Women's Institute of Technology and Innovation

Year 1, Cohort 4, End of Semester I Examinations
CSD 114 , Elements of Data Science & Machine Learning
Tuesday 10th December, 2024
Time Allowed: 72 hours from the start date and time

Rules & Instructions

1. **ATTEMPT THREE QUESTIONS IN ALL. Question *one* is compulsory**. You are free to attempt the bonus question for extra 10 points.

2. **Grading and feedback**: The total points for the questions add up to 90 points. The remaining 10 points are allocated to code style, commit frequency and messages, overall organization, spelling, grammar, etc. There is also an extra credit question that is worth 10 points (if you attempt it).

3. Your solutions must be written up in the Jupyter notebook file saved with your **registration number** only. This file must include your code and write up for each task. Your "submission" will be whatever is in your exam repository at the deadline. Commit and push the code file and the chart outputs of that file.

4. This exam is open book, open internet, closed other people. You may use any online or book based resource you would like, but you must include citations for any code that you use (directly or indirectly). You may not consult with anyone else about this exam other than the instructor myself for this course. You cannot ask direct questions on the internet, or consult with each other, not even for hypothetical questions. At most, all these questions could be completed by consulting all the resources and materials (hand outs and Jupyter notebooks shared during the course of the semester).

5. You have until [DUE DATE] to complete this exam and turn it in via your personal Github repo - late work will not be accepted. Technical difficulties are not an excuse for late work - do not wait until the last minute to knit / commit / push.

6. Each question requires a (brief) narrative as well as a (brief) description of your approach. You can use comments in your code, but do not extensively count on these. I should be able to suppress all the code in your document and still be able to read and make sense of your answers.

7. Even if the answer seems obvious from the .ipynb output, make sure to state it in your narrative as well. For example, if the question is asking what is 2 + 2, and you have the following in your document, you should additionally have a sentence that states "2 + 2 is 4." In short, I am interested in the process that produced the answer or result in your notebook.

8. **A note on sharing / reusing code**: I am well aware that a huge volume of code is available on the web to solve any number of problems. For this exam you are allowed to make use of any online resources (e.g. StackOverflow) but you must explicitly cite where you obtained any code you directly use (or use as inspiration). You are also not allowed to ask a question on an external forum, you can only use answers to questions that have already been answered. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism. All communication with classmates is explicitly forbidden.

1. **Variation and variability**: A random sample of 10,000 observational cases are taken from 2020 US Census data on income. The response variable: yearly income ($). Note that the data has been cleaned to remove people who were not employed and/or income was reported as 0. Thus, $n = 4,365$.

   (a) What is a 'typical' yearly income for these 4,365 people? (3 marks)

(b) Do the yearly incomes of these 4,365 people vary? How do you know? (2 marks)

(c) What would the values of the standard deviation and IQR be if there was **no variability** in yearly income for these 4,365 people? (5 marks)

(d) What would the values of the mean, median, Q1, Q3, minimum and maximum yearly incomes be if there was **no variability**? (5 marks)

(e) What do you think are some possible causes or reasons or potential sources of the variability in the yearly incomes of these 4,365 people? (Hint: Perform bivariate analysis leveraging boxplots between response variable and education level, region of the US. Provide an explanation of your thinking with numerical support) (5 marks)

2. **Exploratory Data Analysis:** The student performance dataset provides a detailed overview of student performance in various schools, focusing on academic achievements and demographic factors.

(a) For each gender, which major has the highest total study hours per week? (2 marks)

(b) Which gender has the highest overall average attendance rate? Return the gender and the average attendance rate? (3 marks)

(c) What percentage of students have part-time jobs, and what percentage have no part-time jobs? (5 marks)

(d) For each gender, what is the mean GPA and variance of age? (5 marks)

(e) Create a visualization that effectively shows if there is a relationship between any two variables of your choice. Your answer must be given in a single pipe. (5 marks)

3. **Linear regression** aims to predict a continuous target variable by finding the best-fit linear relationship between the target and the input features. The algorithm models the relationship using a linear equation: $y = \beta_0 + \beta_1 x + \epsilon$, where; y is the dependent variable, β's are the coefficients (weights), x's are the independent variables, ϵ is the error term. The goal is to find the coefficients (β's) that minimize the difference between the predicted values and the actual values. This is typically done by minimizing the Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - y)^2 \tag{1}$$

(a) Briefly describe the steps involved in fitting a linear regression in Python? (2 marks)

(b) What assumptions does the linear regression take into consideration? (3 marks)

(c) Using the 2020 US Census dataset, train a simple linear regression predicting a continuous target variable using the scikit-learn library. From the output coefficients, generate an equation representing the trained simple linear regression? (3 amrks)

(d) Compute the Mean Absolute Error (MSE) from your model. (2 marks)

(e) Extract and interprete the key outputs from a linear regression model (5 marks)

(f) Plot the linear regression model (5 marks)

4. **Bonus Question**.

(a) Multiply a 5x3 matrix by a 3x2 matrix (real matrix product) (Hint: *create a 5x3 and 3x2 matrices and peform matrix multiplication*).

(b) Create a 3x3 identity matrix

(c) Create a vector with values ranging from 10 to 49).

(d) Create a null vector of size 10.

(e) Create a 3x3x3 array with random values

(f) Create a 5x5 matrix with values 1,2,3,4 just below the diagonal

(g) Normalize a 5x5 random matrix

(h) How to get the dates of yesterday, today and tomorrow?

(i) Find indices of non-zero elements from [1,2,0,0,4,0]. In all the above questions explain your approach.