Unsupervised Learning HW4

Due: Fri Dec 14, 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.

- "Horseshoes in Multidimensional Scaling and Local Kernel Methods" by Diaconis, Goel and Holmes.
- "An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists" by Chazal and Michel.
- "Similarity Search in High Dimensions via Hashing" by Gionis, Indyk and Motwani.
- "A concentration theorem for projections" by Dasgupta, Hsu and Verma.
- "Nearest-neighbor searching and metric space dimensions" by Clarkson.
- "A Topological View of Unsupervised Learning and Clustering" by Niyogi, Smale and Weinberger.

2 **[Density Estimation and Curse of Dimensionality]** Here we will try to establish (nonparametric) density estimation rates and reveal the curse of dimensionality.

Consider the problem of estimating a fixed unknown probability density f over $\mathbb{R}^D$, given an i.i.d. sample $x_1, \ldots, x_n$.

(a) (Estimating cdf in 1D) Let $D = 1$. Since the density function $f(x)$ is simply the derivative of cumulative density $F(x) := \Pr_{x'}[x' \leq x]$, one can start with estimating F from n samples first. A natural estimator (based on n i.i.d. samples) would be:

$$\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[x_i \leq x]. \tag{1}$$

(i) Show that $\hat{F}_n(x)$ is a consistent estimator for $F(x)$.

(ii) For a fixed threshold $x \in \mathbb{R}$, what is the rate of convergence $|\hat{F}_n(x) - F(x)|$? (hint: think about Hoeffding's inequality)

(iii) The rate in part (i) is not terribly useful since it only guarantees that our estimator is good only at one fixed threshold. Ideally we want an estimator that is good for all thresholds (uniformly). That is, derive a *uniform* convergence rate

$$\sup_x |\hat{F}_n(x) - F(x)|.$$

(hint: Design an appropriate hypothesis class of classifiers such that empirical error of a classifier is $\hat{F}_n(x)$ and true error is $F(x)$. Examine the VC dimension of the hypothesis class to derive the uniform convergence rate)

(b) (Estimating cdf, general case) In higher dimensions, the cumulative density is defined as: $F(x) := \Pr_{x'}[x' \preccurlyeq x]$, where $x, x' \in \mathbb{R}^D$ and "$\preccurlyeq$" is simply a component-wise inequality. The estimator Eq. (1) can be extended in a natural way to get

$$\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[x_i \preccurlyeq x].$$

Derive a uniform convergence rate: $\sup_x |\hat{F}_n(x) - F(x)|$.
(hint: Do Part (a)(iii) first)

(c) (Estimating pdf in 1D) Let $D = 1$. Unfortunately one cannot directly use the estimator for the cumulative density Eq. (1) to get an estimator for the density (since the derivate of $\hat{F}_n$ would be zero almost everywhere!). Instead, one can construct an estimator $\hat{f}_n(x)$ by taking a difference ratio:

$$\hat{f}_n^h(x) := \frac{\hat{F}_n(x+h) - \hat{F}_n(x)}{h} = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{1}[x < x_i \leq x + h], \tag{2}$$

where h is a small "bandwidth window", typically depending on the sample size (i.e. $h \equiv h_n$).

(i) Show that:

$$\mathbb{E}(f(x) - \hat{f}_n^h(x))^2 = \underbrace{(f(x) - \mathbb{E}\hat{f}_n^h(x))^2}_{\text{squared bias of estimator (*)}} + \underbrace{\mathbb{E}[(\hat{f}_n^h(x) - \mathbb{E}\hat{f}_n^h(x))^2]}_{\text{variance of estimator (**)}}. \tag{3}$$

Moreover, show that: $(*) \to 0$ (when $h_n \to 0$ as $n \to \infty$), and $(**) \to 0$ (when $nh_n \to \infty$ as $n \to \infty$)

(ii) Assuming F is sufficiently smooth, one can do a second-order Taylor series expansion around $h = 0$ to get: $F(x+h) = F(x) + f(x) \cdot h + \frac{f'(x)}{2}h^2 + o(h^2)$. Using this, show that Eq. (3) can be written as:

$$\mathbb{E}(f(x) - \hat{f}_n^h(x))^2 = \Big(\frac{f'(x)h}{2}\Big)^2 + o(h^2) + \frac{f(x)}{nh} + O\Big(\frac{1}{n}\Big)$$

What setting of bandwidth h yields the fastest rate of convergence? What is the rate of convergence for this setting of the bandwidth?

(d) (Estimating pdf, general case) A multivariate generalization of the estimator Eq. (2) can be given as

$$\hat{f}_n^h(x) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^D}\mathbf{1}[x \preccurlyeq x_i \preccurlyeq x+h].$$

(i) Again, assuming $f(x)$ is sufficiently smooth, one can do a second-order Taylor series expansion: $f(x-hu) = f(x) - h\big(\frac{\partial f(x)}{\partial x'} \cdot u\big) + \frac{h^2}{2}\mathrm{tr}\Big(\frac{\partial^2 f(x)}{\partial x \partial x'}uu^{\mathsf{T}}\Big) + o(h^2)$. Using this, and recalling the bais-variance decomposition of $(f(x) - \hat{f}_n^h(x))^2$ (cf. Eq. (3)), show that:

- (sq. bias term) $(f(x) - \mathbb{E}\hat{f}_n^h(x))^2 = O(h^4)$
- (variance term) $(\hat{f}_n^h(x) - \mathbb{E}\hat{f}_n^h(x))^2] = O(1/nh^D)$

(ii) Using the result from Part (i), what setting of bandwidth h yields the fastest rate of convergence? What is the rate of convergence for this setting of the bandwidth?

(Note: It turns out that this rate is tight, and explictly shows the curse of dimensionality of nonparametric density estimation)

3 **[Fun with Tensors]** We defined in class:

Definition. *If V is a finite-dimensional real vector space, the* dual space *of V is the vector space of linear maps $f : V \to \mathbb{R}$:*

$$V^* = \{f : f \text{ is a linear map from } V \text{ to } \mathbb{R}\},$$

where the linear structure is defined pointwise. That is, $\lambda f + \mu g$ is defined so that

$$(\lambda f + \mu g)(v) \equiv \lambda f(v) + \mu g(v)$$

for all $v \in V$ and $\lambda, \mu \in \mathbb{R}$.

(a) Check for yourself that V^* is a vector space, with $\dim(V^*) = \dim(V)$. No need to write anything for this part.

(b) Define a coordinate-free injective linear map $V \to (V^*)^*$. That is, your map cannot depend on a choice of basis on V.

The space V^{**} is called the *double dual* of V. Note that since V is finite-dimensional, your injective map shows that V and V^{**} are isomorphic; in fact, it is quite likely you defined the *canonical isomorphism* between V and V^{**}.

(c) [Optional (for the algebraically-inclined)] In contrast, no coordinate-free isomorphism exists between V and V^*. Explain why any isomorphism $T : V \to V^*$ must depend on an outside choice of basis.

Below, we'll give the rigorous definition of the tensor product, so there is no ambiguity to what we're working with. However, in actual computation, there is absolutely no problem in just thinking of a tensor T in $V \otimes W$ as a linear combination of $v_i \otimes w_i$'s:

$$\sum_{i=1}^{k} \lambda_i v_i \otimes w_i$$

where $\otimes : V \times W \to V \otimes W$ is an operation that bilinearly attaches vectors from V and W together. That is to say that we may freely move scalars through the tensor product and distribute over addition:

$$(\lambda v + \lambda' v') \otimes w \equiv \lambda v \otimes w + \lambda' v' \otimes w,$$

and similarly for $v \otimes (\lambda w + \lambda' w')$. Further note that it often suffices to prove facts about pure tensors $v \otimes w$ then just linearly extend to their linear combinations.

Definition. *Let V and W be vector spaces. The vector space $V \otimes W$ is the vector space generated by the set $\{v \otimes w : v \in V, w \in W\}$, modulo the equivalence relations:*

$$(\lambda v + \lambda' v') \otimes w \sim \lambda(v \otimes w) + \lambda'(v' \otimes w)$$
$$v \otimes (\lambda w + \lambda' w') \sim \lambda(v \otimes w) + \lambda'(v \otimes w').$$

(Freely generate a vector space over $V \times W$ then quotient by the above equivalence relations).

We stated in class that we may interpret $f \otimes v \in V^* \otimes V$ three ways:

(i) a linear transformation $(V^*)^* \to V$, so a linear transformation $V \to V$, where:

$$(f \otimes v)(u) = f(u)v.$$

(ii) a linear transformation $V^* \to V^*$, where:

$$(f \otimes v)(g) = g(v)f.$$

(iii) a bilinear map $V \times V^* \to \mathbb{R}$, where:

$$(f \otimes v)(u, g) = f(u)g(v).$$

This is equivalent to saying that a matrix $M \in \mathbb{R}^{n \times m}$ has three interpretations:

(i) a linear transformation of column vectors, $\mathbb{R}^m \to \mathbb{R}^m$, where $x \mapsto Mx$.

(ii) a linear transformation of row vectors, $\mathbb{R}^n \to \mathbb{R}^n$, where $y \mapsto yM$.

(iii) a bilinear map $\mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, where $(x, y) \mapsto yMx$.

Let us now prove the *fundamental theorem of linear algebra*, which says that the row rank and the column rank of a matrix is equal.

Definition. *The rank of a tensor $T \in V_1 \otimes \cdots \otimes V_n$ is the minimal r such that:*

$$T = \sum_{i=1}^{r} v_i^{(1)} \otimes \cdots \otimes v_i^{(n)},$$

for some collection of $v_i^{(k)} \in V_k$. For example, when $T \in V \otimes W$, the rank is the minimal r such that:

$$T = \sum_{i=1}^{r} v_i \otimes w_i.$$

(d) Let $T = \sum_{i=1}^r v_i \otimes w_i$. Show that if r is minimal, the sets $\{v_i : 1 \leq i \leq r\}$ and $\{w_i : 1 \leq i \leq r\}$ are linearly independent.

(e) Let $T : V \to V$ be defined by $T = \sum_{i=1}^r v_i \otimes f_i$. What are the row space and column space of T? (with proof). Deduce the fundamental theorem of linear algebra.

(f) Give an example of a tensor $T = \sum_{i=1}^r u_i \otimes v_i \otimes w_i$, where the $u_i, v_i, w_i \in \mathbb{R}^2$ and not all of the sets $\{u_i\}$, $\{v_i\}$, and $\{w_i\}$ are linearly independent. Thus, the analogous statement to *row rank equals column rank* is generally false for tensors.

(g) Give an example of $T \in \mathbb{R}^2 \otimes \mathbb{R}^2 \otimes \mathbb{R}^2$ of rank ≥ 3. Thus, the analogous statement to *if $M \in \mathbb{R}^{m \times n}$, then* $\operatorname{rank}(M) \leq \min\{m, n\}$ is generally false for tensors.

4 **[Tensor Power Method]** In this problem, we will further assume that V is an inner product space. With the additional inner product structure, there is a natural identification of V and V^* (contrast to part c of previous problem).

(a) Give a natural isomorphism between V and V^* (by natural isomorphism, we mean one that is invariant to a choice of basis).

Because we have an inner product, we might as well work with an orthonormal basis, so we can identify V with $\mathbb{R}^n$. Likewise, we can identify V^* with $\mathbb{R}^n$. It is for this reason that much of the literature in machine language working with tensors don't distinguish between $\mathbb{R}^n$ and $(\mathbb{R}^n)^*$, the notion of distance/inner product allowing for this. We will do this too here.

Definition. *Let S^dV denote the space of* symmetric tensors *in $V^{\otimes d} \equiv V \otimes \overset{d \text{ times}}{\cdots} \otimes V$. That is, tensors of the form:*

$$T = \sum_{i=1}^r \lambda_i v_i^{\otimes d} \equiv \sum_{i=1}^r \lambda_i v_i \otimes \overset{d \text{ times}}{\cdots} \otimes v_i. \tag{4}$$

As above, we may identify $V^{\otimes d}$ with space of multilinear maps $V^* \overset{d-1 \text{ times}}{\times \cdots \times} V^* \to V$, which according to our remark, we may identify with multilinear maps of the form:

$$\mathbb{R}^n \overset{d-1 \text{ times}}{\times \cdots \times} \mathbb{R}^n \to \mathbb{R}^n.$$

Then, an eigenvector/eigenvalue pair *is a (v, λ) such that v is unit length and:*

$$T(v, \ldots, v) = \lambda v.$$

(b) Let T as in Equation 4. Denote the inner product on V by $\langle \cdot, \cdot \rangle$. Compute $T(v, \ldots, v)$.

Definition. *A tensor $T \in S^dV$ is* orthogonally decomposable (odeco) *if there is an orthonormal basis $\{v_1, \ldots, v_n\}$ of V and $\lambda_i \in \mathbb{R}$ (possibly $\lambda_i = 0$) such that:*

$$T = \sum_{i=1}^n \lambda_i v_i^{\otimes d}.$$

(c) Characterize the subset of $S^2\mathbb{R}^n$ of odeco tensors (which tensors in $S^2\mathbb{R}^n$ are odeco?).

(d) Notice that we require eigenvectors to be of unit length. Let $T \in S^d\mathbb{R}^n$, and suppose we had some $u \in \mathbb{R}^n$ not of unit length such that:

$$T(u, \ldots, u) = \lambda u.$$

Give an eigenvalue/eigenvector pair of T.

This shows the subspace of vectors $v \in \mathbb{R}^n$ such that $T(v, \dots, v) = \lambda v$ for some $\lambda \in \mathbb{R}$ is no longer a linear subspace when T is an order d tensor, $d > 2$. This should make sense, because when T is viewed as a map $T : \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}^n$, it is a $(d-1)$-linear transformation[1], and not a linear transformation.

One way to compute eigenvectors/eigenvalues for matrices $M \in S^2\mathbb{R}^n$ is the *matrix power method*. The algorithm is:

- uniformly at random, choose u from the unit sphere in $\mathbb{R}^n$.
- while some convergence condition not satisfied:
  - update u with $Mu/||Mu||$
- return u and $\lambda = M(u, u)$

This is a computational form of the Rayleigh quotient:

$$v_1 = \operatorname*{arg\,max}_{v \in S^{n-1}} M(v, v) = \operatorname*{arg\,max}_{v \in \mathbb{R}^n} \frac{M(v, v)}{||v||^2}.$$

Let $v_1, \dots, v_n$ be an orthonormal basis of M, and assume for simplicity that $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. We'll first show that u will almost surely converges to v_1 at a rate linear with respect to the *eigengap*, λ_2/λ_1.

(e) Let $u^{(t)}$ denote the tth iteration of applying $v \mapsto Mv/||Mv||$ to u. Show that:

$$||v_1 - u^{(t)}|| = \left(\frac{\lambda_2}{\lambda_1}\right)^t \cdot O(1).$$

(f) Briefly describe how to extend the matrix power method to (aproximately) retrieve all eigenvector/eigenvalue pairs of M.

We'll now show a convergence bound for the analogous tensor power method, for odeco tensors $T \in S^3\mathbb{R}^n$. The algorithm is:

- uniformly at random, choose u from the unit sphere in $\mathbb{R}^n$.
- while some convergence condition not satisfied:
  - update u with $T(u, u)/||T(u, u)||$
- return u and $\lambda = T(u, u, u)$

Here, we let $T = \sum_{i=1}^n \lambda_i v_i \otimes v_i \otimes v_i$, and $u = \sum_{i=1}^n \mu_i v_i$, where v_i's form an orthonormal basis of $\mathbb{R}^n$.

(g) For simplicity, assume $|\lambda_1\mu_1| > |\lambda_2\mu_2| \geq \cdots \geq |\lambda_n\mu_n| \geq 0$. Denote by $u^{(t)}$ the tth iteration of applying $v \mapsto T(v, v)/||T(v, v)||$ to u. To simplify our calculation, denote by $\overline{u}^{(t)}$ the tth iteration of apply $v \mapsto T(v, v)$. Show that:

$$\overline{u}^{(t)} = \sum_{i=1}^n \lambda_i^{2^t - 1} \mu_i^{2^t} v_i.$$

[1] **Definition.** *Let $V_1, \dots, V_k, W$ be real vector spaces. A* multilinear map $f : V_1 \times \cdots \times V_k \to W$ *is a map that is linear in each coordinate. That is, fixing any $k-1$ coordinates with $v_1 \in V_1, \dots, v_k \in V_k$ produces a linear map $f_i : V_i \to W$,*

$$f_i(v) := f(v_1, \dots, v_{i-1}, v, v_{i+1}, \dots, v_k).$$

A multilinear map on k coordinates is called k-linear.

(h) Now show that $u^{(t)} = \overline{u}^{(t)}/||\overline{u}^{(t)}||$.

(i) Using part (h), show that:

$$||v_1 - u^{(t)}||^2 = \left|\frac{\lambda_2\mu_2}{\lambda_1\mu_1}\right|^{2^{t+1}} \cdot O(1).$$

This shows that convergence of $||v_1 - u^{(t)}|| \to 0$ is quadratic in rate with respect to the analogous eigengap $|\lambda_2\mu_2/\lambda_1\mu_1|$. Notice further that depending on μ_1 and μ_2, different initializations of u will converge to different eigenvectors v_i, while in the order 2 case (i.e. matrices), almost all choices will converge to v_1.