Tutorial 5

for

Data Mining - Simple Modelling

Version 1.0

Adam Ryan (14395076)

COMP47530

October 28, 2021

Contents

1 Question 1

1.1 Questions and answers

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.
    - We see that midterm and final appear to exhibit a linear relationship, with final increasing as midterm increases. We see there's a strong correlation of 0.78 between the two variables. We see that we've twelve records in each with the data skewed towards higher marks. The key summary stats are shown in figure 1.1 while the correlation matrix is shown in figure 1.2 and the seemingly linear relationship between the variables in figure 1.3.

```
------------------
The dataframe is described by:


         count       mean        std   min    25%   50%    75%   max
midterm   12.0  72.166667  17.698656  33.0  63.50  77.5  83.75  94.0
final     12.0  74.000000  13.149490  49.0  71.25  77.0  80.25  90.0
```

Figure 1.1: Dataframe Summary Statistics

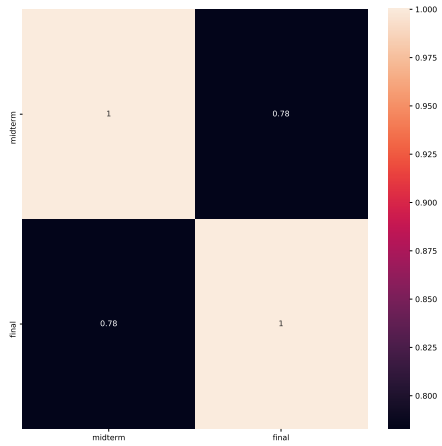

Figure 1.2: Correlation Matrix for Exam

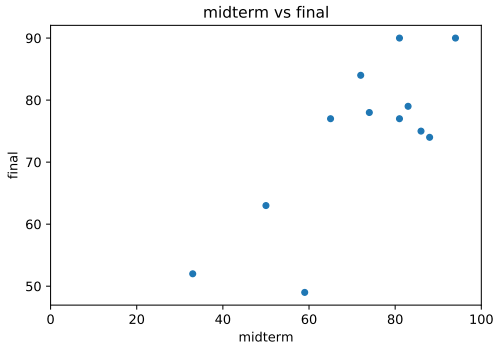

Figure 1.3: Midterm vs Final

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.

   a) I have generated a model using SKLearn. The model is:
   $final(midterm) = (midterm * 0.5816000773918931) + (32.02786108155171)$

   b) This model has an RMSE of 7.8 and an r^2 of 0.61. Over 5-folds, the average RMSE is 8.6 with a standard deviatio of the RMSE of 3.3.

   c) Ultimately the volume of data we have is pretty low.

   d) As per the lecture notes I've trained the model on the entire data set.

   e) The model can be visualised as in 1.4 and 1.5:

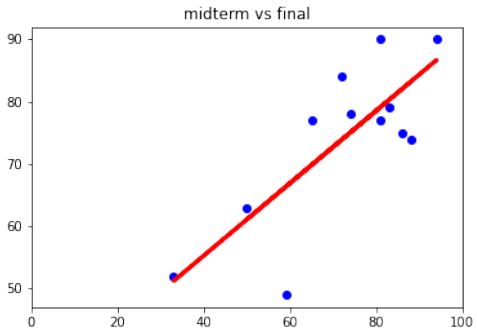

Figure 1.4: Regression Line on Data

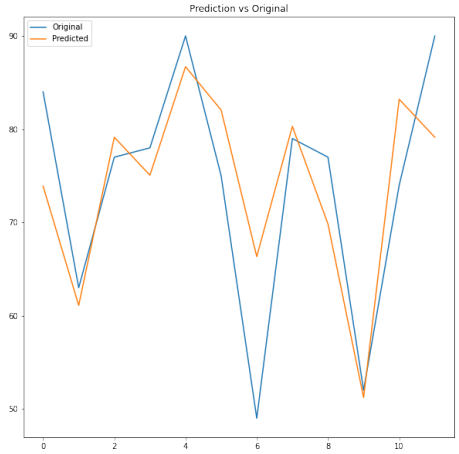

Figure 1.5: Predictions vs Originals per Test Case

3. According to your model, what will be the final exam grade of a student who received an 86 on the midterm exam?
    - My model predicts this student will receive 82 in their final.

2 Question 2

2.1 Questions and Answers

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

a) Filter out the TID attribute, as it is not useful for decision making.

- I use drop to remove the column.

b) 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.

- I pre-process the data to one hot encode the categorical variables as the DecisionTreeClassifier does not work with strings. I then join back onto the dataframe to bring in the income. Using the default values except for Random State, IG as the splitter, and impurity decrease, I create a decision tree. We see with the minimum impurity decrease as 0.5, the decision tree simply classifies everything using the majority class which, in this case, is that you won't default on borrowing. The generated model and tree is visible in 2.1 and 2.2

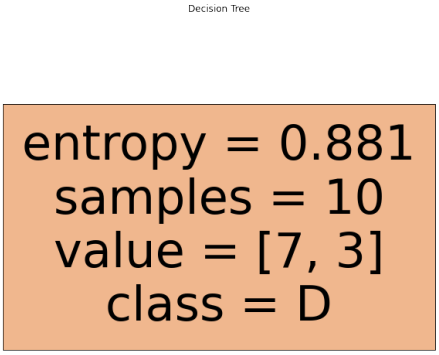

Figure 2.1: Decision Tree High

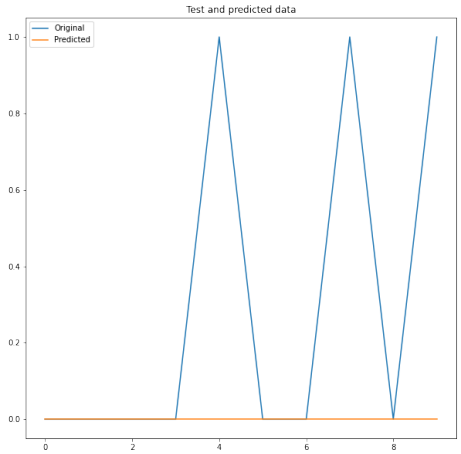

Figure 2.2: Predictions vs Originals High Tree

c) 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.

- I pre-process the data to one hot encode the categorical variables as the DecisionTreeClassifier does not work with strings. I then join back onto the dataframe to bring in the income. Using the default values except

for Random State, IG as the splitter, and impurity decrease, I create a decision tree. We see with the minimum impurity decrease as 0.1, the decision tree gains additional branches. The generated model and tree is visible in 2.3 and 2.4. Ultimately, it classifies not married people with an annual income greater than 77.5 but less than 122.5 into the 'defaulter' class. and everyone else into the 'non-defaulter' class.

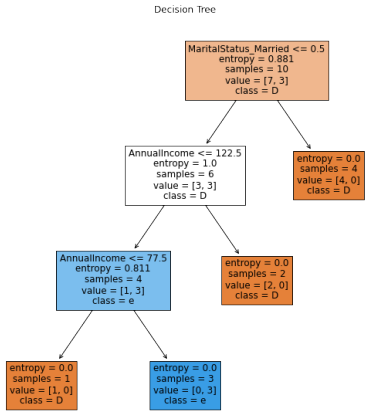

Figure 2.3: Decision Tree Low

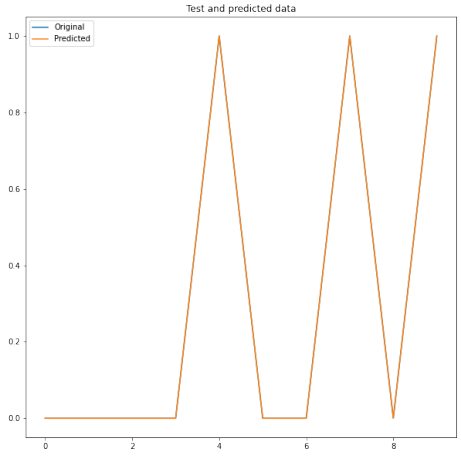

Figure 2.4: Predictions vs Originals High Tree

d) Discuss the generated models in your report.

- Please see discussion above. The low model classifies not married people with an annual income greater than 77.5 but less than 122.5 into the 'defaulter' class. and everyone else into the 'non-defaulter' class, while the high model classifies everyone into the non-defaulter class. Ultimately, the low model is superior, however we have a low volume of data.
- In both cases, the volume of data is far too low to make any proper comparison or insight into the model performance.

e) As a general note, throughout this assignment I have trained/tested/fit the model on the full dataset as the data volume is far too low for training or test splits. This is not a good practice and as such the performance claims need to be taken with caution.