ECE421: Introduction to Machine Learning — Fall 2024

Worksheet 4: K-means Clustering and Gaussian Mixture Model

Q1 **(Lack of Optimality of K-Means)** Consider a K-Means clustering problem instance with $K = 2$ and a dataset of 4 points in $\mathbb{R}$ as follows: $x_1 = -12$, $x_2 = -3$, $x_3 = 3$, and $x_4 = 12$. Initialize K-Means with the centroids $\mu_1 = -3$ and $\mu_2 = 12$. Demonstrate that in the problem instance above, K-Means converges to a solution that is not globally optimal.

Answer. The centroids are initially set to $\mu_1[0] = -3$ and $\mu_2[0] = 12$. It can be easily verified that with these centroids, the K-means algorithm assigns $\mathcal{B}_1[1] = \{x_1, x_2, x_3\}$ and $\mathcal{B}_2[1] = \{x_4\}$. In the next step, the centroids will be updated to $\mu_1[1] = \frac{x_1+x_2+x_3}{3} = -4$ and $\mu_2[1] = x_4 = 12$. It is easy to check that the set memberships will remain unchanged in the next step. Thus, K-Means algorithm converged to the following set membership and centroids:

$$\hat{\mathcal{B}}_1 = \{x_1, x_2, x_3\}, \hat{\mathcal{B}}_2[1] = \{x_4\}, \hat{\mu}_1 = -4, \hat{\mu}_2 = 12$$

. The error at the converged solution is given by $E = (x_1 - \hat{\mu}_1)^2 + (x_2 - \hat{\mu}_1)^2 + (x_3 - \hat{\mu}_1)^2 + (x_4 - \hat{\mu}_2)^2 = 8^2 + 1^2 + 7^2 + 0 = 114$.

To demonstrate that this is not a globally optimal solution, consider the following set memberships and centroids:

$$\mathcal{B}_1^\star = \{x_1, x_2\}, \mathcal{B}_2^\star = \{x_3, x_4\}, \mu_1^\star = -7.5, \mu_2^\star = 7.5.$$

With such set memberships and centroids, the clustering error would be $E^\star = 4(4.5)^2 = 81$, which is lower than E.

Q2 **(K-means algorithm: Problem 6 - Final Exam 2018)** Consider the K-means algorithm. Let $K = 2$ and let $\mathcal{D}$ be a dataset consisting of four data points with $\mathcal{D} = \{0, 0.5, 0.5 + \Delta, 1.5 + \Delta\}$, where $\Delta \geq 0$ is a problem parameter. All data points lie on the real line.

2.a Let $\Delta = 0.5$ and initialize K-means by initializing the two cluster centers at $\mu_1 = 1$ and $\mu_2 = 2$. Run K-means till convergence. For each iteration l until convergence, describe your set membership $\{\mathcal{B}_1[l], \mathcal{B}_2[l]\}$ and cluster centers $\{\mu_1[l], \mu_2[l]\}$. Make sure you identify the final values of the cluster centers and set membership at convergence.

Answer. Observe that $\mathcal{D} = \{0, 0.5, 1, 2\}$ and the centriods are initialized as $\mu_1[0] = 1, \mu_2[0] = 2$. Thus, we would have the following iterations:

- **Iteration** 1**:** $\mathcal{B}_1[1] = \{0, 0.5, 1\}$ and $\mathcal{B}_2[1] = \{2\}$; $\mu_1[1] = 0.5$ and $\mu_2[1] = 2$.
- **Iteration** 2**:** $\mathcal{B}_1[2] = \{0, 0.5, 1\}$ and $\mathcal{B}_2[2] = \{2\}$; $\mu_1[2] = 0.5$ and $\mu_2[2] = 2$. The cluster assignments and cluster centroids do not change. Done!

2.b For this part, find a condition that Δ must satisfy, such that Δ has a small positive value, and K-means (initialized in the same manner as in **2.a**, *i.e.*, $\mu_1 = 1$ and $\mu_2 = 2$) converges to a different solution from that obtained in **2.a**. In your solution, describe:

2.b.i What is this condition on Δ and explain your reasoning/derivation.

Answer. It is easy to check that for any $\Delta > 1$, in the first iteration of K-Means algorithm, we would have $\mathcal{B}_1[1] = \{0, 0.5\}$ and $\mathcal{B}_2[1] = \{0.5 + \Delta, 1.5 + \Delta\}$; $\mu_1[1] = 0.25$ and $\mu_2[1] = 1 + \Delta$. And the set memberships and centroid would remain unchanged in the next iteration. Hence, we would converge to a different solution.

One can easily verify that if $0 < \Delta < 1$, we would always converge to $\mathcal{B}_1 = \{0, 0.5\}$ and $\mathcal{B}_2 = \{0.5 + \Delta, 1.5 + \Delta\}$. Let $x_3 = 0.5 + \Delta$. Observe that for any $0 < \Delta < 1$, $\mathcal{B}_1[1] = \{0, 0.5, 0.5 + \Delta\}, \mathcal{B}_2[1] = \{1.5 + \Delta\}$, and $\mu_1[1] = \frac{1+\Delta}{3}$ and $\mu_2[1] = 1.5 + \Delta$. It is easy to show that $|\, x_3 - \mu_1[1] \,|<|\, \mu_2[1] - x_3 \,|$, when $0 < \Delta < 1$, and we will converge to the same solution as in **2.a**.

2.b.ii As in **2.a**, run the cluster algorithm, describe the values of cluster centers and set membership for each iteration until convergence.

Answer. With $\Delta > 1$,

- **Iteration 1:** $\mathcal{B}_1[1] = \{0, 0.5\}$ and $\mathcal{B}_2[1] = \{0.5 + \Delta, 1.5 + \Delta\}$; $\mu_1[1] = 0.25$ and $\mu_2[1] = 1 + \Delta$.
- **Iteration 2:** $\mathcal{B}_1[2] = \{0, 0.5\}$ and $\mathcal{B}_2[2] = \{0.5 + \Delta, 1.5 + \Delta\}$; $\mu_1[2] = 0.25$ and $\mu_2[2] = 1 + \Delta$. Converged!

Q3 **(Gaussian Mixture Model: Problem 5 - Final Exam 2018)** Consider an already-trained Gaussian Mixture Model (GMM) that is trained to fit data on student performance in a class. The GMM uses two components ($K = 2$) as the class consists of two categories of students: undergraduate students (category 1) and graduate students (category 2). The learned parameters of the GMM are as follows.

- The weights of the two categories (*i.e.*, the responsibilities) are $\pi_1 = \frac{2}{3}$ (undergraduate) and $\pi_2 = \frac{1}{3}$ (graduate).
- The distribution that fits scores in category 1 is $\mathcal{N}\left(x; 70, 10^2\right)$.
- The distribution that fits scores in category 2 is $\mathcal{N}\left(x; 80, 5^2\right)$.

3.a According to the GMM, what is the probability that an arbitrarily selected student scores greater than 80%? That is, compute $\mathbb{P}[X \geq 80]$, where X denotes the score of the student. In your computation, use the approximation that for zero-mean σ^2-variance random variable Z, $\mathbb{P}[|Z| \leq \sigma] = \frac{2}{3}$.

Answer. Note that

$$p(x) = \pi_1 \mathcal{N}\left(x \mid 70, 10^2\right) + \pi_2 \mathcal{N}\left(x \mid 80, 5^2\right).$$

Thus,

$$\begin{aligned}
\mathbb{P}[X \geq 80] &= \pi_1 \mathbb{P}_{X \sim \mathcal{N}(70, 10^2)}[X \geq 80] + \pi_2 \mathbb{P}_{X \sim \mathcal{N}(80, 5^2)}[X \geq 80] \\
&= \frac{2}{3} \cdot \frac{1}{2}\left(1 - \mathbb{P}_{X \sim \mathcal{N}(70, 10^2)}[|X| < 80]\right) + \frac{1}{3} \cdot \frac{1}{2} \\
&= \frac{1}{3}\left(1 - \mathbb{P}_{X \sim \mathcal{N}(0, 10^2)}[|X| < 10]\right) + \frac{1}{6} \\
&\approx \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{6} = \frac{5}{18}.
\end{aligned}$$

3.b If a particular student has a score greater than 80, what is the probability that the student is from category 1? That is, compute $\mathbb{P}[\text{class} = 1 \mid X \geq 80]$. (Use the same approximation as in the previous part.)

Answer. Applying the Bayes' rule, *i.e.*, $\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$, we have

$$\begin{aligned}
\mathbb{P}[\text{class} = 1 \mid X \geq 80] &= \frac{\mathbb{P}[(X \geq 80) \mid \text{class} = 1]\mathbb{P}[\text{class} = 1]}{\mathbb{P}[(X \geq 80)]} \\
&= \mathbb{P}_{X \sim \mathcal{N}(70, 10^2)}[X \geq 80] \times \pi_1 \times \frac{18}{5} \\
&= \frac{2}{5}.
\end{aligned}$$