HW#1 Due Date: 15/10/2024
MLR 570

Q1: Confusion Matrix 7 points

Part 1: Calculating Metrics for a Given Confusion Matrix

You are provided with the following confusion matrix for a binary classification problem:
- True Positives (TP): 50
- False Positives (FP): 10
- True Negatives (TN): 90
- False Negatives (FN): 20

1. Construct the confusion matrix.
2. Calculate the following metrics: Accuracy, Precision, Recall, F1 Score, Arithmetic Mean, and Geometric Mean.
3. Explain the significance of these metrics.

Part 2: Time-Varying Confusion Matrix and Exact Equality

Consider a scenario where the confusion matrix evolves over time. The values of TP, FP, TN, and FN are functions of time t:

$$TP: 100 + 5t$$
$$FP: 20 - 2t$$
$$TN: 200 + 3t$$
$$FN: 30 - t$$

1. Derive expressions for precision, recall, F1 score (harmonic mean), arithmetic mean, and geometric mean as functions of time t.
2. Find the time t where the three means are exactly equal.

Part 3: Approximate Equality of Means with Tolerance

You are tasked to determine when the three means are approximately equal within a specified tolerance (ε = 0.01), where t is limited to 0 to 20 (step size = 0.01).

1. Derive the condition for approximate equality within tolerance.
2. Find the time t when differences between the three means are all ≤ 0.01 in the range 0 ≤ t ≤ 20.

Q2: Pearson Correlation 3 points

You are analyzing the relationship between temperature and ice cream sales. This question involves calculating the Pearson correlation coefficient for a simple dataset, followed by exploring different types of noise.

Part 1:

Given the following data for temperature (in °C) and ice cream sales (in units):
- Temperature: [10, 15, 20, 25, 30]
- Sales: [100, 150, 200, 250, 300]

1. Calculate the Pearson correlation coefficient.

Part 2: Exploring Different Gaussian Noises

Using the same temperature dataset and sales:
- Temperature: [10, 15, 20, 25, 30]
- Sales: [100, 150, 200, 250, 300]

1. Add different Gaussian noise to create two noisy datasets.
2. Calculate Pearson correlation coefficient for both noisy datasets. Ensure the difference between the two coefficients is < 1e-3. Explain why this happens (solution is not scaling the variance of the added noise).

Q3: Windowed Aggregation 4 points

You are tasked with analyzing the stock prices of **Tesla (TSLA)** and **Ford (F)** over a 6-month period. The stock price data for both companies is retrieved using a financial data library. Your tasks involve calculating and plotting rolling means and medians for each company, followed by analyzing the correlation between the stock prices of Tesla and Ford.

```
import yfinance as yf

# Download stock data for Tesla (TSLA) and Ford (F)

tesla = yf.download('TSLA', period='6mo')

ford = yf.download('F', period='6mo')
```

Part 1: Calculation of Rolling Mean and Median

1. **Calculate and plot the rolling mean and rolling median** of the stock prices for each company using the following window sizes: 5-day, 10-day, and 20-day windows.
2. **Plot the stock prices of both companies** along with their respective rolling means and medians for each window size (5-day, 10-day, and 20-day). The x-axis should represent the trading days, and the y-axis should represent the stock prices.

Part 2: Pearson Correlation and Rolling Pearson Correlation

1. **Calculate the Pearson correlation coefficient** between the stock prices of Tesla and Ford.
2. **Compute and plot the rolling Pearson correlation coefficient** using a 10-day window to analyze the dynamic relationship between the stock prices of Tesla and Ford over time. Plot the rolling correlation as a function of time.