# ML Assignment 3

*Author : Aman Bhatia*
*Dated : 12 April 2016*

## Decision Trees for Classification

**(a)** With increasing number of nodes, the training accuracy increases which is the expected behaviour. But on validation and test data, in the last the decrease in accuracies are observed. The reason obviously being overfitting.

***Accuracies***

| | |
|---|---|
| Train data | : 89.34 |
| Validation data | : 84.8 |
| Test data | : 73.51 |

**(b)** Pruning tries to remove the overfitting of the data. Hence, obviously the train data accuracies decreases, but the accuracies of the validation data and test data increases. Infact a significant increase in the validation data accuracy is observed. Following are the accuracies after pruning,

***Accuracies***

| | |
|---|---|
| Train data | : 87.15 |
| Validation data | : 89.6 |
| Test data | : 76.12 |

**(c)** Since, there was no threshold on the number of samples to split, the tree went on spliting till depth 15. As a result, high overfitting of the data is observed. The training accuracy rises upto 100% where as validation data accuracy decreased significantly. Following are the accuracies observed,

| | |
|---|---|
| Train data | : 100.0 |
| Validation data | : 71.2 |
| Test data | : 74.6268656716 |

Model (c) is better that model (a) because we are not changing the actual information given to us. We are building the tree on the data which is more informative than that in model (a). However, model(c) is capable of overfitting to a great extent. Hence, the tree we get ***after pruning*** from model(c) will give us better reuslts as compared to model(a).

**(d)** I plotted various accuracies with the *max_depth* parameter and observed that in all the cases the validation accuracy is decreasing from ~85% and again getting a local maxima at near *max_depth*=4. Similarly, for *max_depth*=4, I plotted all the accuracies with *min_samples_split* parameter and found that nice results in terms of accuracies are observed at *min_samples_split*=6 or 7.

Hence, for maximum validation accuracy, *max_depth* needs to be 1. But, nice results are obtained at ***max_depth=4*** and ***max_samples_split=7***.

## Newsgroup Classification

**(a)** Following accuracies are obcserved on test data,

Iter 1 : Accuracy = 95.36652835408022
Iter 2 : Accuracy = 94.95159059474412
Iter 3 : Accuracy = 94.53665283540802
Iter 4 : Accuracy = 95.50484094052558
Iter 5 : Accuracy = 94.19087136929461

**(b)** On randomly guessing, we are getting an accuracy of 12.5% which is expected as there are 8 classes and probability of guessing the correct class is 1/8=0.125 (which is 12.5%).

**(c)** Cross-posting is a problem for us. Suppose we cross posted all the data. In that case, there will be nothing to learn and accuracy will come down to

randomly guessing accuracy that is 12.5 %. Hence there should be minimum cross posting.

**(d)** The accuracies monotonically increases for the train data and monotonically decreases for the test data. This is the expected behavious.

When we have seen little, and someone asks the same little things which we have seen, than we will answer correctly, where as when asked things which we have not seen, we are not going to answer them correctly. That is why the accuracies in the beginning are high for train data and less for test data.

In contrast, when we have seen all the things there are, then we may answer known things wrong because of so much data. And obviously, as our knowledge is increased now, we will answer things which we have not seen more correctly. This is what is happening in the end of the learning curve.

**(e)** Following Confusion Matrix is obtained,

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rec.sport.hockey | : | 975. | 0. | 3. | 2. | 3. | 9. | 0. | 7 |
| talk.religion.misc | : | 0. | 546. | 6. | 38. | 3. | 0. | 9. | 26 |
| rec.motorcycles | : | 1. | 0. | 965. | 3. | 22. | 2. | 0. | 3 |
| talk.politics.guns | : | 0. | 2. | 2. | 871. | 0. | 1. | 1. | 32 |
| rec.autos | : | 0. | 0. | 15. | 13. | 948. | 4. | 0. | 9 |
| rec.sport.baseball | : | 16. | 1. | 3. | 4. | 2. | 966. | 0. | 2 |
| talk.politics.mideast | : | 2. | 3. | 3. | 5. | 1. | 1. | 907. | 18 |
| talk.politics.misc | : | 1. | 8. | 1. | 61. | 2. | 2. | 12. | 688 |

As can be seen, newsgroup with highest diagonal entry : ***rec.sport.hockey*** and most confused new groups : ***talk.politics.misc*** and ***talk.politics.guns***

On normalizing the data, the same results are seen.

***-------- END OF DOCUMENT---------***