STAT40800 Data Programming with Python (online)

Final project

Due Sunday December 10th 2023 at 11:59 pm

Instructions

- Solutions must be submitted on Brightspace under *Assessments* → *Assignments* → *Final project*.
- Your submission must include your completed Jupyter notebook in **.ipynb and PDF format**.
- All of the results that you wish to include should be viewable without running the Python code. Note that the code may still be run by the grader to check that it functions correctly and as intended.
- Marks will be awarded for complete and correct answers to **all six** questions. An additional 10 marks will be reserved for organisation, presentation and conciseness.
- For full marks, you must justify your answers, clearly explain all steps and computations, label your figures, and write concise code.
- The project must be completed individually.
- To confirm that your have complied with the School of Mathematics and Statistics Honour Code, the following should be written and signed on the final page of your submission:
  "I confirm that all work submitted is my own and that I have neither given, sought, nor received aid in relation to this assignment."
- For this project you will analyse data derived from a study on performances of Portuguese students in maths exams[1]. Two datasets have been provided for this project:
  - `male_stud.csv`: contains information about the group of male students;
  - `female_stud.csv`: contains information about the group of female students.

  A full description of the values stored in the two *.csv files can be found in the text file "student.txt".

[1] Cortez,Paulo. (2014). Student Performance. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

Question 1 *10 marks*

(a) Load the `male_stud.csv` dataset into Python as a pandas DataFrame.

(b) Inspect the data. How many students are included in this dataset? How many different indicators are included? Does this dataset contain any missing values?

(c) Perform an exploratory data analysis, creating both numerical and graphical summaries of the data. Discuss and interpret your results.

Question 2 *10 marks*

(a) Load the `female_stud.csv` dataset into Python as a pandas DataFrame.

(b) Inspect the data. How many students are included in this dataset? Are the indicators the same as those in the male group?

(c) Perform a t-test, for each of the measurements, to test whether any of the indicators differ between the male and the female groups. Use a significance level of $\alpha = 0.01$. Display the t-score and p-value for each measurement. Clearly state the conclusion of your tests and explain your reasoning.

Question 3 *10 marks*

(a) Combine the two datasets into a single DataFrame.

(b) Compute the Pearson correlation coefficient between each of the measurements and identify which indicators are most correlated. List the four most strongly correlated pairs.

(c) Create scatter plots for the each of the correlated pairs identified in Q3(b). Are the relationships as expected from the correlation coefficients?

The full dataset should be used for all subsequent questions.

Question 4 *16 marks*

Logistic regression to predict students failure.

(a) In the Portuguese system a grade from 0 to 9.5 is considered as a FAIL, while grades from 10 to 20 are considered PASS. Create a new column in the dataframe indicating whether the student passed or failed. Use this column as dependent variable for the regression task in the following items *(Remaining indicators except for the final grade have to be used as predictor variables).*

(b) Separate the data into response and predictor variables and standardise the predictor variables.

(c) Fit a logistic regression model and interpret the fitted model.

(d) Perform forward selection for your regression model using the Akaike Information Criterion (AIC). Examine the selected model and discuss your findings in relation to the model fitted in part (b).

The original (without the additional column created for question 4) non-standardised dataset should be used for all subsequent questions.

Question 5 *20 marks*

Random forest regression to predict the final grade of a student.
(Remaining indicators to be used as predictor variables.)

(a) Split the data into appropriate training and test sets.

(b) Fit a random forest regression model with 10 trees using the training data. Include the argument `random_state=101` in the random forest regression function to ensure reproducible results. Determine which variables are most important in predicting the final grade of a student. Discuss your findings in relation to the logistic models fit in question 4.

(c) Use the random forest regression model to predict the final grade of a student for the test set. Create a scatter plot of the true final grade versus the predicted one. Interpret your plot.

(d) Assess the performance of a random forest regression model with 5, 10, 50, 100, 500, 1000, and 5000 trees in predicting the final grade of a student. You should repeat the model fit and prediction 20 times for each number of trees, using a different random state for each repeat. Create a plot of the model performance as a function of the number of trees (use a log axis for the number of trees). The plot should show the mean and standard error of the performance metric for each number of trees. Discuss your findings.

(e) Explain the rationale for fitting the model multiple times with different random states.

Question 6 *24 marks*

Clustering algorithms to identify different student groups

(a) Perform a k-means cluster analysis, using the indicators as the features. Run the clustering algorithm for different numbers of clusters (integers from 1 to 10). Plot the model performance as a function of the number of clusters and identify the optimal number of clusters for this data.

(b) Perform a k-means cluster analysis, using the optimal number of clusters (identified in part (a)), and identify the most discriminatory variables.
(*Hint*: Create histograms for each variable, with the data separated by cluster.)

(c) Create a series of scatter plots for the most discriminatory variables, colouring the points by cluster number. Discuss your findings. Do your findings support the claim that multiple categories of students, with distinctly different characteristics, are included in this dataset?

(d) Identify another clustering algorithm that may be suitable for this data. Give an overview of your chosen algorithm and discuss the type of problems it works best for. Repeat part (a)–(c) using your chosen algorithm. Discuss your results in relation to those from the k-means cluster analysis.
(See https://scikit-learn.org/stable/modules/clustering.html for an overview of other clustering algorithms.)

Organisation, presentation and conciseness *10 marks*