CSCI 145 Problem Set 12

November 20, 2025

Submission Instructions

Please upload *your* work by **11:59pm Sunday November 23, 2025.**

- You are encouraged to discuss ideas and work with your classmates. However, you **must acknowledge** your collaborators at the top of each solution on which you collaborated with others and you **must write** your solutions independently.
- Your solutions to theory questions must be written legibly, or typeset in LaTeX or markdown. If you would like to use LaTeX, you can import the source of this document (available from the course webpage) to Overleaf.
- I recommend that you write your solutions to coding questions in a Jupyter notebook using Google Colab.
- You should submit your solutions as a **single PDF** via the assignment on Gradescope.

Grading: The point of the problem set is for *you* to learn. To this end, I hope to disincentivize the use of LLMs by **not** grading your work for correctness. Instead, you will grade your own work by comparing it to my solutions. This self-grade is due the **Monday** *after* the problem set is due, also on Gradescope.

Problem 1: Mean Estimation

Consider n real numbers $y_1, \ldots, y_n$. The mean is defined as

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{1}$$

In this problem, we will estimate the mean with a subset of m samples $S \subseteq \{1, \ldots, n\}$.

Suppose we have covariates $\mathbf{x}_i \in \mathbb{R}^d$ for each point i. It is expensive to compute y_i, but we know how to learn a function $f : \mathbb{R}^d \to \mathbb{R}$ so that $f(\mathbf{x}_i) \approx y_i$, from cheaper training data. Define the mean of the function so that

$$\bar{f} = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i). \tag{2}$$

Our estimate for the mean will be

$$\hat{y} = \bar{f} + \frac{1}{m} \sum_{i \in S} (y_i - f(\mathbf{x}_i)). \tag{3}$$

Note: For parts A and B, assume your samples are drawn *without* replacement.

Part A: Unbiased

Show that $\mathbb{E}[\hat{y}] = \bar{y}$.

Part B: Variance

Carefully derive an upper bound on $\mathrm{Var}(\hat{y})$.

Part C: Empirical Estimation

Load a regression dataset of your choice, and train a function f on the training set. Now, we'll compare several estimators for the mean:

- Monte Carlo with replacement: $\frac{1}{m} \sum_{i \in S} y_i$ where samples in S are drawn *with* replacement,
- Regression adjustment with replacement: $\bar{f} + \frac{1}{m} \sum_{i \in S} (y_i - f(\mathbf{x}_i)$ where samples in S are drawn *with* replacement,
- Monte Carlo without replacement: $\frac{1}{m} \sum_{i \in S} y_i$ where samples in S are drawn *without* replacement, and
- Monte Carlo without replacement: $\bar{f} + \frac{1}{m} \sum_{i \in S} (y_i - f(\mathbf{x}_i)$ where samples in S are drawn *without* replacement.

For variance budgets $m \leq n$, plot the MSE between each estimate and the mean. The horizontal access should be m, and the vertical access should be the MSE.

Problem 2: SHAP

Part A: Data, Training, SHAP

Using `shap`, load (a subset of) the California dataset. Using `sklearn`, train a linear regression and neural network model on the data. Using `shap`, apply an explainer of your choice to each model and the dataset.

Part B: Waterfall Plot

For the same observation in the dataset, make a waterfall plot with the Shapley values for both models. What do you notice?

Part C: Beeswarm Plot

Make a beeswarm plot with the Shapley values for both models. What do you notice?