Unsupervised Learning HW1

Due: Sat Oct 08, 2022 at 11:59pm

All homeworks (including this one) should be typesetted properly in pdf format. Late homeworks or handwritten solutions will not be accepted. You must include your name and UNI in your homework submission. To receive credit, a typesetted copy of the homework pdf must be uploaded to Gradescope by the due date. You must show your work to receive full credit. Discussing possible solutions for homework questions is encouraged on the course discussion board and with your peers, but everyone must write their own individual solutions. You must cite all external references you used (including the names of individuals you discussed the solutions with) to complete the homework.

1 Readings

Let x be the last digit of your UNI, and define $y := (x \mod 5) + 1$. Read the y^{th} paper from the list below (yes, we'll check). Summarize the main results of your assigned paper, discuss their significance and provide a short proof sketch of their technical results.

(i) "Clustering with Interactive Feedback" by Balcan and Blum

(ii) "Incremental clustering: the case for extra clusters" by Ackerman and Dasgupta

(iii) "Comparing Clusterings – An Axiomatic View" by Meila

(iv) "Hartigan's Method: k-means Clustering without Voronoi" by Telgarsky and Vattani

(v) "Clustering large graphs via the singular value decomposition" by Drineas, Kannan, Frieze, Vempala, and Vinay

2 Hardness of k-center

Show that the k-center problem is NP-hard.

(Hint: You can use the fact that the following variation of the Vertex-Cover problem is NP-hard.

Vertex-Cover*

- **Input:** An undirected unweighted graph $G = (V, E)$, with vertices V and edges E.
- **Output:** $V' \subseteq V$, such that $V' \cup \left(\cup_{v' \in V'} \cup_{e(v',v) \in E} \; v \right) = V$.
- **Goal:** minimize $|V'|$.

)

3 Coverings and Packings

Given a metric space (X, d), and an $\epsilon > 0$, a set $C \subset X$ is called an ϵ-*cover* if $\forall x \in X, \exists c \in C$ such that, $d(x, c) \leq \epsilon$, a set $P \subset X$ is called an ϵ-*packing* if $\forall p, p' \in P$ distinct, $d(p, p') > \epsilon$.

(i) Show that for any compact metric space (X, ρ) and any $\epsilon > 0$, there exists $Y \subset X$, such that Y is both an ϵ-cover and an ϵ-packing.

ϵ-covering number of X, denoted by $N_\epsilon(X)$, is defined to be the size of *smallest* ϵ-cover of X, similarly ϵ-packing number of X, denoted by $P_\epsilon(X)$, is defined to be the size of *largest* ϵ-packing of X.

(ii) Show that for any metric space (X, ρ) and any $\epsilon > 0$,

$$P_\epsilon(X) \leq N_{\epsilon/2}(X) \leq P_{\epsilon/2}(X).$$

(iii) Let $B^d(x, r)$ denote the closed Euclidean ball of radius r centered at x in $\mathbb{R}^d$, that is $B^d(x, r) := \{p \in \mathbb{R}^d \mid \|x - p\|_2 \leq r\}$. Give an estimate of the ϵ-covering number of $B^d(0, 1)$ (ie, closed Euclidean unit ball in $\mathbb{R}^d$, centered at the origin).

(Hint: consider a maximal size packing of $B^d(0, 1)$ and compare the relative d-dimensional volumes of balls centered at elements of the packing with a ball containing all of them, using the fact that vol$(B^d(x, r)) = r^d$vol$(B^d(x, 1)$)

(iv) **[Application: approximating maximum singular value using a cover]** Given a $m \times n$ matrix A, recall that the maximum singular value of A is defined as

$$\sigma_{\max}(A) := \max_{x \in \mathbb{R}^n, \|x\|=1} \|Ax\|$$

An (impractical) solution to finding $\sigma_{\max}(A)$ is to test the condition $\|Ax\|$ on each $x \in S^{n-1} := \{x \in \mathbb{R}^n : \|x\| = 1\}$ and return the maximum value. The obvious problem with this approach is of course that the set S^{n-1} is of infinite size and thus one cannot return $\sigma_{\max}$ in finite time using such an approach.

One obvious remedy is to approximate $\sigma_{\max}$ by testing the value $\|Ax\|$ on a finite (albeit large) collection of x and approximates S^{n-1} well. More concretely, one can construct a *cover* C of S^{n-1}, and find the maximum $\|Ax\|$ on the each $x \in C$ (let's denote that as σ_C) and relate it to $\sigma_{\max}$.

Give a tight upperbound on the approximation $\sigma_{\max}$ as a function of σ_C, and the quality of the cover C.

4 Low-Dimensional Embeddings from Dissimilarity Data

We often encounter problems in machine learning where we do not have access to the data directly, but instead to dissimilarity ratings or comparisons between our datapoints. For example, in a series of medical trials for drug development, we do not have access to a Euclidean representation of the drugs in question, but we may have access to differential trial data which compared the performance of different drugs across a population. For another more concrete example, consider distances between cities. Given interpoint distances between n cities, we'd like to be able to reproduce the 2-dimensional global positions of the cities in question. As a third example, at a wine tasting, you may know only the relative quality or character of each wine, represented as a set of ratings, but you may want to find an embedding of the wines for clustering or visualization purposes, according to these ratings. We will explore how this can be done.

Mathematically, we are given dissimilarity ratings in an $n \times n$ matrix $D \in \mathbb{R}^{n \times n}$ where $D_{ij} = d(\alpha_i, \alpha_j)^2$ for some data $\alpha_1, ...\alpha_n$ (which we do not have access to). We'd like to find a k-dimensional Euclidean representation $x_1, ..., x_n \in \mathbb{R}^k$ such that

$$\sum_{i \neq j}^{n} \left(D_{ij} - \|x_i - x_j\|^2\right)^2$$

is minimized, i.e. the learned (squared) Euclidean distance is as close as possible to the given distance D_{ij}.

(i) First, we will show that, if the underlying data $\alpha_1, ..., \alpha_n$ is Euclidean in $\mathbb{R}^n$ (i.e. there exists a perfect embedding such that $\|\alpha_i - \alpha_j\|^2 = D_{ij}$ for all i, j), we can recover this embedding exactly from the D matrix alone. First, we would like to transform the data matrix D_{ij} into a set of inner products of the form

$$= \langle \alpha_i - \overline{\alpha}, \alpha_j - \overline{\alpha} \rangle$$

where $\overline{x}$ represents the data average. Let $H = I - \frac{1}{n}\mathbb{1}\mathbb{1}^T$. Show that $-\frac{1}{2}H^T D H$ has the desired form. This is called a Gram Matrix.

Hint: $\|\alpha_i - \alpha_j\|^2 = \langle \alpha_i, \alpha_i \rangle + \langle \alpha_j, \alpha_j \rangle - 2\langle \alpha_i, \alpha_j \rangle$. Also try expanding both sides and matching up terms.

(ii) Assume the matrix B_{ij} is in this form. Let Q be the matrix whose columns are the eigenvectors of B, and $\Lambda^{1/2}$ the diagonal matrix whose diagonal entries are roots of the corresponding eigenvalues. Show that the rows of the matrix $Q\Lambda^{1/2} \in \mathbb{R}^{n \times n}$ are a perfect (isometric) embedding of the original data into $\mathbb{R}^n$.

Hint: First prove that B is positive semi-definite. What does this imply? It turns out that the data matrix is in fact isometrically embeddable in $\mathbb{R}^n$ `if and only` if the Gram matrix is positive semi-definite.

(iii) What if we want a lower-dimensional embedding instead? Show that if we can take the top k eigenvectors Q_k and corresponding eigenvalues Λ_k of the centered matrix B, the rows of $Q_k \Lambda_k^{1/2}$ minimize the loss

$$\sum_{i \neq j}^{n} \left(D_{ij} - \|x_i - x_j\|^2\right)^2$$

over all possible k-dimensional embeddings $x_i \in \mathbb{R}^k$ (this is the same as PCA on the original matrix X).

Hint: You may wish to use SVD and the Eckart-Young theorem, or (requiring slightly more work) rewrite this equation as the PCA objective function. Rewriting the objective using the Frobenius norm as

$$\min_{\text{rank}\mathbf{Q}<k} \|D - \mathbf{Q}\|_F^2$$

think about whether applying H^T and H to the left and right sides of the normed term actually changes the minimizing value of Q? How does this change D and Q?

5 Effects of a Gaussian Random Matrix

Given a unit vector $v \in \mathbb{R}^D$, show the following.

Pick any $\epsilon > 0$ and $d \in \mathbb{Z}_+$. Let G be a $d \times D$ matrix with entries i.i.d. $N(0, 1/d)$. Then, (for a universal constant c)

$$\Pr_{\boldsymbol{G}}\left[\|Gv\|^2 > (1+\epsilon)^2\right] < e^{-cd\epsilon^2}.$$

(Hint: Use the result from HW0 for the length deviation)