Lab

Decision tress and SVMs

Part C

The performance (accuracy rate in %) of each classifier using 10-fold cross validation to observe the accuracy of modes is as following :

	Soybeans	Labor
ZeroR	13.47	64.91
OneR	39.97	71.93
J4.8	91.51	73.68

There is a large difference in results.

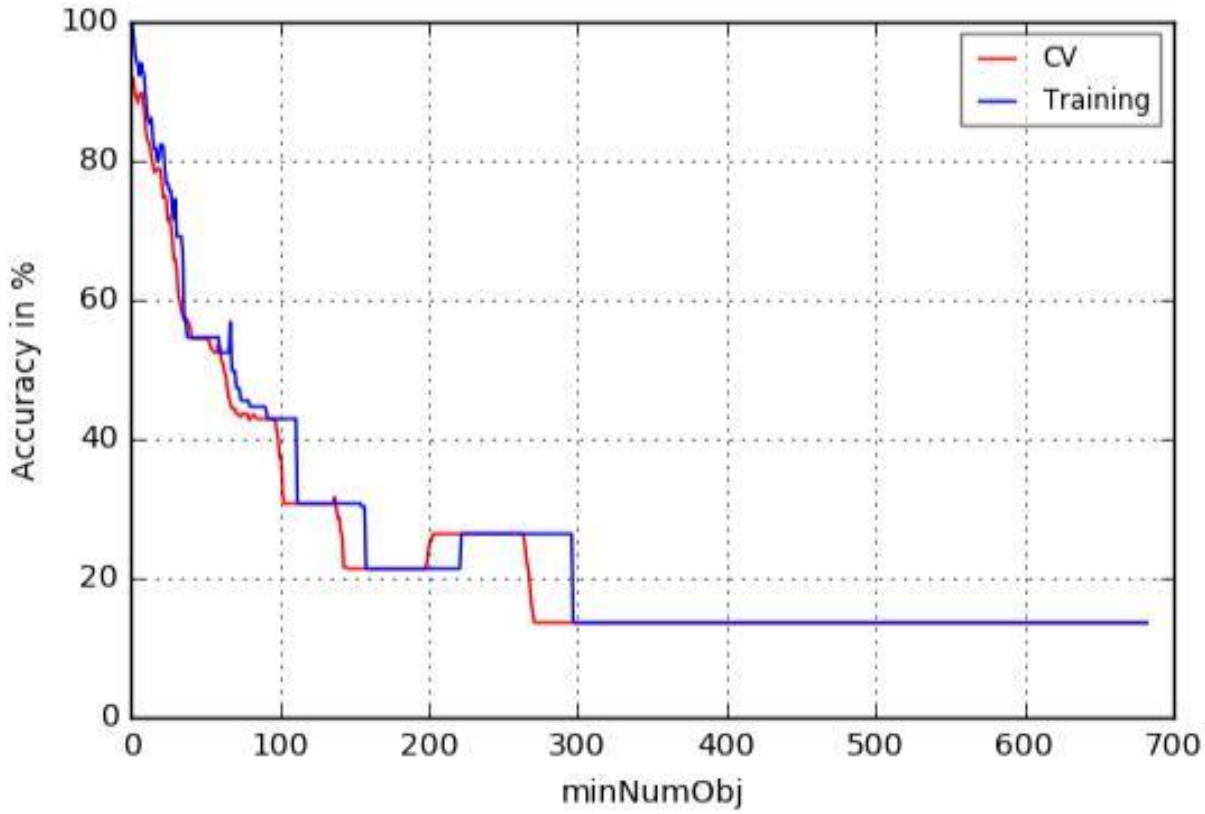
The zeroR classifier performed by far the worst. When classifying the Soybeans dataset it's performance is terrible partly because of the size of dataset and the number of attributes. The zeroR classifier performed much better for the labor dataset but labor dataset consists only of 57 instances with 17 attributes. Such a dataset is by default much more homogeneous than the Soybeans dataset and we can explain the better performance by that.

The oneR classifier which is a single level decision tree should be mainly compared to J4.8, a full decision tree learner. In the Soybeans dataset which has 36 attributes the performance of OneR is much lower than performance of J4.8. This is tells us that not just one attribute is important when classifying soybeans. In the Labor example we can see that performance of OneR is very close to performance of J4.8. We can blame the low number of attributes (just 17) and also a low number of instance (57) for J4.8 not vastly outperforming the OneR classifier.

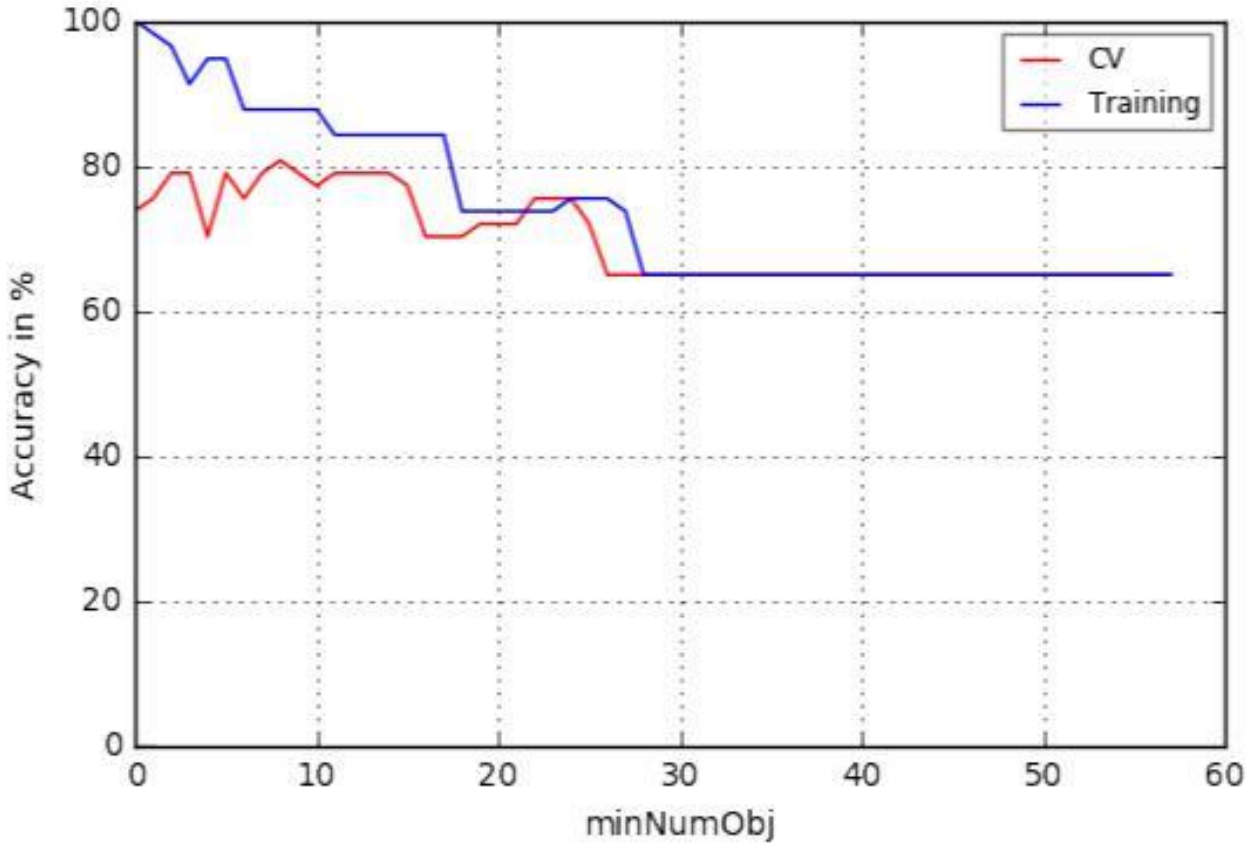
Part D

The graphs below contain on X-axis the hyper parameter - minNumObj which represents the minimal number of objects in a leaf node. It is used as a parameter in pre pruning strategy. The strategy obviously **overfit with low numbers of minNumObj (0-2)** as the training set classification was almost 100% correct and the corresponding rules of decision tree were really complex. Also the strategy strongly **underfits the model with larger minNumObj (n>15)** as the leaf size are too large for any complex decision tree to be used. That effect is especially visible with the soybeans dataset which has many attributes and because of complexity was badly treated with large minimal numbers of objects in a leaf node. Both datasets had the **ideal values in ranges from 3-15 minimal objects per leaf** node. Compared to part C we can see some improvement in the accuracy, especially with the labour data set but not a large difference.

Soyebean dataset



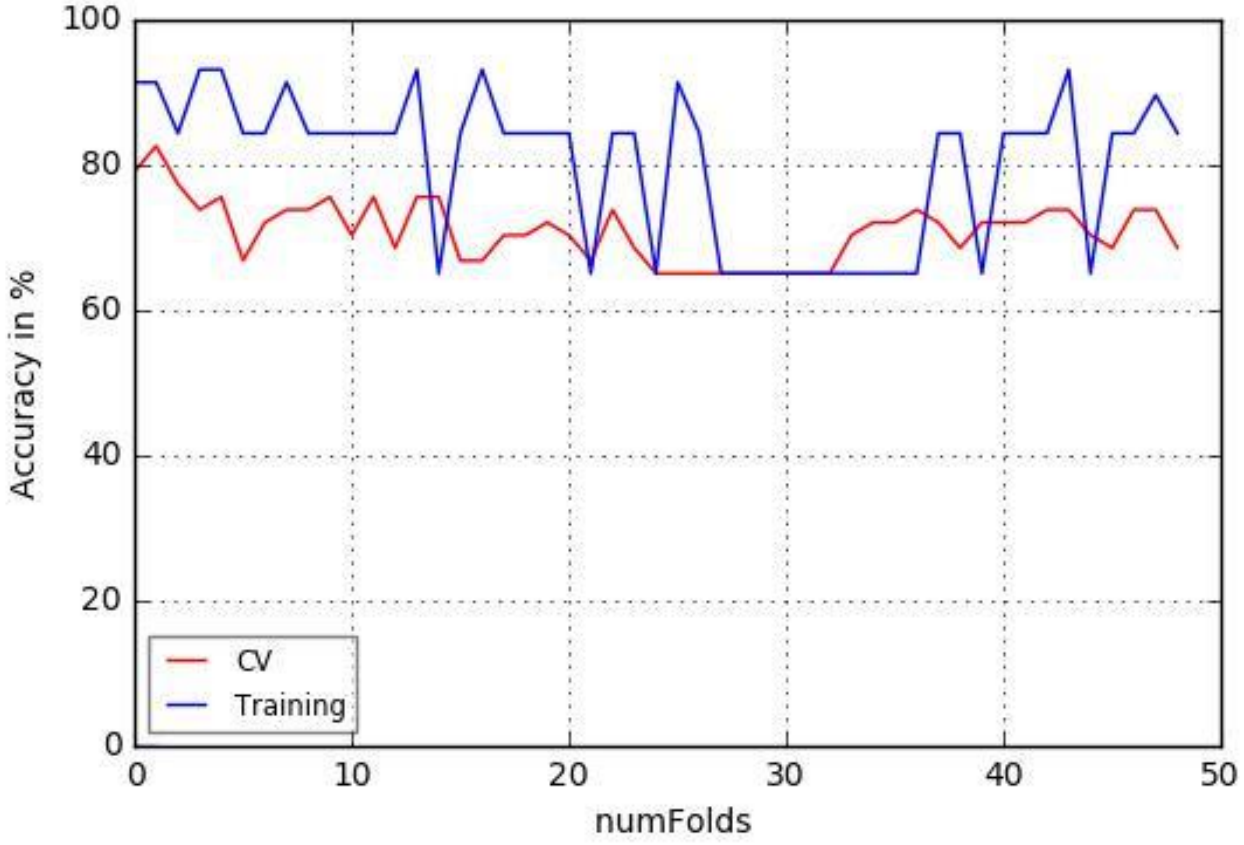
Labor dataset



Part E

The graphs below contain on Y-axis the accuracy of cross validation in % and on the X-axis the hyper parameter - numFolds which determines the amount of data used for reduced-error pruning. This is another pruning strategy. The strategy did overfit the data when low numbers of numFolds were used but it did not overfit as much as the previous strategy. The training set classification was almost 100% correct but also the CV accuracy was high. Also the strategy slightly underfits the model with larger numFolds values but the discrepancies are much smaller compared to the other pruning strategy. Both datasets had the ideal values in ranges from 2-10 minimal objects per leaf node. Compared to part C we can see some improvement in the accuracy. Compared to part D we can see large improvements in overfitting and underfitting.

Labor dataset



Soybean dataset