SCIF30003: Advanced Data Science for Scientific Computing

Exercise 3 – Assessment of Topics 2 and 3

This exercise will count for 15% of the marks for this unit.

It is intended to assess your technical skills in investigating scientific datasets which have a binary target (dependent) variable and numerical (independent) variables. It will assess the 6th and 7th intended learning outcome (ILO) for this unit:

6. Apply some of the more common learning and clustering algorithms used in machine learning

7. Describe and implement advanced data visualisation techniques for multi-dimensional data sets.

Instructions

In this exercise, we ask you to showcase the following skills: -

- Bayesian approaches
- Binary Logistical regression (BLR)
- Analysing data in a scientific context

Submission

Please submit all information needed to check and reproduce your work to the submission point on Blackboard. This would normally involve working code in a Jupyter notebook or a Python IDE, exhibiting best practice, but you could choose to use R or indeed Excel for all or some of your analyses. Provided all work is fully explained and can be repeated, you can choose the platform you find most appropriate.

Make sure that all questions are answered, that you showcase your skills, and that you fully justify the choices made. Marks will be awarded for the selection of approaches, clarity of analysis, quality of presentation and clear links with the scientific content, as well as the quality of code and analysis. The marking criteria are the same as those used for exercise 2 and are available on Blackboard.

Clearly indicate which option you have chosen and submit your work by the deadline of Wednesday at noon in week 20 (15th March 2023) on Blackboard. You can find information about extensions and extenuating circumstances on the assessment page for this unit, but please make sure you use the extension request form provided there, rather than the central route.

Data and context

Please choose one of the options for your work.

Option 1: Exoplanets

To date, over 5000 exoplanets (planets orbiting outside of our solar system) have been discovered within our Galaxy. The detection techniques used are heavily biased towards selecting planets which are large in size and orbiting close to their host stars with short orbital periods.

The attached file (exoplanet-dateset.csv) has a subset of data taken from the exoplanet.eu website which gives information on the units for the respective variables. A column has been added which indicates if the host star is an M-dwarf or not. M-dwarf stars have temperatures in the range of 2400 – 3700 K.

Part 1) Binary Logistical Regression (BLR)

Create and evaluate a BLR model where "star type" (i.e. M-dwarf or otherwise) is the target variable.

You should start by considering how many possible BLR models could be created (i.e. how many possible combinations of independent variables there are). By performing suitable exploratory data analysis on the dataset (*cf.* Topic 1) and considering the scientific context, select a suitable combination of independent variables (no more than 6).

You should evaluate the accuracy and limitations of your model.

Part 2) Bayesian distribution for star type

Let the variable θ denote the probability that a star is M-dwarf.

a. For the whole dataset, compute the Posterior distribution for θ using a suitable likelihood and prior. You should consider the order in which you insert data and track how the most likely value of θ and the width of the posterior distribution changes as you insert data.
b. (More challenging). Now investigate how the posterior distribution changes as a function of distance from the Earth. Create suitable plots involving θ and distance to gain an initial impression of any trends: you should think about suitable "binning" of the distance variable according to the range of values and its distribution. It's recommended to use a maximum of 10 bins.

   As for part a), you should consider the order in which you insert data. Present a table showing the final most likely value of θ and the width of the posterior distribution for each distance "bin".

It is known that ~70% of stars within the local neighbourhood within our Galaxy are M-dwarfs. How do your results compare? Consider the limitations of the dataset and the biases in detection of exoplanets.

Option 2: Wine Quality dataset

For this option, your task is to explore a dataset on the quality of wine, which has various chemical markers (wine-dataset.csv). The dataset is taken from the following web resource: https://archive.ics.uci.edu/ml/datasets/wine+quality.

An extra column has been added which indicates if the wine is "red" or white".

Part 1) Binary Logistical Regression (BLR)

Create and evaluate a BLR model where the wine colour (i.e. red or white) is the target variable.

You should start by considering how many possible BLR models could be created (i.e. how many possible combinations of independent variables there are). By performing suitable exploratory data analysis on the dataset (*cf.* Topic 1) and considering the scientific context, select a suitable combination of independent variables (no more than 6).

You should evaluate the accuracy and limitations of your model.

Part 2) Bayesian distribution for wine quality

Let the variable θ denote the probability that the wine quality is rated as "good". For the purposes of this exercise, "good" is defined as having a quality score which is 6 or higher.

a. For the whole dataset, compute the Posterior distribution for θ using a suitable likelihood and prior. You should consider the order in which you insert data and track how the most likely value of θ and the width of the posterior distribution changes as you insert data.
b. (More challenging). Now investigate how the posterior distribution changes as a function of "citric acid". Create suitable plots involving θ and citric acid to gain an initial impression of any trends: you should think about suitable "binning" of citric acid according to the range of values and its distribution. It's recommended to use a maximum of 10 bins.

   As for part a) you should consider the order in which you insert data. Present a table showing the final most likely value of θ and the width of the posterior distribution for each citric acid "bin".

What conclusions can you draw from your analysis? In answering this, you should consider whether the quality score threshold of 6 you used is sensible.