CSc 177 Project 2

Linear Regression Project and Classification Tree Homework

Alec Resha

A: Regression

For this project, I used a different dataset from the first project. The dataset is available at https://www.kaggle.com/datasets/sohier/calcofi?select=bottle.csv. For simple linear regression, I compared the temperature and salinity of the water. For multiple linear regression, I added the O2 saturation and depth of each record. I used MSE to evaluate the regression models. The simple linear regression had a lower MSE than the multiple linear regression, but only by about O.OO1 difference. Both MSE scores were 2.1 +/- 0.01. The images for the simple linear regression mapping are in the jupyter notebook file, the low MSE score is understandable as the data does not appear to be linearly correlated.

B: Admissions dataset Regression

The data appears to have a lot of relatively high chances of admission (very few between 0 and 0.4, most are between 0.4 and 0.7). For simple linear regression, I used the GRE score as the independent variale, and chance of admit as the dependent variable. MSE was used for evaluation, the multiple linear regression had an MSE around 0.0028 while the simple linear regression had an MSE of 0.0072. This shows that the multiple linear regression is better for this dataset than just the GRE score.

B: Classification Tree

I discretized the chance of admit into three classes, Low(0.0-0.4), Medium(0.4-0.7), and High(0.7-1.0). GRE Score, TOEFL Score, University Rating, SOP, LOR , CGPA, and Research were used as the features used. To make the tree smaller and more accurate, I switched the criterion to entropy, set the minimum number of samples per leaf to 5% of the total number of samples, and limited the number of leaf nodes to 20. This increased the accuracy from an average of 0.77 to 0.85. Rules:

- If CGPA > 8.63, it is almost guarenteed that there is a high chance of admission
- If CGPA <= 8.63, there is only medium chance of admission, regardless of other factors
- Research and/or a high University rating are the main way of getting a high chance of admission if CGPA <= 8.63

C: Entropy

- Solution in pdf file titles "entropy_solution.pdf".
    - Rules:
        - If Green, it is a -
        - if Blue, it is a +
        - Red or blue and square, it is a +
- Changes to entropy
    - If pattern of shirt was added, the entropy would likely remain higher until near the bottom of the tree, rather than many of the paths having an entropy of 0 and being removed.
    - If there is a high confidence/accuracy with a missing attribute, that attribute likely has little to no impact on the outcome.
    - It would make an impact on the manager or CEO if the accuracy was still high, since it would show that that feature actually doesn't matter, so it would be less important later on.

Group

This project was done alone, all of the work in this project was done by me (Alec).