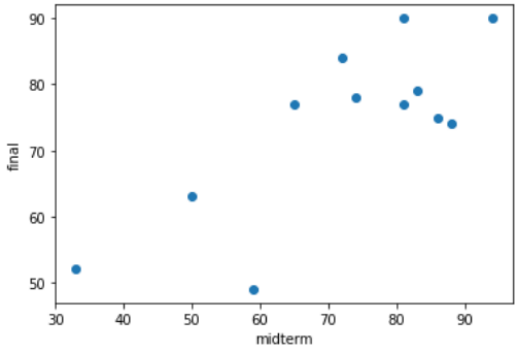
COMP40370 – Data Mining

Weronika Wolska

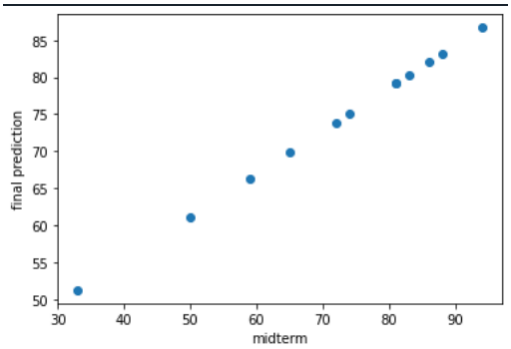
17301623

Question 1: Simple Linear Regression

1. The midterm and final results seem to have a linear relationship, with a few outliers:



2. In order to create a model, I have created two numpy.ndarray objects, a and b, one for storing the midterm results and one for storing the final results, as can be seen in lines 33 and 34 of run.py. The reshape() method was called on array a because it is required to be two-dimensional. The model itself was created using the default parameters of LinearRegression(), as can be seen in line 37. The model is fitted to a and b using the fit() function. This is the result of graphing the midterm results from the graph above with the expected final results generated by the model:



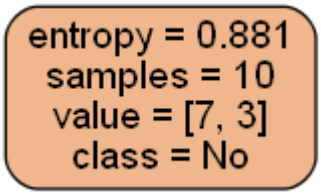
3. According to the model, if a student got 86% in the midterm, they will get 82.05% in their final.

Question 2: Classification with Decision Tree:

1. Column 'TID' was removed in lines 57-58 of run.py
2. The data set needs to be numerical, therefore I manually changed values in 'HomeOwner' from 'No' to 0 and from 'Yes' to 1. Likewise, in 'MaritalStatus', 'Single' was changed to 0, 'Married' was changed to 1, and 'Divorced' was changed to 2. In order to make the decision tree, I used the pydotplus module, which I had to download. The download instructions used are:
    1. Search "Anaconda Prompt (anaconda3)" in start menu
    2. Select "run as administrator"

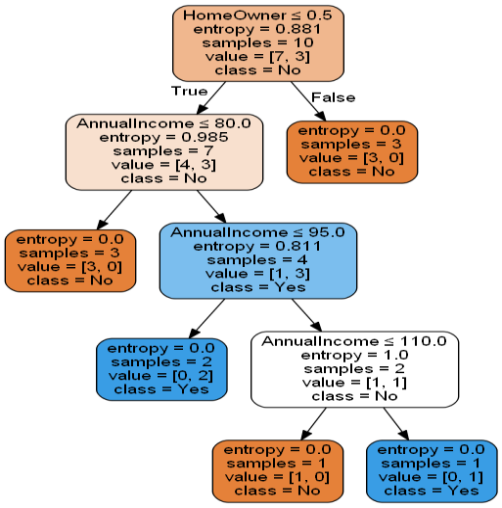
3. Type “conda install -c conda-forge pydotplus” and click enter

The decision tree produced in lines 60-85 is as follows:



So the predictions are made with 88.1% impurity, where 7 out of the 10 table entries are not a defaulted borrower, and 3 of them are. Therefore, the majority class is ‘No’.

3. The associated code for this question can be found in lines 87-95 of run.py. Pydotplus was used to create the decision tree, similarly to above. The decision tree produced is:



This tree is richer in information, as there is a lot more that we can tell from the data, for example: nobody that was not a homeowner was defaulted to be a borrower. Also, leaf nodes in the second tree have an entropy of 0.0, meaning that there is no impurity in the data, unlike in the first tree, where the level of impurity was extremely high. This implies that the lower the minimum impurity decrease, the better the results.

4. Completed in points 2 and 3 above.