Thomas Igoe
17372013

Data Mining - Comp40370

Q1
1 The data seems to have a rough linear relationship, as midterm trends up so does final, (with some notable outliers) indicating that a student with a good midterm score will likely also have a high final score.

2 Linear regression creates a line that best fits the dataset, and then uses that line to best estimate future data points, and what they correspond to. My program prints out the intercept and slope

3 My program outputs 82.04546774 as an estimated grade for a 86, this is done using the .predict() command, although you could easily calculate this with the y=mx+b formula
M - slope
B - intercept
X - 86

Q2
1 Self explanatory

2 The only major difficulty with the setup of this question was mapping all of the discrete values in the dataframe (HomeOwner, MaritialStatus etc) to numerical values, as that was necessary for the decision tree implementation.
The decision tree used 0.5 as the min_impurity_decrease and entropy as the criterion (AKA information gain as per the sklearn documentation) as per the question. The tree generated consists of only one node, as the higher min_impurity_decrese means that there are no attributes that could be selected to split the data that would result in an impurity decrease of 0.5 or greater.

3 Using similar options as the previous question, however the change to 0.1 min_impurity_decrease means that a more traditional tree can be generated, as the less strict impurity decreases threshold, disallows less of the possible splits, and as a result generates an actual tree, with multiple nodes.

4 The key difference between the two trees was obviously the min_impurity_decrease, and while this dataset might indicate that a lower min_impurity_decrease is always better, that is not the case, as an extremely large tree could be generated as we try to split on an even slightly impure dataset. Defining a minimum threshold keeps the software from making those unnecessarily small splits.