Homework 4
Dimensionality Reduction, Clustering

CMU 10-716: Advanced Machine Learning (Spring 2020)

OUT: April 9, 2020
DUE: **April 23, 2020, 3:00 PM**.

Instructions:

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., “Bob explained to me what is asked in Question 4.3”). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.

- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Each derivation/proof should be completed on a separate page. Submissions can be handwritten, but should be labeled and clearly legible. Else, submission can be written in LaTeX. Upon submission, label each question using the tempate provided by Gradescope.

1 PCA

In this problem, we will connect two perspectives on PCA: one as low-dimensional projections that yields the least error (that we covered in class), and another on low-dimensional projections that have the most variance. Let $X \in \mathbf{R}^{d\times n}$ be an uncentered data matrix and let $\bar{x} = \dfrac{1}{n}\sum_i x_i$ be the sample mean of the columns of X. Note that in this case, we have n data points, each of which is d-dimensional. Recall that the k-dimensional PCA of X is the solution P^* of the problem:

$$\min_{P\in\mathcal{P}_k} \|PX' - X'\|_F^2$$

where $\mathcal{P}_k \in \mathbb{R}^{d\times d}$ is the set of all $d \times d$ projection matrices of rank k, X' is the centered data matrix with columns $x_i' = x_i - \bar{x}$, and $\|A\|_F = \|\mathrm{vec}(A)\|_2$ is the element-wise ℓ_2 norm of a matrix (also called the Frobenius norm).

1. Show that the sample variance of one-dimensional projections of the n data points onto an arbitrary vector u equals $u^T Cu$, where $C = \dfrac{1}{n}\sum_i (x_i - \bar{x})(x_i - \bar{x})^T$ is the sample covariance matrix.

2. Show that

$$\underset{u:\|u\|_2=1}{\mathrm{argmin}} \|uu^T X' - X'\|_F^2 = \underset{u:\|u\|_2=1}{\mathrm{argmax}}\, u^T Cu,$$

which says that PCA with $k = 1$ projects the data onto the direction of maximal variance.

2 Johnson Lindenstrauss

In this question, we will show that a large positive definite matrix can be decently approximated by a matrix of much smaller rank with high probability. To this end, define $\epsilon-$rank to be:

$$r_\epsilon(X) = \min\{\mathrm{rank}(A) : A \in \mathbb{R}^{n\times n}, \|X - A\|_{\max} \leq \epsilon\}$$

where $\|B\|_{\max} = \max_{i,j} B_{ij}$. We will show that r_ϵ is small w.h.p.

1. Recall that JL states that if $x_1, ..., x_n \in \mathbb{R}^N$ and take $r = \lceil \frac{8\ln(n)}{\epsilon^2} \rceil$, then there exists a linear map $Q : \mathbb{R}^N \to \mathbb{R}^r$ such that w.h.p:

$$(1 - \epsilon)\|x_i - x_j\|^2 \leq \|Q(x_i - x_j)\|^2 \leq (1 + \epsilon)\|x_i - x_j\|^2$$

preserves pairwise distances.

Using this, we can derive a variant of the JL that will be useful later on.

Let $x_1, ..., x_n \in \mathbb{R}^N$ and consider $r = \lceil \frac{8\ln(n+1)}{\epsilon^2} \rceil$. Prove there exists a linear map $Q : \mathbb{R}^N \to \mathbb{R}^r$ such that w.h.p:

$$|x_i^T x_j - x_i^T Q^T Q x_j| \leq \epsilon(\|x_i\|_2^2 + \|x_j\|_2^2 - x_j^T x_i)$$

$\forall i, j \in [1, n]$ and preserves pairwise inner products.

Hint: Apply JL to points $\{x_1, ..., x_n, 0\}$

2. Using the previous part, show that if $X \in \mathbb{R}^{n\times n}$ positive definite with $\epsilon > 0$, $r = \lceil 72\ln(2n+1)/\epsilon^2 \rceil$, there exist a matrix Y with rank $\leq r$ such that w.h.p:

$$\|X - Y\|_{\max} \leq \epsilon \|X\|_2$$

where $\|A\|_2$ is the spectral norm of the matrix A, equal to the largest singular value of the matrix.

Hint: Consider computing the SVD of $X = U'\Sigma V'^T = UV$ where $U = U'\sqrt{\Sigma}$ and $V = V'\sqrt{\Sigma}$, and apply part 1 to columns of U and V.

Hint 2: Use that $\|X\|_2 \geq \|X\|_{\max}$

3 k-Means^{++}

Recall the k-Means^{++} initialization procedure where, given data $X = \{X_1, \ldots, X_n\}$, we pick the centers making up our codebook C according to the following steps

a) Choose c_1 uniformly at random from X. Let $C = \{c_1\}$.

b) Compute $D(X_i) = \min_{c\in C} \|X_i - c\|$ for each $X_i \in X$.

c) Set $c = X_i$ with probability $\frac{D(X_i)^2}{\sum_{j=1}^n D(X_j)^2}$ and set $C = C \cup \{c\}$.

d) Go back to step b) if $|C| < k$.

Furthermore recall that the empirical risk of a codebook C is

$$R_n(C) = \frac{1}{n}\sum_{i=1}^n \min_{1\leq j\leq k} \|X_i - c_j\|^2$$

and we define $\hat{C}$ to be the codebook that globally minimizes this quantity. In what follows, where $S \subset X$, let $R_{n,S}(C)$ be the portion of the empirical risk that comes from points in S. That is,

$$R_{n,S}(C) = \frac{1}{n}\sum_{x\in S} \min_{1\leq j\leq k} \|x - c_j\|^2$$

In this problem we will show the weaker result that the k-means^{++} initialization technique is competitive when we use it to chooses centers from each cluster of the global optimum $\hat{C}$.

1. For a set of points S with center of mass $c(S) = \frac{1}{|S|}\sum_{x\in S} x$ and an arbitrary point z, show that

$$\sum_{x\in S} \|x - z\|^2 - \sum_{x\in S} \|x - c(S)\|^2 = |S| \, \|c(S) - z\|^2$$

2. Let A be an arbitrary cluster in $\hat{C}$, and let C be a codebook with just one center, which is chosen uniformly at random from A. Show that $\mathbb{E}[R_{n,A}(C)] = 2R_{n,A}(\hat{C})$.

3. Let A be an arbitrary cluster in $\hat{C}$, but now, let C be any arbitrary codebook. Show that if we add a random center to C from A, chosen with the D^2 weighting outlined above, then $\mathbb{E}[R_{n,A}(C)] \leq 8R_{n,A}(\hat{C})$.

   Hint: You may want to show that for arbitrary $a_0, a \in A$ that $D(a_0)^2 \leq \frac{2}{|A|}\sum_{a\in A} D(a)^2 + \frac{2}{|A|}\sum_{a\in A} \|a - a_0\|^2$. A useful inequality to use for this is the so-called power-mean inequality which states that for any real numbers $x_1, \ldots, x_m$ that $\sum x_i^2 \geq \frac{1}{m}\left(\sum x_i\right)^2$.