Sample Problems 6

To be discussed on 05.07.2024

Sample Problem 6.1: Bounds for k-means++

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

$$B(v, r) = \{x \in \mathbb{R}^p \ : \ \|x - v\| \leq r\}, \qquad \text{where } \|x - v\| \text{ is the Euclidean distance.}$$

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 data set $\mathcal{X} = \{x_1, \ldots, x_m\}$ 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 same ball, and the minimum distance between two points in different balls?
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 $\frac{1}{16}$.
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.
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 (you may get an even smaller bound).

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 ith cluster center $c_i \in \mathbb{R}^d$ as

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

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 jth 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

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

   You may now use the following fact: The above result also allows bounding the squared distance between two cluster centers c_s, c_t with high probability. To be precise, one can show that there exists $K_2 > 0$ such that

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

   if you assume $d = K_3 \log k$ for some suitable K_3. *It is not expected of you to prove this fact, but if you want to give it a shot, try using Hoeffding's inequality.*

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 $\Omega(dk^2)$.

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 $\Omega(dk^2)$.