Programming for ML: Mid-semester examination

27th September, 2025

Rules

- No use of AI agents like ChatGPT/Gemini/Claude/Copilot, or inbuilt code agents in your editor.
- Basic tab completion on your editor and documentation lookup etc. are fine.
- You are allowed to use the class material, exercises, and exercise solutions.
- Open book/open notes/open internet.
- No help from other humans.

Questions

1. Consider the Alice in wonderland book (use the text file provided already). Using this data, perform the following tasks:

    a. Write code to read the entire novel, remove all punctuation, and replace multiple whitespaces with a single space. Then count the number of times each word (case insensitive) occurs in the text. Show a histogram of the top 15 words in the text. **Hint:** you can use a `collections.Counter` or use `pd.Series` to get the sorted values. (3 marks)

    b. Now split the book into the respective chapters. You will need to see the text file and think about how to do this. Once you have the text of each of the chapters, we will track the occurrence of 3 characters in the book namely, "Alice", "White Rabbit", and the "Hatter". Find out in which chapters these characters occur, i.e. for each character list the chapters that they occur in and also count the number of times they occur in each chapter. Rank the characters on the number of times they occur in total in the book. Plot these quantities suitably. (5 marks)

2. Recall that we copied the data from this table online to get all the ODI batting data for Virat Kohli during class. We also showed you how to clean this up and save this data to a CSV file. We have provided cleaned data for Virat Kohli and Rohit Sharma for this problem. Perform the following tasks:

a. Write a generic function called `get_batting_data(file_name)` which reads the clipboard and saves the data out to a CSV file using the code we have shared in the class. Use this function to save the data as a CSV file for any other Indian batsman of your choice. (2 marks)

b. Now given these three batsmen, compare their relative performance against the the top cricketing teams, i.e. India, Sri Lanka, Australia, England, South Africa, West Indies, and Pakistan (ignore the other countries). Specifically look at the number of innings they have played against each country, how many runs they have scored, the distribution of the runs, and at what average strike rate. For each of these, plot the data using an appropriate method. Provide a compact representation of the plots and not make too many plots (for example, use subplots or show multiple players together). Make some comments on each player's performance. (4 marks)

c. Write an interactive plotting function that is given a specific country (as a string), an option, which is one of "runs", or "SR" and one option which is either "box" or "violin". The function should compare the three batsmen as per the option provided. (4 marks)

d. Consider Virat Kohli and Rohit Sharma and find out which matches they both played in together. Plot their performance in these common matches as a box plot, and violin plot. Furthermore show a scatterplot of Kohli's score versus that of Rohit. Is there a correlation between their performance (compute the correlation coefficient for their scores in these common matches)? Compute their average strike-rates versus different countries and plot these. Comment on the results with respect to the two players. **Hints:** You can assume that if they played on the same date that they were playing together. Also see the `DataFrame.set_index` function and read the documentation carefully about what it returns and make note of the `inplace` parameter. (5 marks)

e. Let us imagine that Kohli is going to play for a 3 ODI series with Australia. Provide a 90% confidence interval of his average score in the 3 games (use the Australia specific data to do this and not all his runs). Do this without using a for loop or list-comprehension. (4 marks)

3. Consider the function $f(x, y) = sin(x) + cos(y)$. Plot filled contours of this function in the region $(x_{min}, y_{min}) = (0, -2\pi)$, $(x_{max}, y_{max}) = (4\pi, 2\pi)$ by sampling it with a uniform mesh of points. Visually inspect this plot and annotate one of the minima and maxima with text and an arrow. (3 marks)

Submit a single jupyter notebook with your answers in sequence and upload this on the GL interface.