COMP40370 – Practical 5

Linear Regression and Classification

Prof. Tahar Kechadi

Academic year 2021-2022

Assignment Files

- ./practical05.pdf assignment questions (this file)
- ./specs/marks_question1.csv data file for Question1
- ./specs/borrower_question2.csv data file for Question 2

Question 1: Simple linear regression

The file ./specs/marks_question1.csv contains data about midterm and final exam grades for a group of students.

1. Plot the data using matplotlib. Do midterm and final exam seem to have a linear relationship? Discuss the data and their relationship in your report. Save your plot to ./output/marks.png.
2. Use linear regression to generate a model for the prediction of a students' final exam grade based on the students' midterm grade in the course, then describe the model in your report.
3. According to your model, what will be the final exam grade of a student who received an 86 on the midterm exam?

Question 2: Classification with Decision Tree

The file ./specs/borrower_question2.csv contains bank data about customers (borrowers) that may or may not have being defaulted.

1. Filter out the TID attribute, as it is not useful for decision making.
2. Using sklearn decision trees, generate a decision tree using information gain as splitting criterion, and a minimum impurity decrease of 0.5. Leave everything else to its default value. Plot the resulting decision tree, and discuss the classification results in your report. Save the produced tree into ./output/tree_high.png.

3. Train another tree, but this time use a minimum impurity decrease of 0.1. Plot the resulting decision tree, and compare the results with the previous model you trained. Save the produced tree into ./output/tree_low.png.
4. Discuss the generated models in your report.

Expected output and submission data

Your submission should be a single archive file (zip, tar, tgz, ...) containing one folder called output and the following files and directories:

- ./run.py main Python script
- ./report.pdf your PDF report (2 pages maximum)
- ./output/marks.png plot of data from Question 1
- ./output/tree_high.png plot of first decision tree from Question 2
- ./output/tree_low.png plot of second decision tree from Question 2
- ./specs/ the original specs folder included in the assignment archive, containing the input data

The final deadline for the submission is **Thursday, 28th of October**, 2021, at **17:00**. You can submit your solution on Brightspace.

Grading

The grading for the assignment will be assigned as follows:

- Question 1: **35%**
- Question 2: **35%**
- Report quality and content, code quality, submission format: **30%**

Programming requirements and tools

The assignment should be solved in Python, version 3.8 or above (3.9 is recommended). You shall use the following packages for this assignment:

- pandas 1.3+
- matplotlib 3.4+
- sklearn 0.24+ (earlier versions do not support plot_tree)

In particular, the following user guides are available for the required algorithms of the assignment:

- Linear regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear model.LinearRegression.html
- Decision tree: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- Plotting decision tree: https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot tree.html
- Pandas integration with matplotlib: https://pandas.pydata.org/pandas-docs/version/0.13/visualization.html