Unsupervised Learning HW2

Due: Sun Oct 28, 2018 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 piazza 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]** Read any two papers of your choice from the following list, summarize their main results, discuss their significance and provide a short proof sketch of their technical results.

- "A tight bound on approximating arbitrary metrics by tree metrics" by Fakcharoenphol, Rao and Talwar.
- "On the Impossibility of Dimension Reduction in ℓ_1" by Brinkman and Charikar.
- "Bounded geometries, fractals, and low-distortion embeddings" by Gupta, Krauthgamer and Lee.
- "Some theory for ordinal embedding" by Arias-Castro.
- "A lower bound on the distortion of embedding planar metrics into Euclidean space" by Newman and Rabinovich.

2 **[From Distances to Embeddings]** Given n items $\gamma_1, \ldots, \gamma_n \in \Gamma$ (not necessarily in Euclidean representation), and a (symmetric) comparison function $\rho : \Gamma \times \Gamma \to \mathbb{R}$, we will show that the n items are (isometrically) embeddable in n-dimensional Euclidean space with representation $x_1, \ldots, x_n$, such that $\|x_i - x_j\|^2 = \rho(\gamma_i, \gamma_j)$ if and only if the $n \times n$ matrix $G := -\frac{1}{2} H^{\mathsf{T}} D H$ is positive semidefinite, where both D and H are $n \times n$ matrices defined as $D_{ij} := \rho(\gamma_i, \gamma_j)$ and $H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}$.

(a) First, show that

$$G_{ij} = -\frac{1}{2}\Big(\rho_{ij} - \frac{1}{n}\sum_j \rho_{ij} - \frac{1}{n}\sum_i \rho_{ij} + \frac{1}{n^2}\sum_{ij} \rho_{ij}\Big),$$

where ρ_{ij} is the shorthand for $\rho(\gamma_i, \gamma_j)$.

(b) If there exists $x_1, \ldots, x_n \in \mathbb{R}^n$, such that $\|x_i - x_j\|^2 = \rho_{ij}$, then show that

$$G_{ij} = (x_i - \bar{x})^{\mathsf{T}}(x_j - \bar{x}),$$

where $\bar{x} = \frac{1}{n}\sum_i x_i$.

(c) Part (b) shows that for a Euclidean representation, G denotes is the (mean centered) inner product matrix of the Euclidean representation of the n items. Show that for G positive semidefinite, one can factorize $G := X^{\mathsf{T}}X$, where $X = [x_1, \ldots, x_n]$ and thus recovering the Euclidean representation of the given n items.

3 **[ℓ_1 to ℓ_∞ embedding for arbitrary size spaces]** Find an isometric embedding from ℓ_1^d to $\ell_\infty^{2^d}$.

4 **[ℓ_∞ to ℓ_2 embedding]** Recall from lecture that:

Theorem: Let q be any integer ≥ 2, and define $D := 2q - 1$. Let (X, ρ) be an n-point metric space. Then there exists a D-embedding of X into ℓ_∞^d with $d = O(qn^{1/q} \ln n)$.

Using the above result, show that every n-point metric space can be D-embedded into ℓ_2^d with $D = O(\log^2 n)$ and $d = O(\log^2 n)$.

5 **[Isometric embeddings of trees into ℓ_1]** Show that any finite tree (with arbitrary branching factor) can be isometrically embedded into ℓ_1.

(*hint:* use induction!)

6 **[Boolean hypercube into ℓ_2]** In this problem, we'll show that the Boolean hypercube Q_d (for $d \geq 2$) cannot be isometrically embedded into ℓ_2^m.

The d-dimensional Boolean hypercube is the graph on $Q_d = \{0,1\}^d$ with the ℓ_1-distance ρ; an edge joins $u, v \in Q_d$ iff $\rho(u, v) = 1$. We can interpret Q_d as the collection of all possible outcomes of d coin tosses; the distance between two sequences $u, v \in Q_d$ is just the number times they differed, and two sequences are connected by an edge iff they differed only once.

Of course, it suffices to show that (Q_2, ℓ_1) cannot be embedded into ℓ_2^m (and because Q_2 is contained in Q_d for $d \geq 2$, then no higher dimensional hypercube can be embedded isometrically into any ℓ_2^m). However, the route we take is easily modified to show that there are graphs with n vertices whose embeddings into a Euclidean space ℓ_2^m require distortion $\Omega(\log n)$.

Notice that the metric structure of Q_d is independent of its graph structure (ρ and $E(G)$ are independent). And so ρ is also a perfectly valid metric on the complete graph K_{2^d} on 2^d vertices. In the proceeding steps, we'll formalize the intuition that any embedding of Q_d into ℓ_2^m must shrink the edges of K_{2^d} that are not in Q_d.

(a) For any graph with metric, define $\rho^2(G)$ be the sum of squared lengths of its edges:

$$\rho^2(G) := \sum_{e \in E} \rho(e_0, e_1)^2.$$

Compute $\rho^2(Q_d)$ and $\rho^2(K_{2^d})$. *Hint: every edge is connected to exactly two vertices, so summing over edges is equivalent to summing over vertices, up to a factor of 2.*

(b) Given any embedding $f : Q_d \to \ell_2^m$, let σ be the induced metric on Q_d, where

$$\sigma(u, v) = ||f(u) - f(v)||_2.$$

Of course, σ is translation invariant, so we may assume f is zero-centered (i.e. the mean of $f(u)$ for all $u \in Q_d$ is 0). When $m = 1$, then $f : Q_d \to \mathbb{R}$ is just a vertex function. Give the formula for $\sigma^2(Q_d)$ and $\sigma^2(K_{2^d})$ in terms of their Laplacians and f.

(c) When $m = 1$, using the fact that the smallest nontrivial eigenvalue for Q_d is $\lambda_2 = 2$, deduce a lower bound for $\sigma^2(Q_d)$.

(d) When $m = 1$, using the Laplacian formula from part (b), show that $\sigma^2(K_{2^d}) = 2^d \cdot ||f||_2^2$, assuming that f is zero-centered.

(e) If f were an isometric embedding, then the ratio $\rho^2(K_{2^d})/\rho^2(Q_d)$ should be equal to its corresponding ratio after embedding, $\sigma^2(K_{2^d})/\sigma^2(Q_d)$. However, the calculations from parts (a), (c) and (d) show that:

$$\frac{\rho^2(K_{2^d})}{\rho^2(Q_d)} = \frac{\sigma^2(K_{2^d})}{\sigma^2(Q_d)} \cdot \Omega(d).$$

This implies that any embedding $f : Q_d \to \ell_2^1$ will require distortion $D = \Omega(\sqrt{d})$. Explain why this result for ℓ_2^1 implies that any embedding into ℓ_2^m must also have distortion $\Omega(\sqrt{d})$. *Hint: in ℓ_2^m, Pythagorean theorem holds; what is the formula for $\sigma^2(G)$?*