Tutorial Notes

1. 38,160 instances (rows)
2. 20 attributes
3. 1 class/target feature (column contam\_code)
4. obscuredA.arff and obscuredB.arff are identical except the final column has been randomly shuffled
5. Confusion matrix can be used to see whether your model is biased toward certain classes.

How can you tell there is a signal?

1. Train a model (e.g. decision tree, neural net etc) that is more accurate than just looking at the majority class
2. Look at the confusion matrix and seeing if the model just classifies everything under the majority class
3. If you build a decision tree and find that the attributes are being used

Techniques:

1. Majority class as a baseline
2. Use 10-fold cross validation
3. Confusion matrix comparison

**Results**

Confusion Matrix Comparison

* Decision trees were fitted on a random selection of 80% of the data from each set (training set)
* The distribution of test data from each set is overlaid to show that the distributions were similar thus comparisons are possible.
* The confusion matrices are shown below:

The confusion matrix from Obscured set B shows that the decision tree only predicts the majority class correctly. In comparison the confusion matrix from Obscured set A shows that the decision tree correctly predicted more than just the majority class.

* Accuracy and Cross validation comparison

As a baseline, a dummy classifier was created that only predicted the majority class. The accuracy of this classifier was around 88%.

The decision tree fitted on both obscure sets A and B had 10-fold cross validation applied. The accuracy of the model was calculated at the end of each iteration. The results are as shown below:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cross validation iterations | | | | | | | | | | |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Dataset | Obscure A | 0.920 | 0.92 | 0.921 | 0.92 | 0.919 | 0.923 | 0.920 | 0.922 | 0.917 | 0.92 |
| Obscure B | 0.886 | 0.885 | 0.887 | 0.886 | 0.886 | 0.884 | 0.885 | 0.885 | 0.885 | 0.88 |

The accuracy of the decision tree fitted to obscure B stays around 88% which is similar to the accuracy of only predicting the majority correctly. Examining the confusion matrix above, it was observed the decision tree in that iteration predicted only the majority class correctly. It can be postulated that the decision tree fitted to obscure B is only predicting the majority class correctly.

In comparison the accuracy of the decision tree fitted to obscure A is consistently 92%. This is higher than the decision tree fitted to obscure B and the dummy classifier.

Using the confusion matrix and the accuracy evidence I am concluding that Obscure set A is the original dataset and Obsure set B has been shuffled.