Assignment 2
273 Business Intelligence for Analytical Decisions

This assignment must be completed individually. Please submit assignment to Canvas dropbox including your name and student ID number.

1. You have seen how nominal and numeric attributes are specified in an ARFF file. Consider a feature called NUM_OF_DEPENDENT, which in the data file can take any value from the set {0,1,2,3}.

In the ARFF file, this feature could be specified as:

@attribute DEPENDENT Numeric

or as:

@attribute DEPENDENT {0,1,2,3}

What is the difference between the 2 specifications? Which specification would you use and why?

1 point

2. For this practice set, you should be working with the data set called **affiliation.** This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified as 16 different attributes in the data set. We are going to use a portion of this data set (marked for *training*) to train our classification models. The goal of the classification is simple: given the stands (votes) of an individual congressman, can we predict his/her party affiliation? As to be expected with any real-world dataset, there are several records with NULL values. However, we have taken a subset of the dataset with only those records that do not have a NULL value. The table training-no-NULL should be used to train (build) classification models, and the table testing-no-NULL should be used to test the results. The appropriate ARFF files are provided at EEE. You can use Excel for parts (a), (b) and (c). Use Weka for part (d).

(a) Prepare a frequency (count) chart for the data set and populate it based on the training data. See examples covered in class. Frequency chart shows cross-tab of class variables with each of the other variables. Use Excel Pivot tables for this – you may need several pivot tables.

(b) Prepare a populated probability chart (conditional probability) from the frequency chart in part (a). Again see examples covered in class.

(c) Based on the probability chart, apply naïve Bayesian classification to the test cases in the data set (5 rows in testing-no-NULL.ARFF). Compare your classification result with the actual data. What is the percentage of accurate classification from the classifier?

(d) Use WEKA to run the Naïve Bayes classifier on training-no-NULL.ARFF and set test file as testing-no-NULL.ARFF. Report the confusion matrix output by Weka.

7 points

3. After running a classifier in WEKA on some dataset, the following confusion matrix was obtained:

```
==========Confusion Matrix=========
a      b      ← classified as
921    28     | a=yes
17     374    | b=no
```

(a) Based on this confusion matrix, estimate the overall accuracy of the classifier.

(b) Estimate the stratified accuracies of the classifier.

2 points