STAT40800 Data Programming with Python (online)

Final project

Due Sunday December 11th 2022 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 from the Allen Institute, a bioscience research institute located in Seattle. Three datasets have been provided for this project:
  - `neurons_group_1.csv`: contains information about the morphology of a group of brain cell, known as neurons.
  - `neurons_group_2.csv`: contains information about the morphology of a second group of neurons.
  - `neurons_additional_measurements.csv`: contains additional morphology measurements for the neurons included in `neurons_group_1.csv` and `neurons_group_2.csv`.

Question 1 *10 marks*

(a) Load the `neurons_group_1.csv` dataset into Python as a pandas DataFrame.

(b) Inspect the data. How many neurons are included in this dataset? How many different measurements 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 *8 marks*

(a) Load the `neurons_group_2.csv` dataset into Python as a pandas DataFrame.

(b) Inspect the data. How many neurons are included in this dataset? Are the measurements the same as those in `neurons_group_1.csv`?

(c) Perform a t-test, for each of the measurements, to test whether any of the neuron properties differ between the group 1 and group 2. 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 *12 marks*

(a) Load the `neurons_additional_measurements.csv` into Python and combine all three datasets into a single DataFrame.

(b) Comment on the dimensions of the combined dataset. Are all of the neurons from group 1 and 2 included in the dataset `neurons_additional_measurements.csv`?

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

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

The full dataset should be used for all subsequent questions.

Question 4 *16 marks*

Linear regression to predict the total surface area of a neuron (`total_surface`).
(Remaining morphological measurements to be used as predictor variables.)

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

(b) Fit a linear regression model and interpret the fitted model.

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

(d) Perform a forward selection Bayes Information Criterion (BIC) regression. Examine the selected model and discuss your findings in relation to the models fitted in part (b) and (c).

(e) Explain how using BIC for model selection differs from using AIC.

The non-standardised dataset should be used for all subsequent questions.

Question 5 *20 marks*

Random forest regression to predict the total surface area of a neuron (`total_surface`).
(Remaining morphological measurements 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 total surface area of a neuron. Discuss your findings in relation to the linear models fit in question 4.

(c) Use the random forest regression model to predict the total surface area of a neuron for the test set. Create a scatter plot of the true surface area of a neuron versus the predicted surface area. Interpret your plot.

(d) Assess the performance of a random forest regression model with 5, 10, 20, 50, 100, 200, 500 and 1000 trees in predicting the total surface area of a neuron. You should repeat the model fit and prediction 30 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 time with different random states.

Question 6 *24 marks*

Clustering algorithms to identify different neuron types

(a) Perform a k-means cluster analysis, using the morphological measurements 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 neurons, with distinctly different morphological properties, 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*