CS910 Exercise Sheet 4: Classification

Please present your results as a single printed document. Limit your answers to 4 sides of paper – this should be more than enough to express your findings; any more and you are including a lot of unnecessary detail.

Breast Cancer Data

These exercises are best completed using Weka. For the first exercise, we use a medical data set that records incidence of Breast Cancer. This data set is included with the Weka software, and is stored in the 'data' subdirectory of the Weka program directory. It is called breast-cancer.arff. You can also download a copy from
`http://tunedit.org/repo/UCI/breast-cancer.arff`.

The first step is to look at the raw data. In the preprocess tab, there is a list of the ten attributes. The first nine give details about the patient and the treatment. The tenth, labeled "Class", indicates whether there was a recurrence of the cancer, or no recurrence.

1. We begin by studying the individual attributes in the data. [20]
   (a) What kind of attribute is the class distribution?
   (b) Determine the types of each of the other attributes: are they categoric, ordered, numeric, or binary?
   (c) What statistical distribution best fits the distribution of age? Estimate its parameters.
   (d) Which attributes appear to show the most dependence on the class value?

2. For the following questions, build classifiers based on a 66% training split. [30]
   (a) What is the accuracy percentage of the majority class classifier (Zeror)? What is the model given by this classifier?
   (b) How accurate is the NaiveBayes classifier "Bayes/NaiveBayes"?
   (c) Run the nearest neighbour classifier (lazy/IBk) with different values of kNN for the k nearest neighbours, from 1 to 10. (Click on the description of the classifier to change the parameters). Which value(s) gives the best accuracy on this data?
   (d) Run the nearest neighbour classifier with ManhattanDistance instead of EuclideanDistance (under nearestNeighbourSearchAlgorithm) for KNN from 1 to 10. How similar or different are the results, and why?
   (e) Run the J48 classifier with its default parameters. How accurate is the result? Describe the decision tree model that is found in words. Is the model overfitting the data?
   (f) Apply the SVM classifier, as "functions/SMO", and look at the model that is found by Weka. Which attribute is most heavily weighted in the model? Compare this to the results found using the decision tree classifier.

   Weka also keeps a list of results that you can scroll through to revisit the results of past experiments. For some class values, a false positive is more problematic than others. For example, predicting "no-recurrence-events" incorrectly can miss the potential recurrence of the disease.

3. Revisit your results. Which classifier has the lowest False Positive rate for the no-recurrence class? Which has the highest precision for this class? Discuss how acceptable these rates are for the application of predicting recurrence of breast cancer. [10]

Car Data

For the next exercise, we will use the Car Evaluation dataset from the UCI machine learning repository, from `http://archive.ics.uci.edu/ml/machine-learning-databases/car/`.

The file 'car.names' gives some background information about the data set. It describes the attributes, the number of values that they take on, the number of examples, and the overall distribution of the class value (how acceptable is the car).

To open this dataset in Weka, we have to put it into the "ARFF" format. To do this, create a copy of the car.data file, and call it 'car.arff'. Using a text editor, add the following lines to the start of the file:

```
@relation car
@attribute buying { vhigh, high, med, low }
@attribute maint { vhigh, high, med, low }
@attribute doors { 2, 3, 4, 5more }
@attribute persons { 2, 4, more }
@attribute lug_boot { small, med, big }
@attribute safety { low, med, high }
@attribute class { unacc, acc, good, vgood }
@data
```

The first line names the relation. The subsequent lines name the attributes in the order that they occur in the file, and give their possible values: buying takes on possible values vhigh, high, med and low. Then the @data line indicates that the data set follows. Save the file with these lines inserted, and open it in Weka.

4. We first try to predict a class with multiple values. [20]

   (a) Look over the number of occurrences of each attribute value in the preprocess tab in Weka. What do you notice about the number of examples for each value? Explain this behaviour.

   (b) Try out the different classifiers from above, and experiment with different parameter settings. Which classifier gives the best accuracy? Which parameter setting best captures the behaviour in the data? Comment on what you find.

5. Next, convert the data from having a multivalued class variable to a binary class variable: change all occurrences of good and v-good to acceptable ('acc'). Update the header information to change the description of the class to [20]

   ```
   @attribute class { unacc, acc }
   ```

   (a) Certain attribute values determine the class. Find which single attribute values automatically ensure that the vehicle is unacceptable.

   (b) Build a J48 decision tree on the data for the class value, and also an SVM model. Both should obtain high accuracy. Compare the models for the two models: which is more interpretable? Justify your answer with reference to the models found.