Bagging and Boosting

Abhishek Mehra akmehra

Mohit Galvankar mgalvank

Bagging

To implement bagging with classifier as decision tree, we created n bootstrap data sets by sampling with replacement from original provided train dataset where n is the "number of bags" input given . We then generated a decision tree by training each bootstrap dataset. We predicted each data point in the test set with each tree and assigned max count of prediction from all the tree predictions as the final prediction for that data point.

RESULTS –

- For depth 3

  Bag 5

```
C:\Python27\python.exe C:/Users/Mohit/PycharmProjects/Decision
Depth : 3 | Bags : 5

|  N : 2125    |   Predicted : No |   Predicted : Yes |
|--------------+------------------+-------------------|
| Actual : No  |             1597 |               496 |
| Actual : Yes |                0 |                32 |
Accuracy : 0.766588235294
```

Bags 10

```
Depth : 3 | Bags : 10

|  N : 2125    |   Predicted : No |   Predicted : Yes |
|--------------+------------------+-------------------|
| Actual : No  |             1597 |               496 |
| Actual : Yes |                0 |                32 |
Accuracy : 0.766588235294
```

- For depth 5

Bags 5

```
Depth : 5 | Bags : 5

|  N : 2125    |   Predicted : No |   Predicted : Yes |
|--------------+------------------+-------------------|
| Actual : No  |             1645 |               448 |
| Actual : Yes |                0 |                32 |
Accuracy : 0.789176470588
```

Bags 10

```
Depth : 5 | Bags : 10

|  N : 2125    |   Predicted : No |   Predicted : Yes |
|--------------+------------------+-------------------|
| Actual : No  |             1645 |               448 |
| Actual : Yes |                0 |                32 |
Accuracy : 0.789176470588
```

Weka result for bagging

```
=== Summary ===

Correctly Classified Instances        1593               74.9647 %
Incorrectly Classified Instances       532               25.0353 %
Kappa statistic                          0.0812
Mean absolute error                      0.2504
Root mean squared error                  0.5004
Relative absolute error                 45.0609 %
Root relative squared error             90.0298 %
Total Number of Instances             2125

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.746    0.000    1.000      0.746    0.854      0.206    0.873     0.996     0
                 1.000    0.254    0.057      1.000    0.107      0.206    0.873     0.057     1
Weighted Avg.    0.750    0.004    0.986      0.750    0.843      0.206    0.873     0.982

=== Confusion Matrix ===

    a    b   <-- classified as
 1561  532 |    a = 0
    0   32 |    b = 1
```

Boosting

To implement boosting with classifier as decision trees, we have majorly modified the entropy function as a function of the weights of each training sample rather than just merely the probability.

Thus for each node, we now calculate the entropy as,

$$\text{positive}_{\text{weight}} = \frac{\Sigma(\text{weights of all positive samples})}{(\text{total weights})}$$

Similarly we calculate $\text{negative}_{\text{weight}}$ and then calculate the entropy as below:

$$-(\text{positive}_{\text{weight}} \text{ x } \log(\text{positive}_{\text{weight}})) - (\text{negative}_{\text{weight}} \log(\text{negative}_{\text{weight}}))$$

After we have build a tree with the current weights, we calculate the error and update our weights, assigning higher weights to misclassified samples and lower weights to correctly classified samples.

We then build another classifier with the new weight set and keep repeating the process till we have build the number of classifier required.

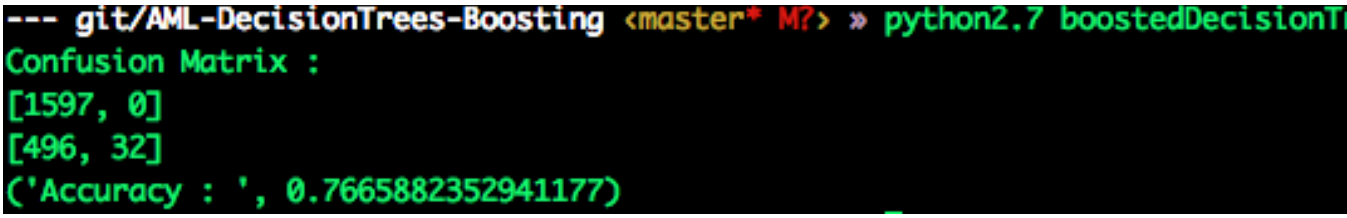
For classification of test sample, we take into consideration the sign of the results of the weighted classification by each classifier.

- Initial weights have been set to 1/n where n is the total number of samples.

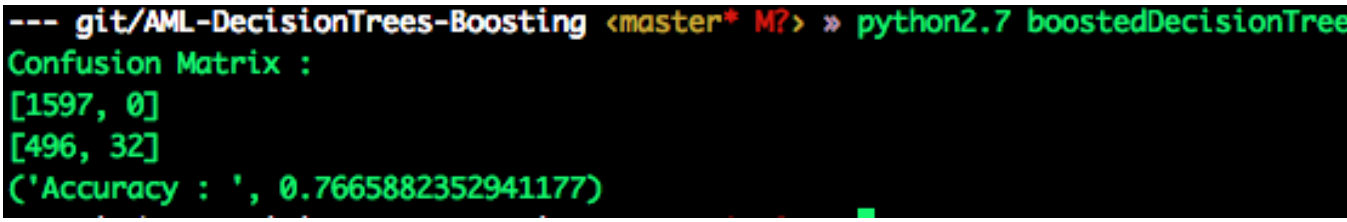
- Labels have been changed from (0,1) to (-1.1) to be able to take the sign in the end into consideration

RESULTS –

- For depth 1 and 10 classifiers

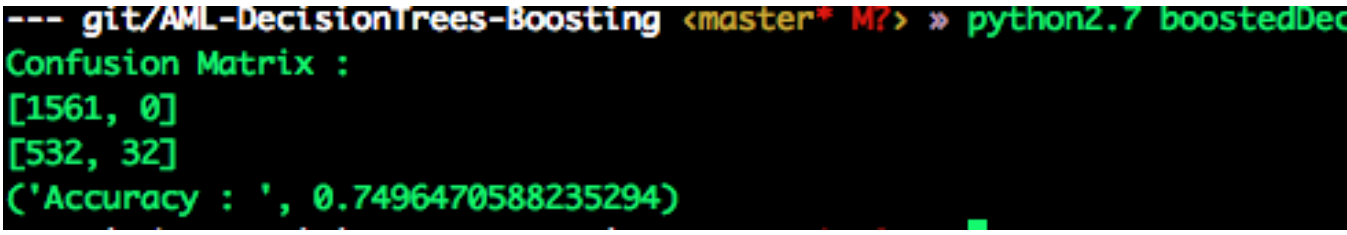


- For depth 1 and 5 classifiers



- For depth 2 and 10 classifiers

- For depth 2 and 5 classifiers

- 

Weka result for boosting

```
=== Summary ===

Correctly Classified Instances        1611               75.8118 %
Incorrectly Classified Instances       514               24.1882 %
Kappa statistic                          0.0847
Mean absolute error                      0.2452
Root mean squared error                  0.4889
Relative absolute error                 44.1395 %
Root relative squared error             87.9716 %
Total Number of Instances             2125

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.754    0.000    1.000      0.754    0.860      0.210    0.994     1.000     0
                 1.000    0.246    0.059      1.000    0.111      0.210    0.994     0.571     1
Weighted Avg.    0.758    0.004    0.986      0.758    0.849      0.210    0.994     0.993

=== Confusion Matrix ===

    a    b   <-- classified as
 1579  514 |    a = 0
    0   32 |    b = 1
```