
Math 173
Problem Set 5
Monday, October 29, 2018

1 Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be independent Rademacher random variables (that is, $\boldsymbol{x}_i = \pm 1$ with equal probability) and $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ be a sequence of real numbers.

(a) Show that $\mathbb{E}e^{s\boldsymbol{x}_i} = \cosh(s)$.

(b) Prove that $\cosh(s) \leq e^{s^2/2}$.

(c) Use (a), (b), and Markov's inequality to prove that

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} a_i \boldsymbol{x}_i\right| \geq \epsilon\right) \leq e^{-\frac{\epsilon^2}{2\|\boldsymbol{a}\|_2^2}}$$

■

2[a] In the k-means clustering problem we are given some input vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m \in \mathbb{R}^n$ and a positive integer k, and we'd like to output a partition P of $\{1, 2, \ldots, n\}$ into k disjoint subsets $P_1, P_2, \ldots, P_k$ and cluster centers $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k \in \mathbb{R}^n$. which minimize the function

$$f(\{P_i\}, \{\boldsymbol{y}_i\}; \{\boldsymbol{x}_i\}) = \sum_{j=1}^{k} \sum_{i \in P_j} \left\| \boldsymbol{x}_i - \boldsymbol{y}_j \right\|_2^2.$$

That is, the x_i are clustered into k clusters according to P. This problem is NP-hard, but good approximation algorithms exist which can return almost-optimal clusterings.

(a) For a fixed partition P, show that the optimal $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k$ is where for every nonempty $P_i \in P$,

$$\boldsymbol{y}_i = \frac{1}{|P_i|} \sum_{j \in P_i} \boldsymbol{x}_i$$

is just the average of the points in P_i. Thus we can restrict ourselves to optimizing over partitions P.

(b) Prove that for a given cluster P_i, the optimal cost is

$$\frac{1}{2|P_i|} \sum_{j,k \in P_i}^{n} \|\boldsymbol{x}_j - \boldsymbol{x}_k\|_2^2$$

(c) Using the Johnson-Lindenstrauss lemma, show that for any $0 < \epsilon < \frac{1}{2}$ there is a linear map $\boldsymbol{S} \in \mathbb{R}^{m \times n}$, $m = \mathcal{O}(\epsilon^{-2} \log m)$ such that for all partitions P simultaneously

$$(1 - \epsilon) f(\{P_i\}; \{\boldsymbol{x}_i\}) \leq f(\{P_i\}; \{\boldsymbol{S}\boldsymbol{x}_i\}) \leq (1 + \epsilon) f(\{P_i\}; \{\boldsymbol{x}_i\})$$

and where $\boldsymbol{S}$ can be found efficiently with a randomized algorithm with small failure probability. Thus if one does not mind worsening the quality of our clusters by a factor $1 + \epsilon$, without loss of generality one can assume that the input vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m \in \mathbb{R}^n$ are in dimension $n = \mathcal{O}(\epsilon^{-2} \log m)$, which can be *much* smaller than the original dimension n. *Hint:* Use the Johnson-Lindenstrauss lemma.

[a] Jelani Nelson, 2013

■