Assignment 1 Decision Trees and Naïve Bayes(75 points)

CSCE A415 and CS F480: Machine Learning

CSCE A490 Applications of PDC

Spring 2020

Due: 13 February 2020

Submission Instructions: Turn in both assignments in a zip or tar file through Blackboard. The zip or tar file should be of the form *last_name_only.zip* Each problem should be in a separate folder before zipping/tarring the two folders.

Problem 1 (25 points): Decision Trees

Using the ID3 algorithm, build a decision tree to predict the habitability of planets based on size and orbit (Martin Azizyan). The final result should be decision tree showing your calculations for each internal node, showing why you selected the node. If you draw the decision tree by hand, make sure your labels and calculations are readable.

Size	Orbi	Habitable	Count
big	near	yes	20
big	far	yes	170
small	near	yes	139
small	far	yes	45
big	near	no	130
big	far	no	30
small	near	no	11
small	far	no	255

Problem 2 (50 points): Implementing Naïve Bayes

For this problem you have two files, wine_test_set.csv and wine_train_set.csv. The goal is to develop a classifier that can distinguish between different qualities of wine from the set of wine features. You can see that the wine features are continuous values, whereas the quality of the

wine is a discrete rating.

When we apply the Naïve Bayes classification algorithm, there are two assumptions about the data: first, we assume that our data is drawn *iid* from a joint probability distribution over the possible feature vectors X and the corresponding class labels Y; second, we assume for each pair of features X_i and X_j with $i \neq j$ and that X_i is conditionally independent of X_j given the class label Y

a. (5 points) Briefly describe and visualize the data
b. (4 points) Determine the prior for each Y_k for the training data.
c. (4 points) Determine the probability of the evidence for each X_i for the training data
d. (4 points) Determine the probability of the likelihood of evidence for each X_i for the training data.
e. (4 points) Evaluate your model using your training data
f. (8 points) Determine the correlation of attributes with each other and assess whether you can fine tune your model by eliminating any attribute(s) to see if you can improve your model.
g. (11 points) Evaluate your model with your test data and discuss your results
h. (10 points) Select a smaller training set of 1000 records and a smaller set (100 records) of from the entire set of data. Complete the same steps (steps a through g) as before. Compare the results from the smaller training set with the larger training set. Describe how you selected the smaller set and explain why the results are similar or different.

Special Notes for problem 2: You may only use MS Excel, Matlab, or Octave(nearly identical to Matlab and it is open source) for this assignment. No other tools may be used. Please turn in your files showing your work in addition to your write up.