CS 189 / 289 Introduction to Machine Learning
Fall 2022 Jennifer Listgarten, Jitendra Malik HW0

Due 09/07/22 11:59 pm PT

- Homework 0 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.
- **Start early. This is a long assignment. Most of the material is prerequisite material not covered in lecture; you are responsible for finding resources to understand it.**

Deliverables:

1. Submit a PDF of your homework to the Gradescope assignment entitled "HW0 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 your code in an appendix**. Begin code for each coding question in 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 appendix to correct questions.

HW0, ©UCB CS 189 / 289, Fall 2022. All Rights Reserved. This may not be publicly shared without explicit permission.

1 Gradients and Derivatives (13 points)

What is the derivative of the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 2x$? From basic calculus, we know the answer is 2 everywhere. Fundamentally, the derivative is a best linear approximation: for $f(x) = 2x$, anywhere we look, the linear transformation that best approximates f is "multiplication by 2" (in fact, f is exactly equal to this linear transformation). More precisely, for any differentiable function f, the derivative at a, $\frac{df}{dx}(a)$, is the *best linear approximation of f at a*. That is, $\frac{df}{dx}(a)$ is equal to a linear function L such that $f(x) \approx f(a) + L(x - a)$ for all x near a. We can view $\frac{df}{dx}(a)$ as the rate of change of f near a.

This perspective is handy in higher dimensions. Take the function $f : \mathbb{R}^n \to \mathbb{R}^k : f(x) = Ax$ where A is a matrix in $\mathbb{R}^{k \times n}$. How can we compute the derivative now that one scalar has been replaced by a matrix, and the other by a vector? We can simply recognize the fact that $\frac{df}{dx}(x)$ must be the linear transformation that best approximates f at x. But f is nothing more than multiplication by A, a linear transformation. Thus, $\frac{df}{dx}(x) = A$ for all $x \in \mathbb{R}^n$.

The gradient of a function of a vector When $f : \mathbb{R}^n \to \mathbb{R}$ maps a vector to a scalar, the derivative at each point is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}$, which can be represented as a row vector $\frac{df}{dx}(x) \in \mathbb{R}^{1 \times n}$, i.e. $f(x + \Delta) \approx f(x) + \frac{df}{dx}(x)\Delta$ for small $\Delta \in \mathbb{R}^n$. The *gradient* in this case is the transpose of the derivative, $\nabla_x f(x) = \frac{df}{dx}(x)^\top \in \mathbb{R}^n$. Why do we bother to define the gradient? The fact that the gradient is the same shape as the input is convenient, and its ith entry is the partial derivative of f with respect to the ith entry of the input:

$$\nabla_x f(x)_i = \frac{\partial f}{\partial x_i}(x).$$

The gradient of a function of a matrix When $f : \mathbb{R}^{n \times m} \to \mathbb{R}$ maps a matrix to a scalar, its derivative at X, $\frac{df}{dX}(X)$, is a linear transformation from $\mathbb{R}^{n \times m}$ to $\mathbb{R}$. For any linear transformation $T : \mathbb{R}^{n \times m} \to \mathbb{R}$, there is a matrix $A \in \mathbb{R}^{n \times m}$ such that:

$$\forall X \in \mathbb{R}^{n \times m},\ T(X) = \langle A, X \rangle := \mathrm{Tr}\left(A^\top X\right).$$

Thus, T can be represented by the matrix A. Don't be intimidated if you haven't seen this before: $\mathrm{Tr}(A^\top X)$ is merely the matrix dot product between A and X, which is very similar to the vector dot product. It is equivalent to multiplying A and X elementwise and summing the entries of the resulting matrix, or flattening out both matrices and computing their vector dot product. We can now define the gradient $\nabla_X f(X) \in \mathbb{R}^{n \times m}$ as the matrix such that $\left(\frac{df}{dX}(X)\right)(A) = \langle \nabla_X f(X), A \rangle$, for all $A \in \mathbb{R}^{n \times m}$. Therefore, for small $\Delta \in \mathbb{R}^{n \times m}$,

$$f(X + \Delta) \approx f(X) + \left(\frac{df}{dX}(X)\right)(\Delta) = f(X) + \langle \nabla_X f(X), \Delta \rangle.$$

The gradient is also expressible as the matrix of partial derivatives of f:

$$\nabla_X f(X)_{ij} = \frac{\partial f}{\partial X_{ij}}(X).$$

The Hessian Finally, we define the Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$ as the matrix

$$\nabla_x^2 f(x)_{ij} = \frac{\partial f}{\partial x_i \partial x_j}(x).$$

The Hessian equals the derivative of the gradient, $\nabla_x^2 f(x) = \frac{d\nabla_x f}{dx}(x)$. If f has continuous second order partial derivatives, this matrix is symmetric.

(a) (1 point) Let $w \in \mathbb{R}^n$. Compute the gradient $\nabla_x f(x)$ of

$$f : \mathbb{R}^n \to \mathbb{R},\ f(x) = w^\top x.$$

(b) (2 points) Let $A \in \mathbb{R}^{n \times n}$. Compute the gradient $\nabla_x f(x)$ of

$$f : \mathbb{R}^n \to \mathbb{R},\ f(x) = x^\top A x.$$

Hint: you can use the product rule of derivatives. For differentiable $f : V \to \mathbb{R}^{l \times k}$ and $g : V \to \mathbb{R}^{k \times m}$ where V is any vector space, we have $\left(\frac{d(fg)}{dx}(x)\right)(v) = \left(\frac{df}{dx}(x)\right)(v)g(x) + f(x)\left(\frac{dg}{dx}(x)\right)(v)$ for any $v \in V$.

(c) (1 point) Compute the Hessian $\nabla_x^2 f(x)$ of the above function.

(d) (2 points) Let $A \in \mathbb{R}^{m \times n}$, $y \in \mathbb{R}^m$. Compute the gradient $\nabla_x f(x)$ of

$$f : \mathbb{R}^n \to \mathbb{R},\ f(x) = ||Ax - y||_2^2.$$

Hint: you can use the chain rule of derivatives. Given vector spaces U, V, W, for $f : U \to V$, $z \mapsto f(z)$ and $g : W \to U$, $x \mapsto g(x)$ differentiable, we have $\frac{d(f \circ g)}{dx}(x) = \frac{df}{dz}(g(x)) \circ \frac{dg}{dx}(x)$.

(e) (1 point) Let $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$. Compute the gradient $\nabla_A f(A)$ of

$$f : \mathbb{R}^{m \times n} \to \mathbb{R},\ f(A) = u^\top A v.$$

(f) (3 points) Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Compute the gradient $\nabla_A f(A)$ of

$$f : \mathbb{R}^{m \times n} \to \mathbb{R},\ f(A) = ||Ax - y||_2^2.$$

Hint: $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

(g) (3 points) Consider the function that maps a vector to its maximum entry, $x \mapsto \max_i x_i$. While this function is non-smooth, a common trick in machine learning is to use a smooth approximation, *LogSumExp*, defined as follows.

$$\mathrm{LSE} : \mathbb{R}^n \to \mathbb{R}, \mathrm{LSE}(x) = \ln\left(\sum_{i=1} e^{x_i}\right).$$

One of the nice properties of this function is that it is convex, which can be proved by showing its Hessian matrix is positive semidefinite. To that end, compute its gradient and Hessian. You do not need to prove that the Hessian is PSD.

2 Linear Algebra Review (10 points)

1. (3 points) Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Prove equivalence between these three different definitions of positive semidefiniteness (PSD).

   (a) For all $x \in \mathbb{R}^n$, $x^\top A x \geq 0$.

   (b) All the eigenvalues of A are nonnegative.

   (c) There exists a matrix $U \in \mathbb{R}^{n \times n}$ such that $A = UU^\top$.

   Mathematically, we write positive semidefiniteness as $A \succeq 0$.

2. (5 points) Now that we're equipped with different definitions of positive semidefiniteness, use them to prove the following properties of PSD matrices.

   (a) If A and B are PSD, then $2A + 3B$ is PSD.

   (b) If A is PSD, all diagonal entries of A are nonnegative: $A_{ii} \geq 0, \forall i \in [n]$.

   (c) If A is PSD, the sum of all entries of A is nonnegative: $\sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij} \geq 0$.

   (d) If A and B are PSD, then $\text{Tr}(AB) \geq 0$, where $\text{Tr}(M)$ denotes the *trace* of M.

   (e) If A and B are PSD, then $\text{Tr}(AB) = 0$ if and only if $AB = 0$.

3. (2 points) Let $A \in \mathbb{R}^{n \times n}$ be a symmetric, PSD matrix. Write $\|A\|_F$ as a function of the eigenvalues of A.
   Hint: Recall that $\|A\|_F = \sqrt{\text{Tr}(A^\top A)}$. If you haven't seen this before, you should try to prove it. However, you can accept this as a given fact for this homework assignment.

3 Probability Potpourri (11 points)

1. (2 points) Recall the covariance of two random variables X and Y is defined as $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. For a multivariate random variable Z (i.e., each index of Z is a random variable), we define the covariance matrix Σ such that $\Sigma_{ij} = \text{Cov}(Z_i, Z_j)$. Concisely, $\Sigma = \mathbb{E}[(Z - \mu)(Z - \mu)^\top]$, where μ is the mean value of the random column vector Z. Prove that the covariance matrix is always positive semidefinite (PSD).
   Hint: Use linearity of expectation.

2. (4 points) The probability that an archer hits her target when it is windy is 0.4; when it is not windy, her probability of hitting the target is 0.7. On any shot, the probability of a gust of wind is 0.3. Find the probability that

   (i) on a given shot there is a gust of wind and she hits her target.

   (ii) she hits the target with her first shot.

   (iii) she hits the target exactly once in two shots.

   (iv) there was no gust of wind on an occasion when she missed.

3. (2 points) An archery target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Arrows striking within the inner circle are awarded 4 points, arrows within the middle ring are awarded 3 points, and arrows within the outer ring are awarded 2 points. Shots outside the target are awarded 0 points.

   Consider a random variable X, the distance of the strike from the center (in feet), and let the probability density function of X be

   $$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

   What is the expected value of the score of a single strike?

4. (3 points) Let $X \sim \text{Pois}(\lambda)$, $Y \sim \text{Pois}(\mu)$. given that $X \perp\!\!\!\perp Y$, derive an expression for $\mathbb{P}(X \mid X + Y = n)$. What well-known probability distribution is this? What are its parameters?

4 Gaussian basics (11 points)

The multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$, denoted $N(\mu, \Sigma)$, has the probability density function

$$f(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}.$$

$|\Sigma|$ denotes the determinant of Σ. Here, we assume that the covariance Σ is invertible, although there are multivariate Gaussians with non-invertible covariances. You may use the following facts without proof:

1. The Gaussian pdf integrates to 1:

$$\int_{\mathbb{R}^d} f(x; \mu, \Sigma)\, \mathrm{d}x = \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\} \mathrm{d}x = 1$$

2. Change of variables formula: let f be a smooth function from $\mathbb{R}^d \to \mathbb{R}$, $b \in \mathbb{R}^d$, and $A \in \mathbb{R}^{d\times d}$ be an invertible matrix. Then, performing the change of variable $x \mapsto z = Ax + b$,

$$\int_{\mathbb{R}^d} f(x)\, \mathrm{d}x = \int_{\mathbb{R}^d} f(A^{-1}z - A^{-1}b)|A^{-1}|\, \mathrm{d}z.$$

   You don't need to worry about smoothness when applying this fact; rest assured that polynomials, exponentials, and products and compositions of smooth functions are smooth.

(a) (2 points) Let $X \sim N(\mu, \Sigma)$. Show that $\mathbb{E}[X] = \mu$.

(b) (4 points) Show that $\mathrm{Cov}(X) = \Sigma$.

(c) (2 points) Compute the moment generating function (MGF) of X: $M_X(\lambda) = \mathbb{E}[e^{\lambda^\top X}]$, where $\lambda \in \mathbb{R}^d$. Note: moment generating functions have several interesting and useful properties, one being that M_X characterizes the distribution of X: if $M_X = M_Y$, then X and Y have the same distribution.

(d) (2 points) Using the fact that MGFs determine distributions, given $A \in \mathbb{R}^{k\times d}$, $b \in \mathbb{R}^k$ identify the distribution of $AX + b$ (don't worry about covariance matrices being invertible).

(e) (1 point) Show that there exists an affine transformation of X that is distributed as the standard multivariate Gaussian, $N(0, I_d)$. (Assume Σ is invertible.)

5 Isocontours of Normal Distributions (6 points)

Let $f(\mu, \Sigma)$ be the probability density function of a normally distributed random variable in $\mathbb{R}^2$. For parts (b) and (c), write code to plot the isocontours of the following functions, each on its own separate figure. Plot at least 5 contours, enough to get a rough sense of the probability density. Default settings of commonly used contour plotting functions probably suffice for this. You are free to use Matplotlib, NumPy, and SciPy.

(a) (3 points) The spectral theorem allows us to factorize

$$\Sigma = UDU^\top, \quad D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad U = \begin{bmatrix} u_1 & u_2 \end{bmatrix} \in \mathbb{R}^{2\times 2},$$

where U has orthonormal columns u_1 and u_2 and D has positive diagonal entries (assume Σ is invertible). Consider the level set

$$S = \{x \in \mathbb{R}^2 : f(x; \mu, \Sigma) = c\},$$

that is, the set of points $x \in \mathbb{R}^2$ such that the probability density of the Gaussian evaluates to c at those points (c is some value $0 < c < \frac{1}{\sqrt{(2\pi)^d |\Sigma|}}$). Show that S is an ellipse, and compute the direction of each axis and its semilength in terms of $u_1, u_2, \lambda_1, \lambda_2$. For background, an ellipse has two perpendicular axes. We consider the direction of an axis to be a unit vector that is a scalar multiple of the vector pointing from the center of the ellipse to either endpoint of the axis. By axis semilength, we mean half the length of the line segment connecting the axis endpoints. For more info, see `https://en.wikipedia.org/wiki/Ellipse`.

(b) (2 points) $f(\mu, \Sigma)$, where $\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.

(c) (1 point) $f(\mu, \Sigma)$, where $\mu = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$.

6 Hands-on with data (10 points)

In the following problem, you will use two simple datasets to walk through the steps of a standard machine learning workflow: inspecting your data, choosing a model, implementing it, and verifying its accuracy. We have provided two datasets in the form of numpy arrays: `dataset_1.npy` and `dataset_2.npy`. You can load each using NumPy's `np.load` method; see https://numpy.org/doc for more information if you are unfamiliar with the numpy library.

Each dataset is a two-column array with the first column consisting of n scalar inputs $X \in \mathbb{R}^{n\times 1}$ and the second column consisting of n scalar labels $Y \in \mathbb{R}^{n\times 1}$. We denote each entry of X and Y with subscripts:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

and assume that y_i is a (potentially stochastic) function of x_i.

(a) (2 points) It is often useful to visually inspect your data and calculate simple statistics; this can detect dataset corruptions or inform your method. For both datasets:

(i) Plot the data as a scatter plot.

(ii) Calculate the correlation coefficient between X and Y:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

in which $\text{Cov}(X, Y)$ is the covariance between X and Y and σ_X is the standard deviation of X.

Your solution may make use of the NumPy library only for arithmetic operations, matrix-vector or matrix-matrix multiplications, matrix inversion, and elementwise exponentiation. It may not make use of library calls for calculating means, standard deviations, or the correlation coefficient itself directly.

(b) (1 point) We would like to design a function that can predict y_i given x_i and then apply it to new inputs. This is a recurring theme in machine learning, and you will soon learn about a general-purpose framework for thinking about such problems. As a preview, we will now explore one of the simplest instantiations of this idea using the class of linear functions:

$$\hat{Y} = Xw. \tag{1}$$

The parameters of our function are denoted by $w \in \mathbb{R}$. It is common to denote predicted variants of quantities with a hat, so $\hat{Y}$ is a predicted label whereas Y is a ground truth label.

We would like to find a w^* that minimizes the **squared error** $\mathcal{J}_{\text{SE}}$ between predictions and labels:

$$w^* = \arg\min_w \mathcal{J}_{\text{SE}}(w) = \arg\min_w \|Xw - Y\|_2^2.$$

Derive $\nabla_w \mathcal{J}_{\text{SE}}(w)$ and set it equal to 0 to solve for w^*. (Note that this procedure for finding an optimum relies on the convexity of $\mathcal{J}_{\text{SE}}$. You do not need to show convexity here, but it is a useful exercise to convince yourself this is valid.)

(c) (1 point) Your solution w^* should be a function of X and Y. Implement it and report its **mean squared error** (MSE) for **dataset 1**. The mean squared error is the objective $\mathcal{J}_{\text{SE}}$ from part (b) divided by the number of datapoints:

$$\mathcal{J}_{\text{MSE}}(w) = \frac{1}{n}\|Xw - Y\|_2^2.$$

Also visually inspect the model's quality by plotting a line plot of predicted $\hat{y}$ for uniformly-spaced $x \in [0, 10]$. Keep the scatter plot from part (a) in the background so that you can compare the raw data to your linear function. Does the function provide a good fit of the data? Why or why not?

(d) (1 point) We are now going to experiment with constructing new *features* for our model. That is, instead of considering models that are linear in the inputs, we will now consider models that are linear in some (potentially nonlinear) transformation of the data:

$$\hat{Y} = \Phi w = \begin{bmatrix} \phi(x_1)^\top \\ \phi(x_2)^\top \\ \vdots \\ \phi(x_n)^\top \end{bmatrix} w,$$

where $\phi(x_i), w \in \mathbb{R}^m$. Repeat part (c), providing both the mean squared error of your predictor and a plot of its predictions, for the following features on **dataset 1**:

$$\phi(x_i) = \begin{bmatrix} x_i \\ 1 \end{bmatrix}.$$

How do the plotted function and mean squared error compare? (A single sentence will suffice.)

Hint: the general form of your solution for w^* is still valid, but you will now need to use features Φ where you once used raw inputs X.

(e) (1 point) Now consider the quadratic features:

$$\phi(x_i) = \begin{bmatrix} x_i^2 \\ x_i \\ 1 \end{bmatrix}.$$

Repeat part (c) with these features on **dataset 1**, once again providing short commentary on any changes.

(f) (2 points) Repeat parts (c)-(e) with **dataset 2**.

(g) (2 points) Finally, we would like to understand which features Φ provide us with the best model. To that end, you will implement a method known as k-fold cross validation. The following are instructions for this method; deliverables for part (g) are at the end.

(i) Split **dataset 2** randomly into $k = 4$ equal sized subsets. Group the dataset into 4 distinct training / validation splits by denoting each subset as the validation set and the remaining subsets as the training set for that split.

(ii) On each of the 4 training / validation splits, fit linear models using the following 5 polynomial feature sets:

$$\phi_1(x_i) = \begin{bmatrix} x_i \\ 1 \end{bmatrix} \quad \phi_2(x_i) = \begin{bmatrix} x_i^2 \\ x_i \\ 1 \end{bmatrix} \quad \phi_3(x_i) = \begin{bmatrix} x_i^3 \\ x_i^2 \\ x_i \\ 1 \end{bmatrix} \quad \phi_4(x_i) = \begin{bmatrix} x_i^4 \\ x_i^3 \\ x_i^2 \\ x_i \\ 1 \end{bmatrix} \quad \phi_5(x_i) = \begin{bmatrix} x_i^5 \\ x_i^4 \\ x_i^3 \\ x_i^2 \\ x_i \\ 1 \end{bmatrix}$$

This step will produce 20 distinct w^* vectors: one for each dataset split and featurization ϕ_j.

(iii) For each feature set ϕ_j, average the training and validation mean squared errors over all training splits.

It is worth thinking about what this extra effort has bought us: by splitting the dataset into subsets, we were able to use all available datapoints for model fitting while still having held-out datapoints for evaluation for any particular model.

Deliverables for part (g): Plot the training mean squared error and the validation mean squared error on the same plot as a function of the largest exponent in the feature set. Use a log scale for the y-axis. Which model does the training mean squared error suggest is best? Which model does the validation mean squared error suggest is best?