Blake Erickson

TJ Durnford

October 24, 2016

Data Mining

Homework 4 Log

As with Homework 3, we started out by reading in the refined “auto-data.txt” file and creating a table from it. This time we also made a table from the “titanic.txt” file. We applied the Naïve Bayes classifier to each of these data sets to predict the values of a certain attribute.

For the auto data, we predicted MPG for randomly selected cars. For step 1 of this, we assumed that all values in the table were categorical. To reflect this, we had to adapt the weight value of each car to fit into one of 5 bins. We also fit the MPG into one of the ten rating bins created by the EPA. This resulted in our program outputting the predicted MPG as a value between 0 and 9. The predicted results and the actual results from each randomly selected row were then displayed. We also used stratified k-folds to asses our accuracy and printed a confusion matrix to compare our predictions to the actual values and asses our recognition percentage.

Once we had the program working well with the values assumed to be categorical, we created a new Naïve Bayes classifier that kept weight as a continuous attribute and used Gaussian distribution to assign it values. To allow us to re-use functions whose implementation would not change between steps 1 and 2, we made continuous Naïve Bayes child classes from the original ones that only changed the functions related to the weight. We repeated the testing process from the earlier step to compare our predictions against the actual values in the dataset.

The third step was to apply the Naïve Bayes and k-nearest neighbors classifiers to the Titanic dataset to predict whether a given passenger would survive. This set was trickier for k-nn because all of the values were nominal, so comparing distance was essentially reduced to a true/false comparison. Because of this, it was difficult to find the truly closest neighbors aside from how many attribute values they had in common. As before, we used random sampling and confusion matrices to assess the accuracy of our predictor and found that even with its limitations, k-nn managed to be around 75% to 80% accurate. Applying Naïve Bayes to the Titanic data also required some modification to our original Naïve Bayes class because there were some functions that were coded to deal with converting the continuous data into categorical data. We found that our Titanic Naïve Bayes classifier is roughly as accurate as our k-nn classifier and suspect that there may be some small issue with it that is making it less accurate than it possibly could be.