Assignment — 2021

Submission

The assignment solution should be submitted electronically, and consist of a report (as a PDF file) that you submit via Turnitin (by one member of the group), and the code that you create can stay in your GitHub repository.

Each member of the group is also required to submit on vUWS a text file listing contributions to the assignment by all members. Individual marks for the assignment will be calculated based on the contributions, in the case of a disagreement between the members, then the marks will be resolved on a case by case bases by the teaching team.

Tips: Your report should have 3 main sections: one section for each of the questions. Use a headline or section name to indicate which question you are answering. You can use subsections or lists to further organise each section.

Short answers are good, and you do not need to write long essays or repeat a lecture. Usually a few sentences to answer a question will be fine, but make sure you answer each of the questions in the text. If two questions require similar descriptions, it is OK to repeat yourself. Do use your own words – it is not OK to just repeat text from elsewhere, and we will check. Please write complete sentences, and please check your spelling.

Submission is due on 27 August 2021, 11:59am.

For this assignment, you will be working on wine quality data set. There are 11 objective tests (e.g. pH values) and one variable based on sensory data (median of at least three evaluation made by wine experts). The attributes of the data set are:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulphur dioxide
7. total sulphur dioxide
8. density

9. pH
10. suphates
11. alcohol
12. quality (scored between 0 and 10)

For attribute 12 (quality), which is rated by wine experts, each expert gave a score between 0 (very bad) and 10 (very excellent).

This assignment has three tasks.

This assignment is worth 10% of the unit assessment tasks.

Question 1

7 points

Create a linear regression model to predict the alcohol level of each wine. Use one part of the full data set as your training data to find the parameters of the model, and use the other part as test data to find the accuracy of the model. Use at least two different type of methods to find the parameters of your model.

Hint: You will not need to use the `quality` variable for this model. You might want to use the `train_test_split` function from `scikit-learn` to split your data set.

What to submit

- Write a brief description of your steps to create the model and your prediction. What did you do? Your description should include, but not limit to, answers to the following questions:
  - What is the accuracy of your model on the training data? What is the accuracy on the test data?
  - Is the model a good model? Why or why not?
  - Any particular choices you made or had to make in creating model or prediction? Why did you make them?
- Submit your Python code as part of your report (Jupyter Notebook) or if your code is in a separate `.py` file, identify it's GitHub location in your report.

Question 2

8 points

Create a logistic regression model to predict the `quality` of each wine sample. As with the previous question, split your data into training and test sets. You should create

models to at least perform one binary and one multi-class classification task on the data set.

Hint: How would you transform the values in `quality` into something you can use for classification?

What to submit

- As with Question 1, submit your Python code as part of your report or if you have separate Python code (`.py` files), write in your report what they are called and where to find them on GitHub.
- Write a brief description of your steps to create this second model, including but not limited to the following:
  - Include a description of what you did to preprocess the data.
  - What alternative options for preprocessing did you consider (if any)?
  - What is the accuracy (proportion of correct predictions) of your model on the training data and test data?

Question 3

5 points

Create a Ridge regression model to predict alcohol content (as in Question 1).

What to submit

- Submit your Python code as part of your report or if you have separate Python code (`.py` files), write in your report what they are called and where to find them on GitHub.
- Write a brief description of your steps to create your model, including but not limited to the following:
  - What is the accuracy of your model on the training and test data?
  - How did you tune the hyperparameter of your model?
  - Compare this model with the linear regression model in Question 1, did you achieve improvement in your result? What's the difference between the two models?