Instructions

- Homework 1 is due November 30th at 16:00 Chicago Time.
  - We will not accept any submissions past 16:00:00, even if they are only one second late.
- You **must** upload the following files to the class Canvas:
  - `LASTNAME_FIRSTNAME.pdf`
  - `LASTNAME_FIRSTNAME.ipynb`
- Your code notebook **must** be runnable using my environment outlines in class 1 (Python 3.14, and the `requirements.txt`).
- You **must** use this template file and fill out your solutions for the written portion.
- Please note that your last name and first name should match what you appear on Canvas as.
- Include code snippets where required, as well as math and equations.
- Be *concise* where possible, all of the homework probelms can be answered in a few lines of math, code, and words.

Problem 1: One-Dimensional Data

Load in the data from the GitHub repository for this class.

```python
import pandas as pd

df = pd.read_csv(
    "https://raw.githubusercontent.com/tobiasdelpozo/data-analysis-2025/refs/heads/master/homework/homework_1/homework_1_data.csv"
)
```

Problem 1.1

For the feature labeled `X1`, compute the mean, median, variance, and standard deviation. Report your numbers below (rounded to at least 4 decimal places).

Answer:

Problem 1.2

Display a histogram of the feature `X1` using 50 bins. Do you think that the statistics you computed in 1.1 are good descriptors of the data? Include the graph below, and explain your reasoning in 1-2 sentences. [1]

Answer:

Problem 1.3

Using the same feature `X1`, come up with some metrics that are descriptive of the distribution of the data. Note, this is open-ended, so think about what the data looks like, and how a human would describe it.

Answer:

[1] Hint: you can use `\includegraphics{}` to include images in LaTeX.

Problem 2: kNN Regression

This problem uses the same dataset as Problem 1.

We're going to implement a k-Nearest Neighbors regression model. Unless otherwise specified, use an 80/20 train/test split for all parts of this problem.

Problem 2.1

Display a plot of `X2` versus the `target` variable. What do you notice about the relationship between these two variables?

Answer:

Problem 2.2

Implement a kNN regression model from scratch. You may use `numpy` and `pandas`, but you may not use any machine learning libraries (e.g. `scikit-learn`).

Your model should take in 4 parameters:

- `X_train`: training features
- `y_train`: training target variable
- `X_test`: testing features
- `k`: number of neighbors to use

And it should output the predicted values for `X_test`.

The algorithm you should use is as follows:

1. For each test point, compute the Euclidean distance to all training points.
2. Identify the k-nearest neighbors based on these distances.
3. Compute the predicted value as the mean of the target variable of these k-nearest neighbors.

Note that this we are only considering a single feature for this problem, so the Euclidean distance is simply $\sqrt{(x_{\text{test}} - x_{\text{train}})^2}$.

Include your code implementation below.[2]

Answer:

[2] Hint: you can use `\lstlisting[language=python]` to include Python code snippets.

Problem 2.3

Randomly split the data into training and testing sets (80/20 split), and report the Mean Squared Error (MSE) of your kNN regression model on the test set for $k = 5$.

Answer:

Problem 2.4

For $k \in \{1, 5, 10, 20, 50, 100\}$, compute the MSE on the test set and plot the results (k values on the x-axis, MSE on the y-axis). What value of k gives the best performance on the test test?

Answer:

Problem 2.5

Which value of k do you think has the highest bias? And which has the highest variance? Explain your reasoning in 1-2 sentences.

Answer:

Problem 3: Linear Regression

Using the same dataset as Problems 1 and 2, we are going to explore linear regression.

Problem 3.1

Using `statsmodels`, fit a linear regression model to predict `y` using `X3`. You should use **not** use an intercept term in your model. Report your β coefficient below:

Answer:

Problem 3.2

Re-run the linear regression model from 3.1, but this time include an intercept term. What are your new β coefficients (intercept and slope)?

Answer:

Problem 3.3

Do the following data transformations:

$$\tilde{y} = y - \bar{y} \qquad \tilde{X}_3 = X_3 - \bar{X}_3$$

Re-run the linear regression model using $\tilde{y}$ and $\tilde{X}_3$, without an intercept term. What is your β coefficient? How does it compare to your answer in 3.1?

Answer:

Problem 3.4

Inspect your data. Display a scatter plot of `X3` versus `target`. What do you notice about the relationship between these two variables? Is a linear model appropriate for this data? Explain your reasoning in 1-2 sentences.

Answer:

Problem 3.5

Define a new feature `X3_sin` as follows:

$$\texttt{X3\_sin} = \sin(\texttt{X3})$$

Fit a linear regression model to predict `target` using `X3_sin`, including an intercept term. Report your β coefficients (intercept and slope) below:

Answer:

Problem 3.6

Display a plot of `X3_sin` versus `target`. Do you think a linear model is appropriate for this data? Explain your reasoning in 1-2 sentences.

Answer: