Statistical Foundations of Learning - Sample Problems 6

CIT4230004 (Summer Semester 2024)

Sample Problem 6.1: Bounds for k-means++

We define $B(v, r)$ as a ball centered at $v \in \mathbb{R}^p$ with radius $r > 0$, that is,

$$B(v, r) = \{x \in \mathbb{R}^p : \|x - v\| \leq r\},$$

where $\|x - v\|$ is the Euclidean distance.

Fix $r > 0$ and let $v_1, v_2 \in \mathbb{R}^p$ be two points with $\|v_1 - v_2\| = 10r$. Suppose we have a dataset $X = \{x_1, \ldots, x_m\}$ such that $\frac{m}{2}$ points lie in $B(v_1, r)$ and the other $\frac{m}{2}$ points lie in $B(v_2, r)$, where m is even.

1. What is the maximum distance between two points in the same ball, and the minimum distance between two points in different balls?

- Maximum distance between two points in the same ball:

$$\max_{x,y \in B(v,r)} \|x - y\| \leq 2r$$

- Minimum distance between two points in different balls:

$$\min_{x \in B(v_1,r), y \in B(v_2,r)} \|x - y\| \geq \|v_1 - v_2\| - 2r = 10r - 2r = 8r$$

2. Assume that two centers are selected using the k-means++ algorithm. Show that the probability of selecting both centers from the same ball is at most $\frac{1}{16}$.

The k-means++ algorithm selects the first center uniformly at random. Without loss of generality, assume the first center is chosen from $B(v_1, r)$.

For the second center, the probability of selecting a center from the same ball $B(v_1, r)$ is:

$$\frac{1}{2}\left(\frac{1}{2}\right) = \frac{1}{4}$$

The second center is more likely to be selected from $B(v_2, r)$ due to the distance-based probability distribution.

Thus, the probability of selecting both centers from the same ball is at most:

$$\frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$$

3. Derive an upper bound on the k-means cost when k-means++ chooses both centers from the same ball, and also when both centers are chosen from different balls.

- When both centers are chosen from the same ball $B(v_1, r)$:

$$\text{k-means cost} = \sum_{x \in B(v_2,r)} \|x - v_1\|^2 \leq \frac{m}{2}(8r)^2 = 64mr^2$$

- When centers are chosen from different balls:

$$\text{k-means cost} = \sum_{x \in B(v_1,r)} \|x - v_1\|^2 + \sum_{x \in B(v_2,r)} \|x - v_2\|^2 \leq \frac{m}{2}r^2 + \frac{m}{2}r^2 = mr^2$$

4. Combine the previous steps to show that the expected cost of the k-means++ solution is smaller than $9mr^2$, where expectation is with respect to the random choice of centers.

Let $P(\text{both centers from same ball}) = \frac{1}{16}$ and $P(\text{centers from different balls}) = \frac{15}{16}$.

The expected cost is:

$$\mathbb{E}[\text{k-means cost}] = \left(\frac{1}{16} \cdot 64mr^2\right) + \left(\frac{15}{16} \cdot mr^2\right)$$

$$= 4mr^2 + \frac{15}{16}mr^2 = 4mr^2 + 0.9375mr^2 = 4.9375mr^2$$

Thus, the expected cost of the k-means++ solution is smaller than $9mr^2$.

Sample Problem 6.2: Explainable k-means cost

In this exercise, we will construct a dataset $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ such that the ratio between the optimal explainable k-means cost and the (unrestricted) k-means cost is $\Omega(k)$. For some fixed $k \in \mathbb{N}$ and a dimension $d \in \mathbb{N}$ that we will fix later, let us begin by defining the k cluster centers. To this end, consider d independent random permutations $\pi_1, \ldots, \pi_d$ of the set $[k]$. For each $i \in [k]$, choose the i-th cluster center $c_i \in \mathbb{R}^d$ as

$$c_i = \begin{pmatrix} \pi_1(i) \\ \vdots \\ \pi_d(i) \end{pmatrix}$$

Now, we assign $2d$ points to each cluster center c_i, all of them being of the form $c_i \pm e_j$ for $j \in [d]$. Thus, our dataset will have $2dk$ points in total, and since every point has a squared distance of 1 from its corresponding cluster center, the optimal k-means cost is also $2dk$. To show that the explainable k-means cost is of order $\Omega(dk^2)$, proceed as follows.

1. Bound the expected distance between any two cluster centers c_s, c_t along the j-th axis from below. Show that there exists a constant $K_1 > 0$ such that for every pair $s \neq t \in [k]$ and every dimension $j \in [d]$, we obtain

$$\mathbb{E}[|\pi_j(s) - \pi_j(t)|^2] \geq K_1 k^2$$

Solution:

For any pair $s \neq t \in [k]$ and any dimension $j \in [d]$, the distance along the j-th axis is given by:

$$\mathbb{E}[|\pi_j(s) - \pi_j(t)|^2] = \frac{k^2 - 1}{3}$$

Thus, $K_1 = \frac{1}{3}$ and:

$$\mathbb{E}[|\pi_j(s) - \pi_j(t)|^2] \geq \frac{k^2 - 1}{3} \approx K_1 k^2$$

2. Use a union bound to prove that with some positive probability independent of k, all centers are at a squared distance of at least $\Omega(dk^2)$.

Using the union bound, we sum the probabilities of each pair of centers being close:

$$\mathbb{P}(\|c_s - c_t\|^2 \leq dK_2 k^2) \leq \frac{1}{k^2}$$

There are $\binom{k}{2} \approx \frac{k^2}{2}$ pairs of centers. Thus, the probability that any pair of centers is closer than $\Omega(dk^2)$ is:

$$\binom{k}{2} \cdot \frac{1}{k^2} \leq \frac{k^2}{2} \cdot \frac{1}{k^2} = \frac{1}{2}$$

Thus, with positive probability independent of k, all centers are at a squared distance of at least $\Omega(dk^2)$.

3. Now consider explainable k-means on these centers. Argue that any decision tree will always make at least one mistake, in the sense that at least one point will not end up in the same leaf as its friends from the same cluster. Argue that this will push the k-means cost up to $\Omega(dk^2)$.

Solution:

Given that any decision tree must split the data along the coordinate axes, each split can only distinguish between points that differ in a single dimension. However, since the cluster centers c_i are constructed such that each dimension j contains a permutation of $[k]$, it is highly probable that any split will not perfectly separate the points assigned to different cluster centers.

1. **Mistakes in the decision tree:** - For each cluster center c_i, the points are $c_i \pm e_j$ for $j \in [d]$. - Any decision tree that attempts to separate these points will inevitably place some points from different clusters into the same leaf due to the permutations along each dimension. - Specifically, due to the permutations, it is likely that for any dimension j, some points $c_i \pm e_j$ and $c_s \pm e_j$ will be misclassified, as the decision tree cannot perfectly distinguish between them.

2. **Impact on k-means cost:** - The misclassified points will be assigned to the nearest centroid based on the decision tree's splits. - As a result, at least one point from each cluster will end up being assigned to the centroid of a different cluster. - Given the separation distance of $\Omega(\sqrt{d}k)$ between any two cluster centers, the squared distance for each misclassified point will be at least $\Omega(dk^2)$.

Therefore, the k-means cost due to these misclassifications will be at least:

$$\Omega(dk^2)$$

Combining the above steps, we conclude that the explainable k-means cost is $\Omega(dk^2)$, which is significantly higher than the optimal k-means cost of $2dk$. This demonstrates that the ratio between the explainable k-means cost and the optimal k-means cost is $\Omega(k)$.