Homework #0

Due: January 26, 2024 at 11:59 PM

Welcome to CS181! The purpose of this assignment is to help assess your readiness for this course. It will be graded for completeness and effort. **Areas of this assignment that are difficult are an indication of areas in which *you* need to self-study. During the term, the staff will be prioritizing support for new material taught in CS181 over teaching prerequisites.**

1. Please type your solutions after the corresponding problems using this LaTeX template, and start each problem on a new page.
2. Please submit the **writeup PDF to the Gradescope assignment 'HW0'**. Remember to assign pages for each question.
3. Please submit your **LaTeX file and code files (i.e., anything ending in `.py`, `.ipynb`, or `.tex`) to the Gradescope assignment 'HW0 - Supplemental'**.

Problem 1 (Modeling Linear Trends - Linear Algebra Review)

In this class we will be exploring the question of "how do we model the trend in a dataset" under different guises. In this problem, we will explore the algebra of modeling a linear trend in data. We call the process of finding a model that capture the trend in the data, "fitting the model."

Learning Goals: In this problem, you will practice translating machine learning goals ("modeling trends in data") into mathematical formalism using linear algebra. You will explore how the right mathematical formalization can help us express our modeling ideas unambiguously and provide ways for us to analyze different pathways to meeting our machine learning goals.

Let's consider a dataset consisting of two points $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$, where x_n, y_n are scalars for $n = 1, 2$. Recall that the equation of a line in 2-dimensions can be written: $y = w_0 + w_1 x$.

1. Write a system of linear equations determining the coefficients w_0, w_1 of the line passing through the points in our dataset $\mathcal{D}$ and analytically solve for w_0, w_1 by solving this system of linear equations (i.e., using substitution). Please show your work.

2. Write the above system of linear equations in matrix notation, so that you have a matrix equation of the form $\mathbf{y} = \mathbf{X}\mathbf{w}$, where $\mathbf{y}, \mathbf{w} \in \mathbb{R}^2$ and $\mathbf{X} \in \mathbb{R}^{2\times 2}$. For full credit, it suffices to write out what $\mathbf{X}$, $\mathbf{y}$, and $\mathbf{w}$ should look like in terms of x_1, x_2, y_1, y_2, w_0, w_1, and any other necessary constants. Please show your reasoning and supporting intermediate steps.

3. Using properties of matrices, characterize exactly when an unique solution for $\mathbf{w} = (w_0\ w_1)^T$ exists. In other words, what must be true about your dataset in order for there to be a unique solution for $\mathbf{w}$? When the solution for $\mathbf{w}$ exists (and is unique), write out, as a matrix expression, its analytical form (i.e., write $\mathbf{w}$ in terms of $\mathbf{X}$ and $\mathbf{y}$).

   Hint: What special property must our $\mathbf{X}$ matrix possess? What must be true about our data points in $\mathcal{D}$ for this special property to hold?

4. Compute $\mathbf{w}$ by hand via your matrix expression in (3) and compare it with your solution in (1). Do your final answers match? What is one advantage for phrasing the problem of fitting the model in terms of matrix notation?

5. In real-life, we often work with datasets that consist of hundreds, if not millions, of points. In such cases, does our analytical expression for $\mathbf{w}$ that we derived in (3) apply immediately to the case when $\mathcal{D}$ consists of more than two points? Why or why not?

Solution

1. To begin you can take the equation $y = w_0 + w_1x$ and then plug in the values x_1, x_2, y_1, y_2 to get $y_1 = w_0 + w_1x_1$ and $y_2 = w_0 + w_1x_2$. From there, we can manipulate the first equation to get that $w_0 = y_1 - w_1x_1$. Then, we can use substitution to plug the value of w_0 into the second equation and get that $y_2 = y_1 - w_1x_1 + w_1x_2$.
This can be manipulated to get $-w_1x_1 + w_1x_2 = y_2 - y_1$ then $w_1(-x_1 + x_2) = y_2 - y_1$ which gets simplified to $w_1 = \frac{y_2 - y_1}{x_2 - x_1}$.
Now that we have w_1 solved for, it is easy to substitute that value into the equation $w_0 = y_1 - w_1x_1$ to get that $w_0 = y_1 - \frac{x_1y_2 - x_1y_1}{x_2 - x_1}$

2. For this system of linear equations, we can express it as $y = Xw$ such that
$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}, w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$
This is because if you do the multiplication on the right side of the equation, you get the augmented matrix:
$\begin{bmatrix} y_1 & | & w_0 + w_1x_1 \\ y_2 & | & w_0 + w_1x_2 \end{bmatrix}$ Which is the same as the system of linear equations found in Part 1.

3. In order for there to be a unique solution for $w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$, it must be true that $x_1 \neq x_2$ and $y_1 \neq y_2$. In other words, the matrix X must be invertible in order for w to have a unique solution. This can be expressed with $w = X^{-1}y$

4. To begin, we must calculate the value of X^{-1} and get that $X^{-1} = \frac{1}{x_1 - x_2}\begin{bmatrix} -x_2 & x_1 \\ 1 & -1 \end{bmatrix}$ Then we can multiply $X^{-1}y$ to get that $w = \frac{1}{x_1 - x_2}\begin{bmatrix} x_2y_1 + x_1y_2 \\ y_1 - y_2 \end{bmatrix}$
As linear equations this means there is $w_0 = \frac{-x_2y_1 + x_1y_2}{x_1 - x_2}$ and $w_1 = \frac{y_1 - y_2}{x_1 - x_2}$, both of which agree with the solution in Part 1. One of the advantages of phrasing the problem this way, in terms of matrix notation is that it is a cleaner process that produces simultaneous equations for w_0 and w_1 rather than the messier process of having to use substitution and manipulating two separate linear equations.

5. In the case of a dataset with hundreds or more points, the analytical expression for w does not automatically apply. This is because the expression for w found in part 3 implies and works based on a linear relationship between the points, which as stated occurs when X is invertible. That is not to say that the relationship never holds, if all points in the dataset have a linear relationship than the relationship still holds since they will all follow the same linear properties.

Problem 2 (Optimizing Objectives - Calculus Review)

In this class, we will write real-life goals we want our model to achieve into a mathematical expression and then find the optimal settings of the model that achieves these goals. The formal framework we will employ is that of mathematical optimization. Although the mathematics of optimization can be quite complex and deep, we have all encountered basic optimization problems in our first calculus class!

Learning Goals: In this problem, we will explore how to formalize real-life goals as mathematical optimization problems. We will also investigate under what conditions these optimization problems have solutions.

In her most recent work-from-home shopping spree, Nari decided to buy several house plants. *Her goal is to make them to grow as tall as possible.* After perusing the internet, Nari learns that the height y in mm of her Weeping Fig plant can be directly modeled as a function of the oz of water x she gives it each week:

$$y = -3x^2 + 72x + 70.$$

1. Based on the above formula, is Nari's goal achievable: does the plant have a maximum height? Why or why not? Does her goal have a unique solution - i.e. is there one special watering schedule that would acheive the maximum height (if it exists)?

   Hint: plot this function. In your solution, words like "convex" and "concave" may be helpful.

2. Using calculus, find how many oz per week should Nari water her plant in order to maximize its height. With this much water, how tall will her plant grow?

   Hint: solve analytically for the critical points of the height function (i.e., where the derivative of the function is zero). For each critical point, use the second-derivative test to identify if each point is a max or min point, and use arguments about the global structure (e.g., concavity or convexity) of the function to argue whether this is a local or global optimum.

Now suppose that Nari want to optimize both the amount of water x_1 (in oz) *and* the amount of direct sunlight x_2 (in hours) to provide for her plants. After extensive research, she decided that the height y (in mm) of her plants can be modeled as a two variable function:

$$y = f(x_1, x_2) = \exp\left(-(x_1 - 2)^2 - (x_2 - 1)^2\right)$$

3. Using `matplotlib`, visualize in 3D the height function as a function of x_1 and x_2 using the `plot_surface` utility for $(x_1, x_2) \in (0, 6) \times (0, 6)$. Use this visualization to argue why there exists a unique solution to Nari's optimization problem on the specified intervals for x_1 and x_2.

   Remark: in this class, we will learn about under what conditions do *multivariate* optimization problems have unique global optima (and no, the second derivative test doesn't exactly generalize directly). Looking at the visualization you produced and the expression for $f(x_1, x_2)$, do you have any ideas for why this problem is guaranteed to have a global maxima? You do not need to write anything responding to this – this is simply food for thought and a preview for the semester.

Solution

1. Based on this formula, Nari's goal is achievable. The plant has a maximum height since the graph for $-3x^2+72x+70$ is concave down and so by nature of concave graphs, will inevitably have a global maximum, which will tell Nari the maximum height and optimal amount of water per week.

2. In order to find the optimal amount of water per week, we can take the derivative of the equation, getting that $y' = -6x+72$. Then we can solve for the zeroes of this equation to get the critical points. This leaves us with the only critical point being at $x = 12$. Then, with this critical point, we can take the second derivative, $y'' = -6$ to discover that at $x = 12$, the concavity is negative, which means that the critical point is a maximum. Lastly, by the property of a concave down parabolic graph, there is only one maximum point, so we know that $x = 12$ is a global maximum. From these, we can determine that if Nari waters her plant 12 oz of water per week, the plant can reach a maximum of 502 mm.

3. There exists a unique solution to Nari's optimization problem because when you look at the graph that is made with this equation, it is apparent that there is only one local maximum on the interval of interest and so there is only optimal answer within the area that we are concerned.

Problem 3 (Reasoning about Randomness - Probability and Statistics Review)

In this class, one of our main focuses is to model the unexpected variations in real-life phenomena using the formalism of random variables. In this problem, we will use random variables to model how much time it takes an USPS package processing system to process packages that arrive in a day.

Learning Goals: In this problem, you will analyze random variables and their distributions both analytically and computationally. You will also practice drawing connections between said analytical and computational conclusions.

Consider the following model for packages arriving at the US Postal Service (USPS):

- Packages arrive randomly in any given hour according to a Poisson distribution. That is, the number of packages in a given hour N is distributed $Pois(\lambda)$, with $\lambda = 3$.
- Each package has a random size S (measured in in^3) and weight W (measured in pounds), with joint distribution

$$(S, W)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ with } \boldsymbol{\mu} = \begin{bmatrix} 120 \\ 4 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix}.$$

- Processing time T (in seconds) for each package is given by $T = 60 + 0.6W + 0.2S + \epsilon$, where ϵ is a random noise variable with Gaussian distribution $\epsilon \sim \mathcal{N}(0, 5)$.

For this problem, you may find the `multivariate_normal` module from `scipy.stats` especially helpful. You may also find the `seaborn.histplot` function quite helpful.

1. Perform the following tasks:
   (a) Visualize the Bivariate Gaussian distribution for the size S and weight W of the packages by sampling 500 times from the joint distribution of S and W and generating a bivariate histogram of your S and W samples.
   (b) Empirically estimate the most likely combination of size and weight of a package by finding the bin of your bivariate histogram (i.e., specify both a value of S and a value of W) with the highest frequency. A visual inspection is sufficient − you do not need to be incredibly precise. How close are these empirical values to the theoretical expected size and expected weight of a package, according to the given Bivariate Gaussian distribution?
2. For 1001 evenly-spaced values of W between 0 and 10, plot W versus the joint Bivariate Gaussian PDF $p(W, S)$ with S fixed at $S = 118$. Repeat this procedure for S fixed at $S = 122$. Comparing these two PDF plots, what can you say about the correlation of random variables S and W?
3. Give one reason for why the Gaussian distribution is an appropriate model for the size and weight of packages. Give one reason for why it may not be appropriate.
4. Because T is a linear combination of random variables, it itself is a random variable. Using properties of expectations and variance, please compute $\mathbb{E}(T)$ and $\mathrm{Var}(T)$ analytically.
5. Let us treat the *total* amount of time it takes to process *all* packages received at the USPS office within *an entire day* (assuming a single day is 24 hours long) as a random variable T^*.
   (a) Write a function to simulate draws from the distribution of T^*.
   (b) Using your function, empirically estimate the mean and standard deviation of T^* by generating 1000 samples from the distribution of T^*.

Solution

1b. From my histogram, I have found that the most likely combination of weight and size is at W $\approx$ 3.75 and $S \approx 120$. This values are extremely close to the theoretical expected size and weight of the package, with those values being W = 4 and S = 120.

2. Comparing the two PDF plots, one can see that since the graph of W remains virtually the same across different values of S, with only a horizontal shift, which suggests that W and S have a linear correlation, The graph illustrates this because when the value of S increases, the pdf of the weight retains the shape but moves slights to the right, meaning that with the increase in size, there is an equally likely increase in weight.

3. One reason who a Gaussian distribution is a fair representation of this is because size and weight of packages are not unlikely to naturally have a Gaussian distribution, and we can use the Gaussian distribution to model roughly how the packages should behave, if there is some deviation. It may be flawed on the other hand by way of how the world works, especially with box sizes not being all equally valid. There are some sizes of boxes that are much more produced and therefore more packages go out in those packages, so the graph for the distribution of size may not be as continuous as a Gaussian distribution may suggest, which is a fault in the abstraction.

4. To begin to find E[T] and Var[T], we can use the property of Expected Values to get that $E[T] = E[60] + 0.6E[W] + 0.2E[S] + E[\epsilon]$ we also know from the problem that $E[S] = 120$ and $E[W] = 4$, so we already have that $E[T] = 60 + 0.6(4) + 0.2(120) + E[\epsilon]$, so all that's left to do is find $E[\epsilon]$, and since ϵ is given to have mean 0, we know that $E[T] = 60 + 2.4 + 24 + 0 = 86.4$.
Similarly, we can use the property that $Var(T) = 0.6^2 Var[T] + 0.2^2 Var[T] + Var[\epsilon] + 2(0.6)(0.2)Cov[S, W]$. Now that we have this equation, we can use the fact that $Cov[S, W] = 1$, $Var[S] = Var[W] = 1.5$ and $Var[\epsilon] = 25$ to get that $Var[T] = 0.36(1.5) + (0.04)(1.5) + 25 + (.24)(1) = 25.84$
Overall, this gives us $E[T] = 86.4$ and $Var[T] = 25.84$