PRINCETON UNIV. F'18 COS 521: ADVANCED ALGORITHM DESIGN

Homework 3: Dimensionality Reduction and High-Dimensional Geometry (**60pts**)

Due: **Monday, November 19th, 2018, 11:59pm**

Collaboration is allowed on this problem set, but solutions must be written-up individually. Please list collaborators for each problem separately, or write "No Collaborators" if you worked alone. Collaboration is not allowed on bonus problems.

Please prepare your problem sets in LaTeX and compile to a PDF for your final submission. A LaTeX template is available on the course webpage.

§1 **(10 pts)** The ℓ_1 distance between vectors $x, y \in \mathbb{R}^d$ is defined as $\|x-y\|_1 = \sum_{i=1}^{d} |x_i - y_i|$. Consider the Johnson-Lindenstrauss dimensionality reduction method described in lecture: $x \to \Pi x$ where each entry in $\Pi \in \mathbb{R}^{m \times d}$ equals

$$\Pi_{ij} = c \cdot g_{ij},$$

for some fixed scaling factor c and $g_{ij} \sim \mathcal{N}(0,1)$. Describe an example (i.e., a set of points in $\mathbb{R}^d$) which shows that, for any choice of c, this method does *not* preserve ℓ_1 distances, even within a factor of 2.

Extra credit: Show that no *linear transformation* suffices, let alone JL.

§2 **(10 pts)** Recall that a hyperplane in $\mathbb{R}^d$ is defined by parameters $a \in \mathbb{R}^d, c \in \mathbb{R}$ and contains all points x such that $\langle a, x \rangle = c$.

Suppose we have n unit vectors in $\mathbb{R}^d$ separated into two sets X, Y with the guarantee that there exists a hyperplane such that every point in X is on one side and every point in Y is on the other. Furthermore, supppose that the ℓ_2 distance of each point in X and Y to this hyperplane is at least ϵ. When this is the case, the seperating hyperplane is said to have "margin" ϵ.

Show that if we use a Johnson-Lindenstrauss map to reduce the points to $O(\log n/\epsilon^2)$ dimensions, then the dimension reduced data can still be separated by a hyperplane with margin $\epsilon/4$, with high probability.

§3 **(5 pts)** A k-sparse vector is any vector with k nonzero entries. Let $\mathcal{S}_k$ be the set of all k-sparse vectors in $\mathbb{R}^d$. Show that, if Π is chosen to be a Johnson-Lindenstrauss embedding matrix (e.g. a scaled random Gaussian matrix) with $s = O(\frac{k \log d}{\epsilon^2})$ rows then, with high probability,

$$(1-\epsilon)\|\Pi x\|_2 \leq \|x\|_2 \leq (1+\epsilon)\|\Pi x\|_2$$

for all $x \in \mathcal{S}_k$, simultaneously.

§4 **(10 pts)** Given a data matrix $X \in \mathbb{R}^{n \times d}$ with n rows (data points) $x_1, \ldots, x_n \in \mathbb{R}^d$, the *k-means clustering problem* asks us to find a partition of our points into k disjoint sets (clusters) $\mathcal{C}_1, \ldots \mathcal{C}_k \subseteq \{1, \ldots, n\}$ with $\bigcup_{j=1}^{k} \mathcal{C}_j = \{1, \ldots, n\}$.

Let $c_j = \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x_i$ be the centroid of cluster j. We want to choose our clusters to minimize the sum of squared distances from every point to its cluster centroid. I.e. we want to choose $\mathcal{C}_1, \ldots \mathcal{C}_k$ to minimize:

$$f_X(\mathcal{C}_1, \ldots \mathcal{C}_k) = \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \|c_j - x_i\|_2^2.$$

There are a number of algorithms for solving the k-means clustering problem. They typically run more slowly for higher dimensional data points, i.e. when d is larger. In this problem we consider what sort of approximation we can achieve if we instead solve the problem using dimensionality reduced vectors in place of $x_1, \ldots, x_n$.

To receive full credit, solve one of (a) or (b). If you solve both, you will get extra credit.

Let $OPT_X = \min_{\mathcal{C}_1, \ldots \mathcal{C}_k} f_X(\mathcal{C}_1, \ldots \mathcal{C}_k)$.

(a) Suppose that Π is a Johnson-Lindenstrauss map into $s = O(\log n/\epsilon^2)$ dimensions and that we select the optimal set of clusters for $\Pi x_1, \ldots, \Pi x_n$. Call these clusters them $\tilde{\mathcal{C}}_1, \ldots \tilde{\mathcal{C}}_k$. Show that they obtain objective value $f_X(\tilde{\mathcal{C}}_1, \ldots \tilde{\mathcal{C}}_k) \leq (1 + \epsilon) OPT_X$, with high probability.

(Hint: reformulate the objective function to only involve ℓ_2 distances between data points.)

(b) Instead, suppose we reduce our points to k dimensions using the SVD. I.e. let $V_k \in \mathbb{R}^{d \times x}$ have the first k right singular vectors of X. Show that, if $\tilde{\mathcal{C}}_1, \ldots \tilde{\mathcal{C}}_k$ are the optimal clusters for $V_k^T x_1, \ldots, V_k^T x_n$, then $f_X(\tilde{\mathcal{C}}_1, \ldots \tilde{\mathcal{C}}_k) \leq 2OPT_X$.

(Hint: show that for every set of clusters, there is an orthonormal matrix $C \in \mathbb{R}^{n \times k}$ such that $f_X(\mathcal{C}_1, \ldots \mathcal{C}_k) = \|X - CC^T X\|_F^2$. I.e. reformulate k-means as a k-rank approximation problem.)

(c) (Extra extra credit.) Show that the optimal clustering for $V_{O(k/\epsilon)}^T x_1, \ldots, V_{O(k/\epsilon)}^T x_n$ gives a $(1 + \epsilon)$ approximation to OPT_X.

§5 **(10 pts)** A *matroid* on $[n]$ elements is a collection of sets that generalized the concept of linear independence for vectors. Specifically, a matroid $\mathcal{I}$ satisfies:

- **Non-trivial:** $\emptyset \in \mathcal{I}$.
- **Downwards-closed:** If $S \in \mathcal{I}$, then $T \in \mathcal{I}$ for all $T \subseteq S$.
- **Augmentation:** If $S, T \in \mathcal{I}$, and $|S| > |T|$, then there exists an $i \in S \setminus T$ such that $T \cup \{i\} \in \mathcal{I}$.[1]

[1] Think of this as a generalization of linear independence: if I give you a set S of k linearly independent vectors, and T of $< k$ linearly independent vectors, then there is some vector in S not spanned by T.

Prove that the following collections are matroids:

(a) Sets of size at most k (that is, the elements are $[n]$, and $\mathcal{I} = \{X \subseteq [n] | \, |X| \leq k\}$).

(b) Acyclic subgraphs of any undirected graph $G = (V, E)$ (that is, the elements are E and $\mathcal{I} = \{X \subseteq E | X \text{ contains no cycles}\}$).

(c) Let $G = (L, R, E)$ be a bipartite graph. The elements are L, and $\mathcal{I} = \{X \subseteq L | \, |N(S)| \geq |S| \; \forall S \subseteq X\}$ ($N(S)$ are the neighbors of S: $\{x \in R | \, \exists y \in S, (x, y) \in E\}$). That is, $X \in \mathcal{I}$ if and only if all nodes in X can be simultaneously matched to R.

§6 **(5 pts)** Given weights $w_i \geq 0, i \in [n]$, and some collection of feasible sets $\mathcal{I}$, your goal is to find the max-weight feasible set: $\arg\max_{S \in \mathcal{I}} \{\sum_{i \in S} w_i\}$. Consider a greedy algorithm that first sorts the elements in decreasing order of w_i (i.e. picks a permutation σ such that $w_{\sigma(i)} \geq w_{\sigma(i+1)}$ for all i), then iteratively does the following (initializing $A = \emptyset$, $i = 1$, go until $i > n$): Check if $A \cup \{\sigma(i)\} \in \mathcal{I}$. If so, add $\sigma(i)$ to A. Update $i := i + 1$. Prove that when $\mathcal{I}$ is a matroid, the greedy algorithm finds the max-weight feasible set.

§7 **(10 pts)** This problem asks you to prove a simplified (and slightly weaker) version of Theorem 7 from the Lecture 14 notes. Bounds on random matrices like this one are valuable in analyzing many other randomized algorithms.

Construct a random *symmetric* matrix $R \in \mathbb{R}^{n \times n}$ by setting $R_{ij} = R_{ji}$ to $+1$ or -1, uniformly at random. Prove that, with high probability,

$$\|R\|_2 \leq c\sqrt{n \log n},$$

for some constant c. This is much better than the naive bound of $\|R\|_2 \leq \|R\|_F = n$.

Hint: For a symmetric matrix R, there is another way to write the spectral norm besides $\|R\|_2 = \max_x \frac{\|Rx\|_2}{\|x\|_2}$. It also holds that $\|R\|_2 = \max_x \frac{|x^T R x|}{x^T x}$.

Hint: Try to bound $\frac{x^T R x}{x^T x}$ for one particular x, and then extend the result to hold for all x, simultaneously. It is possible to solve this problem using the standard Hoeffding bound for bounded random variables – you do not need exotic concentration bounds!