CS 189 / 289 Introduction to Machine Learning
Fall 2023 Jennifer Listgarten, Jitendra Malik HW4

Due 10/20/23 11:59 pm PT

- Homework 4 consists of both written and coding questions.
- We prefer that you typeset your answers using LaTeX or other word processing software. If you haven't yet learned LaTeX, one of the crown jewels of computer science, now is a good time! Neatly handwritten and scanned solutions will also be accepted for the written questions.
- In all of the questions, **show your work**, not just the final answer.

Deliverables:

1. Submit a PDF of your homework to the Gradescope assignment entitled "HW 4 Write-Up". **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.
    - In your write-up, please state with whom you worked on the homework. This should be on its own page and should be the first page that you submit.
    - In your write-up, please copy the following statement and sign your signature underneath. If you are using LaTeX, you can type your full name underneath instead. We want to make it *extra* clear so that no one inadvertently cheats.

      "I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."
    - **Replicate all of your code in an appendix**. Begin code for each coding question on a fresh page. Do not put code from multiple questions in the same page. When you upload this PDF on Gradescope, *make sure* that you assign the relevant pages of your code from the appendix to correct questions.

1 PCA and Least Squares

Consider the ridge regression estimator,

$$\widehat{w}_{\text{ridge}} := \arg\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_2^2,$$

where $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$. Suppose that X has singular value decomposition $X = U\Sigma V^\top = \sum_{i=1}^d \sigma_i u_i v_i^\top$, where $U \in \mathbb{R}^{n \times d}$, $\Sigma \in \mathbb{R}^{d \times d}$, and $V \in \mathbb{R}^{d \times d}$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$ are the diagonal components of Σ.

(a) Show that

$$\widehat{w}_{\text{ridge}} = \sum_{i=1}^d \rho_\lambda(\sigma_i) v_i u_i^\top y$$

for some function $\rho_\lambda(\sigma)$ that you will determine. What is $\rho_\lambda(\sigma)$ for $\widehat{w}_{\text{ridge}}$? What is $\rho_\lambda(\sigma)$ for $\widehat{w}_{\text{OLS}} = \arg\min_w \|Xw - y\|_2^2$?

(b) The ordinary least squares regression problem on the reduced k-dimensional PCA feature space (PCA-OLS) can be written

$$\hat{w}_{\text{PCA}} = \arg\min_{w \in \mathbb{R}^k} \|XV_k w - y\|^2,$$

where V_k is a matrix whose columns are the first k right singular vectors of X. This expression embeds the raw feature vectors onto the top k principal components by the transformation $V_k^\top x_i$. Assume the PCA dimension is less than the rank of the data matrix, $k \leq r$, which implies that the matrix of PCA embedded data matrix XV_k has full rank. Write down the expression for the optimizer $\hat{w}_{\text{PCA}} \in \mathbb{R}^k$ in terms of U, y and the singular values of $\mathbf{X}$. Then, rewrite $\hat{w}_{\text{PCA}}$ in the same form as in part (a). In doing so, you should define the form of the PCA-OLS spectral function ρ_k.

Hint: Just as V_k is a "shortened" version of V, you may want shortened version of U and Σ. Knowing that $V^\top V = I$, what is the value of $V_k^\top V$?

(c) Compare $\hat{w}_{\text{OLS}}$, $\hat{w}_{\text{PCA}}$, and $\hat{w}_{\text{ridge}}$ that you derived above as functions of σ and explain how the relationships between them vary for different values of λ. How do ridge regression and PCA-OLS deal with overfitting?

Now, we will consider a slightly different way of performing PCA, using the matrix $P_k = V_k V_k^\top$. Throughout, you may assume $\sigma_k > \sigma_{k+1}$.

(d) Show that you can express $XP_k = U\Sigma_k V^\top$, where $\Sigma_k \in \mathbb{R}^{d \times d}$ is a diagonal matrix for you to define. What is XP_k if $k \geq \text{rank}(X)$?

(e) Consider the ridge-PCA estimator

$$\widehat{w}_{\text{RP}} := \arg\min_w \|XP_k w - y\|_2^2 + \lambda \|w\|_2^2.$$

Show that the predictor $X\widehat{w}_{\text{RP}}$ can be written as

$$X\widehat{w}_{\text{RP}} = \sum_{i=1}^{d} \rho_{k,\lambda}(\sigma_i) u_i u_i^\top y,$$

where $\rho_{k,\lambda}$ is a function for you to define.

Hint: The instructions ask you to find a formula for the predictor $X\widehat{w}_{\text{RP}}$*, not for the parameter* $\widehat{w}$.

2 Random Feature Embeddings

In this question, we revisit the task of dimensionality reduction. Dimensionality reduction is useful for several purposes, including visualization, storage, faster computation, etc. We can formalize dimensionality reduction as an embedding function, or *embedding*, $\psi : \mathbb{R}^d \to \mathbb{R}^k$, which maps data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with d-dimensional features to reduced data points $\psi(\mathbf{x}_1), \ldots, \psi(\mathbf{x}_n)$ with k-dimensional features.

For the reduced data to remain useful, it may be necessary for the reductions to preserve some properties of the original data. Often, geometric properties like distance and inner products are important for machine learning tasks. And as a result, we may want to perform dimensionality reduction while ensuring that we approximately maintain the pairwise distances and inner products.

While you have already seen many properties of PCA so far, in this question we investigate whether random feature embeddings are a good alternative for dimensionality reduction. A few advantages of random feature embeddings over PCA can be: (1) PCA is expensive when the underlying dimension is high and the number of principal components is also large (however note that there are several very fast algorithms dedicated to doing PCA), (2) PCA requires you to have access to the feature matrix for performing computations. The second requirement of PCA is a bottleneck when you want to take only a low dimensional measurement of a very high dimensional data, e.g., in FMRI and in compressed sensing. In such cases, one needs to design an embedding scheme before seeing the data. We now turn to a concrete setting to study a few properties of PCA and random feature embeddings.

Suppose you are given n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^d$.

Notation: The symbol $[n]$ stands for the set $\{1, \ldots, n\}$.

(a) Now consider an arbitrary embedding $\psi : \mathbb{R}^d \mapsto \mathbb{R}^k$ which preserves all pairwise distances and norms up-to a multiplicative factor for all points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in the data set, that is,

$$(1-\epsilon)\|\mathbf{x}_i\|^2 \le \|\psi(\mathbf{x}_i)\|^2 \le (1+\epsilon)\|\mathbf{x}_i\|^2 \qquad \text{for all } i \in [n], \quad \text{and} \tag{1}$$

$$(1-\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \le \|\psi(\mathbf{x}_i) - \psi(\mathbf{x}_j)\|^2 \le (1+\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \qquad \text{for all } i, j \in [n], \tag{2}$$

where $0 < \epsilon \ll 1$ is a small scalar. Further assume that $\|\mathbf{x}_i\| \le 1$ for all $i \in [n]$. **Show that the embedding ψ satisfying equations** (2) **and** (1) **preserves *each pairwise inner product*:**

$$|\psi(\mathbf{x}_i)^\top \psi(\mathbf{x}_j) - (\mathbf{x}_i^\top \mathbf{x}_j)| \le C\epsilon, \qquad \text{for all } i, j \in [n], \tag{3}$$

for some constant C. Thus, we find that if an embedding approximately preserves distances and norms up to a small multiplicative factor, and the points have bounded norms, then inner products are also approximately preserved upto an additive factor.

Hint: Break up the problem into showing that $\psi(\mathbf{x}_i)^\top \psi(\mathbf{x}_j) - (\mathbf{x}_i^\top \mathbf{x}_j) \ge C\epsilon$, and $\psi(\mathbf{x}_i)^\top \psi(\mathbf{x}_j) - (\mathbf{x}_i^\top \mathbf{x}_j) \le C\epsilon$. The constant $C = 3$ should work, though you can use a larger constant if you need. You may also want to use the Cauchy-Schwarz inequality.

(b) Now we consider the *random feature embedding* using a Gaussian matrix. In next few parts, we work towards proving that if the dimension of embedding is moderately big, then with high

probability, the random embedding preserves norms and pairwise distances approximately as described in equations (2) and (1).

Consider the random matrix $\mathbf{J} \in \mathbb{R}^{k \times d}$ with each of its entries being i.i.d. $\mathcal{N}(0,1)$ and consider the map $\psi_{\mathbf{J}} : \mathbb{R}^d \mapsto \mathbb{R}^k$ such that $\psi_{\mathbf{J}}(\mathbf{x}) = \frac{1}{\sqrt{k}}\mathbf{J}\mathbf{x}$. **Show that for any *fixed non-zero vector* $\mathbf{u}$, the random variable $\dfrac{\|\psi_{\mathbf{J}}(\mathbf{u})\|^2}{\|\mathbf{u}\|^2}$ can be written as**

$$\frac{1}{k}\sum_{i=1}^{k} Z_i^2$$

where Z_i's are i.i.d. $\mathcal{N}(0,1)$ random variables.

(c) For any fixed pair of indices $i \neq j$, define the events

$$A_{ij} := \left\{ \frac{\left\|\psi_{\mathbf{J}}(\mathbf{x}_i) - \psi_{\mathbf{J}}(\mathbf{x}_j)\right\|^2}{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2} \in (1-\epsilon, 1+\epsilon) \right\}.$$

which corresponds to the event that the embedding $\psi_{\mathbf{J}}$ approximately preserves the angles between $\mathbf{x}_i$ and $\mathbf{x}_j$. In this part, we show that A_{ij} occurs with high probability.

To do this, you will use the fact that for independent random variables $Z_i \sim \mathcal{N}(0,1)$, we have the following probability bound

$$\mathbb{P}\left[\left|\frac{1}{k}\sum_{i=1}^{k} Z_i^2\right| \notin (1-t, 1+t)\right] \leq 2e^{-kt^2/8}, \qquad \text{for all } t \in (0,1).$$

Note that this bound suggests that $\sum_{i=1}^{k} Z_i^2 \approx k = \sum_{i=1}^{k} \mathbb{E}[Z_i^2]$ with high probability. In other words, sum of squares of Gaussian random variables concentrates around its mean with high probability. **Using this bound and the previous subproblem, show that**

$$\mathbb{P}\left[A_{ij}^c\right] \leq 2e^{-k\epsilon^2/8},$$

where A_{ij}^c denotes the complement of the event A_{ij}.

(d) Using the previous problem, now **show that if $k \geq \frac{16}{\epsilon^2}\log\left(\frac{N}{\delta}\right)$, then**

$$\mathbb{P}\left[\text{for all } i, j \in [n], i \neq j, \frac{\left\|\psi_{\mathbf{J}}(\mathbf{x}_i) - \psi_{\mathbf{J}}(\mathbf{x}_j)\right\|^2}{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2} \in (1-\epsilon, 1+\epsilon)\right] \geq 1 - \delta.$$

That is show that for k large enough, with high probability the random feature embedding $\psi_{\mathbf{J}}$ approximately preserves the pairwise distances. Using this result, we can conclude that random feature embedding serves as a good tool for dimensionality reduction if we project to enough number of dimensions. This result is popularly known as the *Johnson-Lindenstrauss Lemma*.

Hint 1: Let

$$\mathcal{A} := \left\{ \text{for all } i, j \in [n], i \neq j, \frac{\left\|\psi_{\mathbf{J}}(\mathbf{x}_i) - \psi_{\mathbf{J}}(\mathbf{x}_j)\right\|^2}{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2} \in (1-\epsilon, 1+\epsilon) \right\}$$

denote the event whose probability we would like to lower bound. Express the complement $\mathcal{A}^c$ in terms of the events A_{ij}^c, and try to apply a union bound to these events.

(e) Suppose there are two clusters of points $S_1 = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ and $S_2 = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ which are far apart, i.e., we have

$$d^2(S_1, S_2) = \min_{u \in S_1, v \in S_2} \|u - v\|^2 \geq \gamma.$$

Then using the previous part, **show that the random feature embedding $\psi_{\mathbf{J}}$ also approximately maintains the distance between the two clusters if k is large enough, that is, with high probability**

$$d^2(\psi_{\mathbf{J}}(S_1), \psi_{\mathbf{J}}(S_2)) = \min_{u \in S_1, v \in S_2} \left\|\psi_{\mathbf{J}}(\mathbf{u}) - \psi_{\mathbf{J}}(\mathbf{v})\right\|^2 \geq (1-\epsilon)\gamma \quad \text{if } k \geq \frac{C}{\epsilon^2}\log(m+n)$$

for some constant C. Note that such a property can help in several machine learning tasks. For example, if the clusters of features with different labels were far in the original dimension, then this problem shows that they will remain far in the smaller dimension.Therefore, a machine learning model can perform well even with the randomly projected data.

3 Interpreting Neural Nets Using T-SNE

For this question, please go through the Google Colab Notebook here to complete the code.

In lecture, you have learned about how t-SNE is a method for nonlinear dimensionality reduction. This is particularly useful for analyzing many real-world datasets in which the data can be categorized according to underlying labels. In this question, you will examine the effect that a neural network has on the t-SNE of such a dataset.

(a) We will work with the CIFAR-10 dataset for this problem, in which the image data is categorized into 10 classes. Flatten the images and take the t-SNE of the training dataset. Plot the t-SNE embeddings and color-code each data point according to its class. Explain what you observe.

(b) Now, we have provided a trained neural network for you to analyze. Save it to your Google Drive so that you can access it from the Colab notebook. The model consists of several convolutional layers and a few linear layers. Calculate its accuracy on the test data.

(c) Instead of taking the t-SNE of the training dataset directly, we will take the t-SNE of the *features* of the neural network when the training dataset is given as input. Using the "hook" functions provided in the notebook, save the outputs of the third convolutional layer of the network for each input data point. Take the t-SNE of these outputs and color-code each point according to its class. Explain what you observe.

(d) Do the same as the above except for both the first and second linear layers of the network. Overall, what does it look like the network is doing to the data? How might this change depending on the network's accuracy?

4 Astronomer's conundrum

(Numbers completely made up; don't use for reference.)

As machine learning invades everything in the world, you find that you can use machine learning to classify celestial bodies (surprise, surprise!). Conveniently, you lose all your precious data and only have the rates at which these celestial bodies lose their mass via stellar wind.

There are three types of celestial bodies that you want to classify: dwarfs, giants, and black holes. Dwarfs slowly lose their mass; giants rapidly lose their mass; black holes, on the other hand, gain mass by absorbing stuff (i.e., they lose mass at negative rates). We also know that 60% of all the celestial bodies are dwarfs, 30% are giants, and 10% are black holes.

Before we continue, let's familiarize ourselves with a kind of probability distribution called the *exponential distribution*. The probability density function of an exponential distribution with parameter λ has the following form:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Note the pdf decreases monotonically on $[0, +\infty)$.

The rate at which a dwarf or a giant loses its mass is exponentially distributed with parameters λ_d and λ_g, respectively. The rate at which a black hole *gains* mass is also exponentially distributed, with parameter λ_b.

(a) Knowing only that the expected value of a exponentially distributed random variable with parameter λ is $\frac{1}{\lambda}$, estimate the parameter for dwarfs, λ_d, using a sample of the mass loss rates of dwarfs: $\{0.6, 1.3, 0.1, 0.3, 0.2\}$.

(b) Given that $\lambda_g = 1$, $\lambda_b = 3$, and using the λ_e you estimated in the previous question, find the Bayes classifier for the three classes. Assume we employ a 0-1 loss.

(c) Following on from the previous question, find the risk of your Bayes classifier. Feel free to use WolframAlpha or some other software for the integration.

5 Risk Minimization with Doubt

Suppose we have a classification problem with classes labeled $1, \ldots, c$ and an additional "doubt" category labeled $c+1$. Let $f : \mathbb{R}^d \to \{1, \ldots, c+1\}$ be a decision rule. Define the loss function

$$L(f(\mathbf{x}), y) = \begin{cases} 0 & \text{if } f(\mathbf{x}) = y \quad f(\mathbf{x}) \in \{1, \ldots, c\}, \\ \lambda_c & \text{if } f(\mathbf{x}) \neq y \quad f(\mathbf{x}) \in \{1, \ldots, c\}, \\ \lambda_d & \text{if } f(\mathbf{x}) = c+1 \end{cases} \tag{4}$$

where $\lambda_c \geq 0$ is the loss incurred for making a misclassification and $\lambda_d \geq 0$ is the loss incurred for choosing doubt. In words this means the following:

- When you are correct, you should incur no loss.
- When you are incorrect, you should incur some penalty λ_c for making the wrong choice.
- When you are unsure about what to choose, you might want to select a category corresponding to "doubt" and you should incur a penalty λ_d.

In lecture, you saw a definition of risk over the expectation of data points. We can also define the risk of classifying a new individual data point $\mathbf{x}$ as class $f(\mathbf{x}) \in \{1, 2, \ldots, c+1\}$, and reason about what the risk would be for all possible values of $\mathbf{x}$. We define the risk as

$$R(f(\mathbf{x})|\mathbf{x}) = \sum_{i=1}^{c} L(f(\mathbf{x}), i)\, P(Y = i|\mathbf{x}).$$

(a) Show that the following policy $f_{opt}(x)$ obtains the minimum risk:

- (**R1**) Find the non-doubt class i such that $P(Y = i|\mathbf{x}) \geq P(Y = j|\mathbf{x})$ for all j, meaning you pick the class with the highest probability given x.
- (**R2**) Choose class i if $P(Y = i|\mathbf{x}) \geq 1 - \frac{\lambda_d}{\lambda_c}$
- (**R3**) Choose doubt otherwise.

Hint: It will first help you to approach the risk function on a case-by-case basis to help simplify the expression. What is the risk if we choose the "doubt" class? What is it if we choose a non-doubt class as our prediction?

In order to prove that $f_{opt}(x)$ minimizes risk, consider proof techniques that show that $f_{opt}(x)$ "stays ahead" of all other policies that *don't* follow these rules. For example, you could take a proof-by-contradiction approach: assume there exists some other policy, say $f'(x)$, that minimizes risk more than $f_{opt}(x)$. What are the scenarios where the predictions made by $f_{opt}(x)$ and $f'(x)$ might differ? In these scenarios, and based on the rules above that $f_{opt}(x)$ follows, why would $f'(x)$ not be able to beat $f_{opt}(x)$ in risk minimization?

(b) How would you modify your optimum decision rule if $\lambda_d = 0$? What happens if $\lambda_d > \lambda_c$? Explain why this is or is not consistent with what one would expect intuitively.