Assignment 1

CS36110 – Machine Learning

Jamie Hall – jah79

1a

Naïve Bayes

The main Bayes algorithm is as follows:

Where P is probability, 'A' is a class and 'b' is a feature…

$$P(A|b) = \frac{P(b|A) * P(A)}{P(b)}$$

The probability of class 'A' given that the selected feature is 'b' is retrieved from finding the probability of feature 'b' given that class 'A' is the selected class. Multiply this with the total probability that class A is selected and divide by the total probability of feature 'b'.

It is known as 'Naïve' as it assumes the occurrence of a specific feature is independent of others.

Advantages: Naïve Bayes algorithm is a simple but effective model. It also returns a degree of certainty which can be useful in certain situations.

J48 (Decision Tree)

A tree where nodes are features, links are values and leaf nodes are labels. Values lead to either another feature or a label. The tree should be as small as possible so it is quick and easy to reach a leaf node.

The tree follows a greedy heuristic algorithm. At the top of the top-down tree should be the most valuable feature which splits the set of data depending on their values. This repeats, using the next most valuable feature and continues until all sets of data from the 'training data' are classified.

Entropy is used to calculate homogeneity of a set of data and partitions this based on the values of said data. The entropy algorithm is used to find out information gain.

Advantages: Builds smallest and fastest tree for the data. Leaf nodes, when found can allow test data to be pruned, this reduces the number of tests required.

1b

In the set of data there are 500 passes and 268 fails.
When using Naïve Bayes classifier, the number of correctly classified instances was 587 (76.43%). The number of correctly classified instances is significantly higher with Bayes classifier over the J48 classifier which had 543 (70.70%). This means the Bayes classifier is more accurate than the J48 classifier.

I looked at the weighted average precision to get an idea of how precise the classifiers are. The J48 classifier produced a weighted average precision of 0.708, however the naïve Bayes classifier was noticeably more precise with a weighted average precision of 0.760. This further backs the point that the Bayes classifier is overall more accurate in comparison to the J48 decision tree classifier.

If I understand correctly, the relative absolute error is the distance of the difference between the target value and the approximation, this is then divided by the distance of the target value. for Naïve Bayes classifier this is 62.52% and for the J48 classifier it is 71.66%. Bayes is significantly less.

1c

Zero R classifier

Can be used to check accuracy by comparing the results of another classifier with that of the Zero R classifier regarding the same data. The Zero Rule classifier excludes any predictors involved in the process and focuses solely on the target feature. It looks for the most frequent value in that set of data, for example if 20 items of data are classified as Pass and 5 as Fail then it will just consider the data as Pass and look at it as True Positives and False Positives.

After executing the Zero R classifier, the precision results it returned were fairly low. The precision of 'pass' in this case was 0.651 and the weighted average for precision was 0.424. Any other classifier results that drop below this baseline are useless as using the Zero R classifier would be an improvement.

2a

When executing the following experiments, I assumed that the 0's were outliers or actually missing values. In some experiments these missing values and outliers may wrongly influence the results, in other cases it may be validly part of the set of data. To determine if it is an outlier we must look at the rest of the data in that feature. If the surrounding data is of low value then it is possible the 0's data is valid. If, however, the 0's data is surrounded by high values then that may indicate missing data.

To remove the zeroes from the data I attempted to use filters, more specifically the numeric cleaner filter however I was not sure what to change or add when editing the settings of the filter and so did not successfully implement this. Instead I manually edited the data. I would recommend using filters to edit data for this situation or similar situations as manually editing it was quite time consuming. After removing all the '0's from the data, I used the 'ReplaceMissingValues' filter to replace the missing values with the numerical data. The numerical data that replaced the missing values was based on the means and modes of the training data.

2b

I do not believe a method of replacing missing values is necessary in terms of the data involved with this assignment as the zeros included in the data are surrounded by low values which suggests that the data there is of value 0 and that this is not an anomaly but a valid part of the data. However, at the beginning of the visual data for InPlaneShearStrength there is a 34 value followed by a 0 value. Although this could be the value is actually 0, it's possible it is missing as the drop from 35 to 0 is significant. Even so, it may not be a large enough difference to act upon in comparison to other data value differences such as 36 to 136. There are two main ways to deal with this, either remove the 0's and leave it, or give them new values based on the training data as I have done as part of question 2.a. This can be done with filters, or as I have done for the previous question, can also be achieved manually.

For this assignment data set I would replace the 0's and assume they are pieces of missing data. There are many zero values in the HeatingCoolingFatigueTest feature, this suggests that these are not missing values. This does not need to be changed as the data is fine as it is, but I have replaced the 0's for this assignment so I can see the changes it makes to compare later with the first set of results. The ResinVsFibreConcentrationIndex feature only contains 1 zero value, this suggests that this is a missing value as the surrounding data is of higher value. This should be acted on by either completely removing the value or replacing it as previously mentioned.

3a

J48 (Decision Tree)
After modifying the data set (as mentioned in question 2), the J48 classifier now returns the following results. The number of correctly classified instances is 552 (71.87%), and incorrectly classified instances is 216 (28.12%). The Relative absolute error is 70.41% and Weighted average precision is 0.721.

Naïve Bayes
Using the same modified dataset, the Naïve Bayes classifier produced the following results. The number of correctly classified instances is 570 (74.21%) and incorrectly classified instances is 198 (25.78). Relative absolute error is 62.94% and the weighted average precision is 0.738.

We can use the Zero R baseline classifier mentioned in question 1.c to check the results against and as both classifiers are producing better results than that of the Zero R classifier it shows that these classifiers are still efficient and useful.

3b

Comparing naïve Bayes results
With the initial set of data, the naïve Bayes classifier had 76.43% of correctly classified instances however with the modified data set this parameter dropped to 74.21%. The percentage dropped slightly as there were no 0's in the data at this stage and they were replaced with various other values based on the training data, this meant that there was a larger variety of values to check and some incorrect instances may have been added.
The weighted average precision for the initial set of data was 0.760, with the modified data was lowered slightly to 0.738. As there was more variance in the modified data set it might not be as accurate.

Comparing J48 results
In the results for the initial set of data, the percentage of correctly classified instances was at 70.70% and this was raised to 71.87% with the modified data set. This surprised me slightly as I thought it would have decreased slightly with the addition of more values. This could be due to a change in the amount of values split at one point, as some may have moved to another branch due to their modified value. As for the weighted average precision, with the initial data it was at 0.708 and with

the modified data was at 0.721. Again an increase, this could be due to the same reason as the increase in correctly classified instances.

Summary

To summarize this report, I have briefly explained the basic idea of the Naïve Bayes and J48 classifiers. Compared the results produced by both classifiers and came to the conclusion that Bayes classifier is better in many aspects especially accuracy and precision. I discussed how the Zero R classifier can be used as a suitable baseline to check if a classifier is useful and also discussed what possible options you have when dealing with missing data. In this case I dived further into replacing the values instead of just removing them from the data.