# Optical Digital Recognition Data set:

In this report, ensembles method will be used to study the accuracy using different combination of trees. The analysis will try to understand the effect of two major components:

1. The effect of how good/bad the underlying method
2. The number of ensembles and the effect of accuracy.

# Feature Engineering

In this dataset, there will be no feature engineering. As already known from previous analysis, the classes are well spread across the dataset and there are no outliers or noise in the data.

# Analysis

The ensembles used in this report is a set of decision trees. To recall, based on previous assignment in decision trees the results were as follows:

|  |  |
| --- | --- |
| Tree Depth | 11 levels |
| Splitting Criteria | Entropy |
| Execution Time | 0.06 seconds |
| Accuracy on the Training Set | 99% |
| Accuracy on the Test Set | 87% |
| Cross Validation Mean | 89% |

Table : Previous Results for Decision Tree Classifier

To understand the behavior of the previous two points. Different combination of tree depth and number of estimators tested as follows:

|  |  |
| --- | --- |
| Tree Depth | Number of estimators |
| 1,2,3,…,11 | 50,150,…,450,500 |

Table : Tested combination

## Effect of underlying ensemble complexity

Ensemble method suggest the underlying model to do better than a chance (>50%). While 51% is higher than 50% and 100% is also higher than chance. Overfitting might appear in case of very complicated over fitted underlying model that achieve almost 100% accuracy on the test set. And this is what this report will start to illustrate. In other words, the effect of a complicated underlying models.

The figure below depicts the behavior by increasing the tree depth over different number of estimators.

Figure : Accuracy vs Tree Depth

Regardless the number of estimators, the effect of tree depth shows similar behavior and trend over the accuracy. It also obvious that ensemble achieve better performance than individual stump. In addition to that, as the number of tree depth increases i.e trees become more complicated, the difference between ensemble approach and the stump model is almost the same. In other words, the more the underlying ensemble becomes more complicated the same model will be returned every time. and this explains the behavior after tree depth of 10 that there is no difference between the stump’s accuracy and the ensemble.

## Number of Estimators

In this experiment, the accuracy over different number of estimators studied while fixing the tree depth. In order to understand the behavior of number of estimators in the given data set.

Figure : Accuracy over different number of estimators

As shown above, the number of estimators in this case has no effect on the accuracy as it was the case of tree depth. The meaning behind that, the ensemble converges in a very early stage even before 50 estimators used and this explain the static accuracy value over different tree depth while increasing the accuracy.

Based on the previous analysis and the overall results the selected model will include an underlying tree with depth 6 which will result in 82% accuracy on training set and 75% on test set. The number of estimators are selected to 50 which will result in ensemble with accuracy of 100% on training set and 94% on test set.

# Results

|  |  |
| --- | --- |
| Number of Estimators | 50 |
| Stump Tree Depth | 6 |
| Stump Accuracy Training set | 82% |
| Stump Accuracy Testing set | 75% |
| Ensemble Accuracy Training set | 100% |
| Ensemble Accuracy Testing set | 94% |
| Training Time | 1.5 seconds |
| Cross Validation (70%, 30%, 5 folds) | 96% |

# Amazon Baby Product Review Data set

Similar approach will be followed for this dataset, which is:

1. The effect of how good/bad the underlying method
2. The number of ensembles and the effect of accuracy.

# Feature Engineering

In this dataset, the following steps applied to the testing data set rows:

1. The title and product columns combined together
2. Drop stop words
3. Take the stem for each word
4. Combine the text for each target variable value (category) into one row. This operation reduces the number of instances from thousands to 5.
5. Compute the word count for each row
6. Consider the words with count more than 150.

This will transform the dataset to a matrix of 145927 rows and 3149 columns.

# Analysis

The ensembles used in this report is a set of decision trees. To recall, based on previous assignment in decision trees the results were as follows:

|  |  |
| --- | --- |
| Tree Depth | 11 levels |
| Splitting Criteria | Entropy |
| Execution Time | 17 seconds |
| Accuracy on the Training Set | 63% |
| Accuracy on the Test Set | 61% |
| Cross Validation Mean | 62% |

Table 3: Previous Results for Decision Tree Classifier

To understand the behavior of the previous two points. Different combination of tree depth and number of estimators tested as follows:

|  |  |
| --- | --- |
| Tree Depth | Number of estimators |
| 1,2,3,…,11 | 50,150,…,450,500 |

Table 4: Tested combination

## Effect of underlying ensemble complexity

Ensemble method suggest the underlying model to do better than a chance (>50%). While 51% is higher than 50% and 100% is also higher than chance. Overfitting might appear in case of very complicated over fitted underlying model that achieve almost 100% accuracy on the test set. And this is what this report will start to illustrate. In other words, the effect of a complicated underlying models.

The figure below depicts the behavior by increasing the tree depth over different number of estimators.

Figure : Accuracy vs Tree Depth

As shown from the figure above, the higher the tree depth the higher the accuracy. The significant difference in accuracy between 500 estimator and 50 estimator on tree depth of 11 indicates that an overfitting starts to appear but not obvious by looking at 1 curve. Also, the training time is close to 1 hour and 10 minutes in cases where the tree depth is 11 and the number of ensembles is high like 500. The table below shows the training time over a high complexity stump with 11 trees.

|  |  |  |  |
| --- | --- | --- | --- |
| Depth | Estimators | Accuracy/Training | Time in minutes |
| 11 | 50 | 75% | 6.621180415 |
| 11 | 100 | 80% | 12.9602242 |
| 11 | 150 | 81% | 19.67040237 |
| 11 | 200 | 82% | 26.606824 |
| 11 | 250 | 83% | 34.13043549 |
| 11 | 300 | 84% | 41.12266126 |
| 11 | 350 | 86% | 48.04401247 |
| 11 | 400 | 87% | 55.29088905 |
| 11 | 450 | 87% | 63.12885042 |
| 11 | 500 | 88% | 69.38265448 |

Table : Execution time over complex trees with different number of estimators

As shown in table above, the training time is very large because of the complexity of the stumps and the number of estimators, which will be another metric to consider when selecting combination for the last model.

## Number of Estimators

In this experiment, the accuracy over different number of estimators studied while fixing the tree depth. In order to understand the behavior of number of estimators in the given data set.

Figure : Accuracy vs Number of Estimators

This confirms with previous figure where the more the complexity and the more trees used the higher the accuracy. And the more likely overfitting might appear.

Looking at overall runs the smaller the tree’s depth the closer the results are over different number of estimators. While the higher the tree depth will make the stumps considerably complex and over-fitted to one of the complex trees.

# Results

The selection will be based on very naïve simple stumps but using high number of estimators. The best candidates are:

|  |  |  |  |
| --- | --- | --- | --- |
| Depth | Estimators | Accuracy | Time(sec) |
| 1 | 400 | 66% | 231 |
| 2 | 400 | 67% | 357 |
| 3 | 400 | 68% | 506 |
| 1 | 450 | 66% | 268 |
| 2 | 450 | 67% | 402 |
| 3 | 450 | 68% | 569 |
| 1 | 500 | 66% | 286 |
| 2 | 500 | 67% | 447 |
| 3 | 500 | 68% | 632 |

Table : Candidates for selection

The highest accuracy with lowest acceptable training time is 400 estimators with tree depth of 3.

|  |  |
| --- | --- |
| Number of Estimators | 400 |
| Stump Tree Depth | 3 |
| Stump Accuracy Training set | 68% |
| Stump Accuracy Testing set | 65% |
| Ensemble Accuracy Training set | 60% |
| Ensemble Accuracy Testing set | 59% |
| Training Time | 506 seconds |
| Cross Validation Accuracy | 64% |