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.

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.

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.
- **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.

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 are $w_1 = \frac{2}{3}$ (undergraduate) and $w_2 = \frac{1}{3}$ (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, $\mathbb{P}[|Z| \leq \sigma] = \frac{2}{3}$.

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.)