Assignment 5

Due: 11.07.2024, 23:59

Points: 14

The solutions have to be handed in via Moodle. We do not accept late submissions.

We would recommend using LaTeX for writing your submission but also accept handwritten solutions, but please note that if we can not read or understand it, we cannot grade it.

To get full points, always provide the steps in your derivation/proofs and make clear when/how you use known results, for example, from the lecture (e.g. already proven concentration inequalities).

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

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

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

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

that is, the average distance to the closest center approaches zero.

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

(3+3=6 points)

Exercise 5.2: Approximation for k-centre clustering

In this exercise, we consider k-center clustering, which is defined in the following way. Given $\mathcal{X}$ and a metric d, find $T = \{t_1, \ldots, t_k\} \subset \mathcal{X}$ such that

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

is minimised. In other words, we associated every x to its closest center t and wish to minimise the maximum distance between any x and its associated center.

Consider the following algorithm, known as **farthest point clustering**.

- Pick $x \in \mathcal{X}$ arbitrarily, and initialise $t_1 = x$.
- For $i = 2, \ldots, k$:
  Find $x \in \mathcal{X}$ 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. Answer the following to derive an approximation guarantee for the above algorithm.

1. Show that $G_i \leq G_{i-1}$ for every i.
2. Give an example of a case where G_i does not reduce over the iterations.
   Hint: The metric d need not be the Euclidean distance.
3. After selecting k centres, let $t_{k+1} \in \mathcal{X}\backslash T_k$ be the farthest point from T_k. Define $T_{k+1} = T_k \cup \{t_{k+1}\}$. Show that, for every $i = 2, \ldots, k+1$, the centers in T_i are least as a distance of G_{i-1} from each other.
4. Let S be any set of k centres. Show that there exist $t, t' \in T_{k+1}$ and $s \in S$ such that s is the closest centre for both t, t'.

5. Use these results to show that $G(T) \leq 2G(S)$. Conclude that the algorithm returns a 2-factor approximation.

(1+1+2+2+2=8 points)