l\|_2^2 + \sum_{i,j \in C_l} \|x_j - \mu_l\|_2^2 + 2 \sum_{i,j \in C_l} (x_i - \mu_l)^T(\mu_l - x_j) \tag{3}$$

For the last termin in equation (3), we have:

$$\sum_{i,j \in C_l} (x_i - \mu_l)^T(\mu_l - x_j) = \sum_{i \in C_l} (x_i - \mu_l)^T \sum_{j \in C_l} (\mu_l - x_j) \tag{4}$$

, and notice that $\mu_l = \frac{1}{|C_l|} \sum_{j \in C_l} x_j$, for $\sum_{i \in C_l} (x_i - \mu_l)$, we then have:

$$\sum_{i \in C_l} (x_i - \mu_l) = \sum_{i \in C_l} x_i - |C_l|\mu_l = \sum_{i \in C_l} x_i - \sum_{j \in C_l} x_j = 0 \tag{5}$$

, and it is similar to anothe term $\sum_{j \in C_l} (\mu_l - x_j)$, so we have:

$$\sum_{i,j \in C_l} (x_i - \mu_l)^T(\mu_l - x_j) = 0 \tag{6}$$

Thus, for the first equation given in this problem, we have:

$$\begin{aligned} Z(C_1, \cdots, C_k) &= \sum_{i=1}^{k} \frac{1}{2|C_l|} \sum_{i,j \in C_l} \|x_i - x_j\|_2^2 \\ &= \sum_{i=1}^{k} \frac{1}{2|C_l|} \sum_{i,j \in C_l} (\|x_i - \mu_l\|_2^2 + \|\mu_l - x_j\|_2^2) \\ &= \sum_{i=1}^{k} \frac{1}{2|C_l|} (|C_l| \sum_{i \in C_l} \|x_i - \mu_l\|_2^2 + |C_l| \sum_{j \in C_l} \|\mu_l - x_j\|_2^2) \\ &= \sum_{i=1}^{k} \frac{1}{2} (2 \sum_{i \in C_l} \|x_i - \mu_l\|_2^2) \\ &= \sum_{i=1}^{k} \sum_{i \in C_l} \|x_i - \mu_l\|_2^2 \end{aligned} \tag{7}$$

, which is equivalent to the second equation given in this problem.

Proved.

1.2 Answer:
Given the definition of goodness of clustering as

$$Z(C, z) = \sum_{i \in C} \|x_i - z\|_2^2 \tag{8}$$

to find the optimal center of the cluster C, we can take the derivative of $Z(C, z)$ with respect to z and set it to 0:

$$\frac{\partial Z(C, z)}{\partial z} = \frac{\partial}{\partial z} \sum_{i \in C} \|x_i - z\|_2^2 = \sum_{i \in C} 2(x_i - z) = 0 \tag{9}$$

, and we can get:

$$\sum_{i \in C} x_i - |C| z = 0 \tag{10}$$

, thus we have:

$$z = \frac{1}{|C|} \sum_{i \in C} x_i = \mu \tag{11}$$

, which is the average of all the points in the cluster C. Thus, the optimal center of the cluster C is the average of all the points in the cluster C, which is equivalent to

$$Z(C, \mu) = \min_z Z(C, z) \tag{12}$$

Proved.

1.3 Answer:
Given that the center is of the cluster z is uniformly sampled from the points from the cluster C, the probability of the center z to be x_i for any $x_i \in C$ is $\frac{1}{|C|}$. Thus, for the point distribution ρ, we can have:

$$E_{z\sim\rho}[Z(C,z)] = E_{z\sim\rho}[\sum_{i\in C} \|x_i - z\|_2^2] = \sum_{i\in C} E_{z\sim\rho}[\|x_i - z\|_2^2] \tag{13}$$

For any point $x_i \in C$, we can have:

$$E_{z\sim\rho}[\|x_i - z\|_2^2] = \sum_{j\in C} \frac{1}{|C|}\|x_i - x_j\|_2^2 \tag{14}$$

Thus, we then have:

$$E_{z\sim\rho}[Z(C,z)] = \sum_{i\in C}\sum_{j\in C} \frac{1}{|C|}\|x_i - x_j\|_2^2 \tag{15}$$

And, from the conclusion of problem 1.1, we have:

$$E_{z\sim\rho}[Z(C,z)] = \sum_{i\in C}\sum_{j\in C} \frac{1}{|C|}\|x_i - x_j\|_2^2 = 2\sum_{i=1}^{k} \frac{1}{2|C_l|} \sum_{i,j\in C_l} \|x_i - x_j\|_2^2 = 2Z(C,\mu) \tag{16}$$

Proved.

1.4 When using the EM algorithm to solve the Gaussian Mixture Model, we can separate it into the expectation stage and the maximiation stage.

For the expectation stage, i.e., the soft cluster assignment, we can use the contribution of the j-th cluster to the total probability of x_i as the soft assignment of x_i to the j-th cluster:

$$z_{ij} = \frac{\pi_l \mathcal{N}(x_i|\mu_l, \Sigma_l)}{\sum_{j=1}^{k} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)} \tag{17}$$

The parameter θ can then be written as $\theta = \{\pi_l, \mu_1, \cdots, \mu_k, \Sigma_1, \cdots, \Sigma_k\}$, and the soft assignment z_{ij} can be written as $z_{ij} = P(z_i = j|x_i, \theta)$. Then, for the maximiation stage, we can use the soft assignment z_{ij} to update the parameters θ, and from GMM, we have:

$$\begin{aligned}
\mu_l &= \frac{\sum_{i=1}^{m} z_{il} x_i}{\sum_{i=1}^{m} z_{il}} \\
\Sigma_l &= \frac{\sum_{i=1}^{m} z_{il}(x_i - \mu_l)(x_i - \mu_l)^T}{\sum_{i=1}^{m} z_{il}} \\
\pi_l &= \frac{\sum_{i=1}^{m} z_{il}}{m}
\end{aligned} \tag{18}$$

When we set $\Sigma_l = \sigma^2 I$ for all $l \in [k]$, the Gaussian density becomes:

$$\mathcal{N}(x_i|\mu_l, \sigma^2 I) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{|x_i - \mu_l|_2^2}{2\sigma^2}) \tag{19}$$

Then, substituting this into the expectation stage, we have:

$$\begin{aligned} z_{il} &= \frac{\pi_l \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{|x_i-\mu_l|2^2}{2\sigma^2})}{\sum j = 1^k \pi_j \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{|x_i-\mu_j|_2^2}{2\sigma^2})} \\ &= \frac{\pi_l \exp(-\frac{|x_i-\mu_l|2^2}{2\sigma^2})}{\sum j = 1^k \pi_j \exp(-\frac{|x_i-\mu_j|_2^2}{2\sigma^2})} \end{aligned} \tag{20}$$

Now, as $\sigma \to 0$, the exponential terms dominate. For any data point x_i, let $l^* = \arg\min_j |x_i - \mu_j|_2^2$ be the index of the closest centroid. Here, for $\forall j \neq l^*$, we have:

$$|x_i - \mu_j|_2^2 > |x_i - \mu_{l^*}|_2^2 \tag{21}$$

And thus, we have:

$$\lim_{\sigma \to 0} z_{il} = \begin{cases} 1 & l = l^* \\ 0 & l \neq l^* \end{cases} \tag{22}$$

, which is the hard assignment of points to clusters, and it is equivalent to the assignment of k-means where each point is assigned to its nearest centroid.

Proved.

Problem 2: PCA

2.1 Answer:
We may first use the eigenvalue decomposition of the covariance matrix Σ:

$$\Sigma = U\Lambda U^T \tag{23}$$

Also, notice that the maximum variance expression given in the problem can be written as:

$$\max_{u:\|u\|_2=1, u^\top u_1=0} \frac{1}{m-1} \sum_{i=1}^{m} (x_i^\top u)^2 = \max_{u:\|u\|_2=1, u^\top u_1=0} u^\top S u \tag{24}$$

for the second principal component u_2 since it is orthogonal to the first principal component u_1. And writing the unit vector u into a linear combination of all the eigenvectors of the covariance matrix Σ:

$$u = \sum_{i=1}^{m} \alpha_i u_i \tag{25}$$

where $\sum_{i=1}^{m} \alpha_i^2 = 1$. Then, we can have:

$$\begin{aligned} u^\top S u &= \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j u_i^\top S u_j \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j u_i^\top (U\Lambda U^T) u_j \end{aligned} \tag{26}$$

since all the eigenvectors u_i are orthogonal to each other, we can have:

$$u_i^\top U \Lambda U^T u_j = \lambda_i u_i^\top u_j \tag{27}$$

, and thus we can have:

$$\begin{aligned} u^\top S u &= \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \lambda_i u_i^\top u_j \\ &= \sum_{i=1}^{m} \alpha_i^2 \lambda_i \end{aligned} \tag{28}$$

Since we have $u^\top u_1 = 0$, which implies that $\alpha_1 = 0$, when we set $\alpha_2 = 1$ and all other $\alpha_i = 0$ for $i = 3, 4, \cdots, m$, we can have:

$$u^\top S u = \lambda_2 \tag{29}$$

, which yields the maximum variance of the second principal component u_2.

Proved.

2.2 Answer:
For the argmin term, we can have:

$$\|x_i - (u^\top x_i)u\|_2^2 = \|x_i\|_2^2 - 2(u^\top x_i)(u^\top x_i) + (u^\top x_i)^2 \|u\|_2^2 \tag{30}$$

And since we have the constraint $\|u\|_2^2 = 1$, we can have:

$$\|x_i - (u^\top x_i)u\|_2^2 = \|x_i\|_2^2 - (u^\top x_i)^2 \tag{31}$$

Thus, by summing over all the points x_i, we can have:

$$\sum_{i=1}^{m} \|x_i - (u^\top x_i)u\|_2^2 = \sum_{i=1}^{m} \|x_i\|_2^2 - \sum_{i=1}^{m} (u^\top x_i)^2 \tag{32}$$

When we try to minimize this term, we can see that the first term $\sum_{i=1}^{m} \|x_i\|_2^2$ is a constant and does not depend on u, thus we can ignore it and only need to minimize the second term $\sum_{i=1}^{m} (u^\top x_i)^2$, which is equivalent to minimize the variance of the projection of the data points x_i onto the direction u since the only difference is the constant term $\sum_{i=1}^{m} \|x_i\|_2^2$ and the coefficient $1/m$ and $1/(m-1)$. That is to say, minimizing the first term is equivalent to maximizing the second term.

Proved.

2.3 Answer:
For the term of the left hand side of the equation to be proved, we have:

$$\|Ux_i\|_2^2 = (Ux_i)^\top [Ux_i] = x_i^\top U^\top U x_i \tag{33}$$

And for the term of the right hand side of the equation to be proved, we have:

$$\begin{aligned}\|x_i - U^\top U x_i\|_2^2 &= (x_i - U^\top U x_i)^\top [x_i - U^\top U x_i] \\ &= x_i^\top x_i - 2x_i^\top U^\top U x_i + (U^\top U x_i)^\top [U^\top U x_i] \\ &= x_i^\top x_i - 2x_i^\top U^\top U x_i + x_i^\top U^\top U U^\top U x_i \\ &= x_i^\top x_i - 2x_i^\top U^\top U x_i + x_i^\top U^\top U x_i \\ &= x_i^\top x_i - x_i^\top U^\top U x_i\end{aligned} \tag{34}$$

Then, we can make a summation over all the points x_i:

$$\sum_{i=1}^{m} \|x_i - U^\top U x_i\|_2^2 = \sum_{i=1}^{m} x_i^\top x_i - \sum_{i=1}^{m} x_i^\top U^\top U x_i \tag{35}$$

since the argmax and argmin are taken over U, rather than x, thus we can treat $\sum_{i=1}^{m} x_i^\top x_i$ as a constant and ignore it, and the remaining term in both sides of the equation is equivalent: the left hand side of the equation is maximizing the term $\sum_{i=1}^{m} x_i^\top U^\top U x_i$, and the right hand side is minimizing the term$-\sum_{i=1}^{m} x_i^\top U^\top U x_i$. Thus, maximizing the variance is equivalent to minimize the reconstruction error.

Proved.