Statistical Foundations of Learning - Assignment 5

CIT4230004 (Summer Semester 2024)

Exercise 5.1: The k-means cost of shifted Rademachers

Given $\mu_1, \ldots, \mu_k \in \mathbb{R}$, consider k independent Rademacher random variables Y_i with means μ_i, in the sense that

$$P(Y_i = \mu_i + 1) = P(Y_i = \mu_i - 1) = \frac{1}{2}$$

1. Show that there exists a sequence $(\mu_i)_{i=1}^{\infty}$ such that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}\left[\min_{j\in[k]} |Y_i - \mu_j|^2\right] = 0$$

Consider the sequence $\mu_i = \frac{i}{k}$. For large k, the μ_i are dense in the interval $[0, 1]$. The Rademacher random variables Y_i take values in $\mu_i \pm 1$.

The expected squared distance to the closest center:

$$\mathbb{E}\left[\min_{j\in[k]} |Y_i - \mu_j|^2\right]$$

For large k, the centers are very close to each other, thus:

$$\min_{j\in[k]} |Y_i - \mu_j| \to 0$$

Thus, we have:

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}\left[\min_{j\in[k]} |Y_i - \mu_j|^2\right] = 0$$

2. What happens if the Y_i's are uniformly distributed on $[\mu_i - 1, \mu_i + 1]$?

If Y_i are uniformly distributed on $[\mu_i - 1, \mu_i + 1]$, the expected squared distance to the closest center can be calculated as follows:

$$\mathbb{E}\left[\min_{j\in[k]} |Y_i - \mu_j|^2\right]$$

For large k, similar reasoning holds:

$$\min_{j \in [k]} |Y_i - \mu_j| \to 0$$

Thus, we have:

$$\lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}\left[\min_{j \in [k]} |Y_i - \mu_j|^2\right] = 0$$

Exercise 5.2: Approximation for k-centre clustering

Consider k-center clustering, defined as follows: Given X and a metric d, find $T = \{t_1, \ldots, t_k\} \subseteq X$ such that

$$G(T) = \max_{x \in X} \min_{t \in T} d(x, t)$$

is minimized.

Algorithm: Farthest point clustering 1. Pick $x \in X$ arbitrarily, and initialize $t_1 = x$. 2. For $i = 2, \ldots, k$: Find $x \in X$ that is farthest from $t_1, \ldots, t_{i-1}$ and set $t_i = x$.

Denote $T_i = \{t_1, \ldots, t_i\}$ as the set of first i centers, and G_i as an intermediate cost after choosing i centers.

1. Show that $G_i \leq G_{i-1}$ for every i.

By the algorithm, each new center t_i is chosen to be the point farthest from the current set of centers T_{i-1}. Therefore, adding t_i cannot increase the maximum distance:

$$G_i \leq G_{i-1}$$

2. Give an example of a case where G_i does not reduce over the iterations.

Consider a set X where all points are equidistant from each other. For example, if X is a set of vertices of a regular polygon with the same distance between any two vertices, then G_i does not change as the new center is equally far from all previous centers.

3. Show that the centers in T_i are at least a distance of G_{i-1} from each other.

After selecting k centers, let $t_{k+1} \in X \setminus T_k$ be the farthest point from T_k. Define $T_{k+1} = T_k \cup \{t_{k+1}\}$. For every $i = 2, \ldots, k+1$, the centers in T_i are at least a distance of G_{i-1} from each other because t_i is chosen to be the farthest point from T_{i-1}.

4. Show that there exist $t, t' \in T_{k+1}$ and $s \in S$ such that s is the closest center for both t and t'.

Consider T_{k+1}. Since T_{k+1} is constructed by choosing the farthest points, the distances between points in T_{k+1} are maximized. Therefore, for any set S

of k centers, there must be at least one center $s \in S$ that is the closest center to at least two points $t, t' \in T_{k+1}$.

5. Show that $G(T) \leq 2G(S)$. Conclude that the algorithm returns a 2-factor approximation.

By the triangle inequality, for any $x \in X$ and any centers $t, t' \in T$ and $s \in S$ such that s is the closest center for both t and t':

$$d(x,t) \leq d(x,s) + d(s,t)$$

$$d(x,t') \leq d(x,s) + d(s,t')$$

Since $t, t' \in T$ are at least a distance of G_{k+1} apart:

$$d(t,t') \geq G_{k+1}$$

Thus:

$$G(T) \leq 2G(S)$$

Therefore, the algorithm returns a 2-factor approximation.