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

Problem 1

1. We write a system of linear equations using x_n, y_n and w_0, w_1 and applying the hint above.

$$y_1 = w_0 + w_1 x_1$$

$$y_2 = w_0 + w_1 x_2$$

We solve for w_0 in the first equation $w_0 = y_1 - w_1 x_1$. Then substituting this into equation 2,

$$y_2 = y_1 - w_1 x_1 + w_1 x_2$$

$$y_2 - y_1 = w_1(x_2 - x_1)$$

This clearly resembles the slope form, $w_1 = \frac{y_2 - y_1}{x_2 - x_1}$ Solving back for w_0,

$$y_1 = w_0 + \left(\frac{y_2 - y_1}{x_2 - x_1}\right) x_1$$

Unsimplified we get,

$$w_0 = y_1 - \left(\frac{y_2 - y_1}{x_2 - x_1}\right) x_1$$

Continuing,

$$w_0 = \frac{y_1(x_2 - x_1)}{x_2 - x_1} - \frac{x_1(y_2 - y_1)}{x_2 - x_1} = \frac{y_1 \cdot x_2 - y_1 \cdot x_1 - x_1 \cdot y_2 + y_1 \cdot x_1}{x_2 - x_1}$$

$$w_0 = \frac{y_1 \cdot x_2 - x_1 \cdot y_2}{x_2 - x_1}$$

2. We write the following such that such that $y = Xw$.

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}$$

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

3. There exists a unique solution for $w = (w_0 w_1)^T$ when X is invertible, but more specifically when $x_1 \neq x_2$ as that would lead to the determinant being 0.

4. We would use the equation $y = Xw$ to find w. First, knowing that there only exists a unique solution for w when X is invertible, we can use the invertible matrix theorem to find w.

$$y = Xw \rightarrow X^{-1}y = w$$

I know from math 22a that an invertible 2x2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \text{ has an inverse} \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

So we do multiplication as such, $X^{-1}y = \begin{bmatrix} x_2 & -x_1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$$w = \frac{1}{x_2 - x_1} \begin{bmatrix} x_2 & -x_1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{x_2 - x_1} \begin{bmatrix} x_2 \cdot y_1 - x_1 \cdot y_2 \\ y_2 - y_1 \end{bmatrix}$$

This agrees with my solution found in part 1. Yay!
Fitting the model in matrix notation allows us to take advantage of theorems of linear algebra that makes handling data points (which could be significantly large in size) much easier than it would in equation form. It is also easier to view as it is compact and easy to point out patterns that otherwise would be more difficult to view in equation form. More elaboration will

5. With hundreds of points there might not exist a line to connect the dots without curves or disobeying the rules of algebra. This leads to the usefulness of invertibility since when the matrix is not invertible since a non-square matrix is not invertible, then it is impossible to find an analytical expression for w.

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

Problem 2

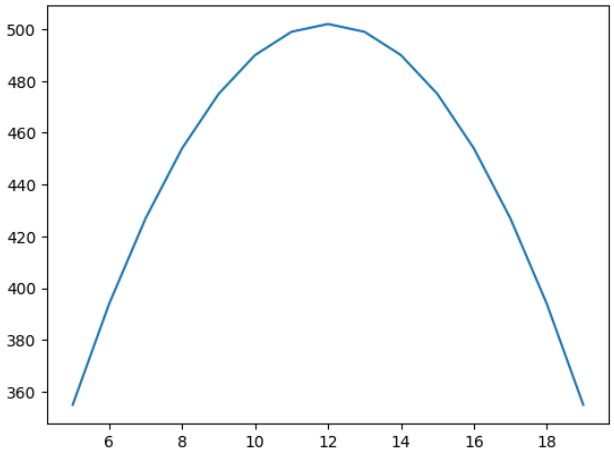

Figure 1: Quadratic

1. There only exists one unique solution as the function is concave with an absolute maximum. Since the formula is quadratic we know that there only exists one part of the function to be a minimum/maximum. Thus the answer is yes.

2. Using calculation we can use the first derivative test, setting $\frac{dy}{dx} = 0$.

$$0 = -6x + 72 \rightarrow 72 = 6x \rightarrow x = 12$$

So plugging in 12 into our function we get

$$y = -3(12)^2 + 72(12) + 70 = 502$$

Now we can check the left and right side of the function with only one critical point of $x = 12$. Since we know that when $x = 0$, or on the left of of the number line, $\frac{dy}{dx} = 72$ and therefore positive. Then on the right of the number line of 12 when x is for example 20, $\frac{dy}{dx}$ is negative. Since at our only critical point our function goes from increasing to decrease it shows that is a local and global maximum. Also using the second derivative test setting $\frac{d^2y}{dx^2}$ in which $\frac{d^2y}{dx^2} = -6$ for all values of x. This means the function is concave, which implies that the function's slope is decreasing for all values of x and there exists one unique maximum.

3. From the graph, the function appears concave and there is a peak which provides sufficient evidence why there is unique solution for Nari's optimization.

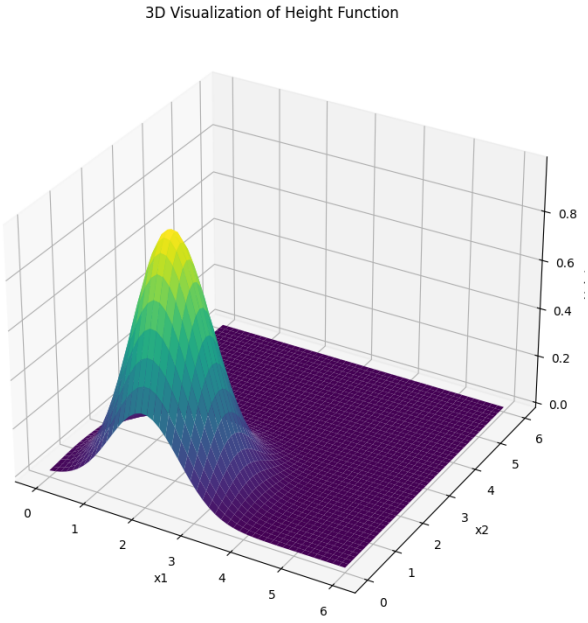

Figure 2: Nari's Optimization function

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 $\text{Var}(T)$ 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

Problem 3.

1. a) This is the histogram I came up with to answer the question

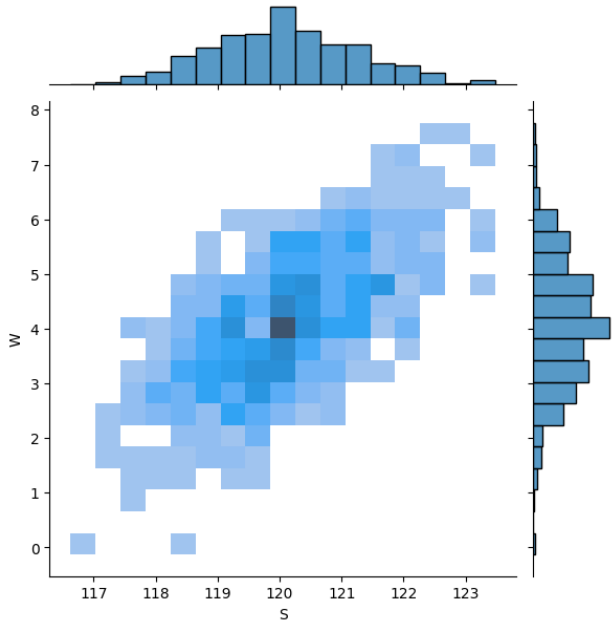

Figure 3: Heatmap for Bivariate Gaussian distribution for Packages

b) The empirical mean for size, S, is $120in^3$ and the mean for weight, W, in pounds is 4lb. This is quite literally the values given by the problem which is no surprise since the central limit theorem says that the sampling distribution of the mean will be normally distributed around the mean, so the value with the greatest frequency will be the mean, of course with large enough sample size (500 is big enough I guess).

2. S and W are positively correlated since at S=118 the value with the highest probability density for W is around 2.3, but it increases when S=122, which implies that W increases with S, thus a positive correlation between the random variables S and W.

3. A Gaussian distribution can be an appropriate model for the size and weight because it allows us to measure or approximate things whose distributions are unknown. In theory, the Gaussian distribution allows us to take account for error amongst different weights or sizes, with this and enough samples we can use the central limit theorem to approximate the mean weight of a package. However, this said a Gaussian distribution might not be appropriate because it allows for negative values. I was thinking that if these packages were a little bit smaller in scale, the Gaussian distribution might provide a negative value for weight or size, even though this is not possible in the real world.

4. $$E(T) = E(60 + 0.6W + 0.2S + \epsilon)$$

   By linearity,

   $$E(T) = E(60) + E(0.6W) + E(0.2S) + E(\epsilon)$$

   By definition of expectation

   $$E(T) = 60 + .6(4) + .2(120) + 0$$

$$E(T) = 86.4$$

$$\text{Var}(T) = \text{Var}(60 + .6W + .2S + \epsilon)$$

The constants cancel out to 0 in variance, and our error is independent of W and S so we can just pull it out. However we have to account for variance between W and S because they are correlated.

$$\text{Var}(T) = .36\text{Var}(W) + .04\text{Var}(S) + 2\text{Cov}(.6W, .2S) + 5$$

We can find the covariance through the cross (top right to bottom left) diagonal of the matrix.

$$\text{Var}(T) = .36(1.5) + .04(1.5) + 2 \cdot .36 \cdot .04 \cdot 1 + 5$$

$$\text{Var}(T) = .54 + .06 + .24 + 5$$

$$\text{Var}(T) = 5.84$$

5. For this problem look in jupyter notebook for solution

6. The empirical mean and standard deviation of T* found was 6229.20, 736.86 respectively.