**A report on fitting a classifier to the MIT8 dataset**

Henry Tu, 17-oct-2015

**Q1. Load the feature table.**

|  |  |
| --- | --- |
|  | Figure 1: The snapshot of the data table  We work on the data table with 2688 data rows.  There are 63 input features (columns) one source column to contain the image source file (for rendering purpose) and one tag column to contain the output tag / category  See data.csv or data.xlsx for details. |

number of columns: 65 (63 input feature + 1 source file + 1 output tag)

number of rows: 2688

We try to use color averaging to explore if we can use color and spatial information to classify the images

**Q2. Design a classifier to give a tag for each image.**

We try to explore 3 models (tree, forest, and boosting) and we have the following data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| features | rpart-A | rpart-B | rf-A | rf-B | gbm-A | gbm-B |
| 10 | 0.76567 | 0.815436 | 0 | 0.401007 | 0.147436 | 0.508389 |
| 20 | 0.760513 | 0.81407 | 0 | 0.331658 | 0.062723 | 0.413735 |
| 50 | 0.683535 | 0.670017 | 0.000713 | 0.249581 | 0.013542 | 0.291457 |
| 63 | 0.727208 | 0.729866 | 0 | 0.253782 | 0.012802 | 0.271044 |

And we can visualize the table in 3 charts

|  |  |
| --- | --- |
|  | For decision tree, we saw that both training (A) and testing error (B) are high.  We cannot use tree for this dataset  See exploring-chart.xlsx for details. |
|  | For random forest, we saw that training error (A) is quite low, almost zero. and the testing error (B) is around 0.25 when we reach 50 features. We don't even have to use 63 features.  See exploring-chart.xlsx for details. |
|  | For boosting, we saw that the training error (A) is quite high (0.3) and the testing error (B) is also non zero when we use small number of features.  See exploring-chart.xlsx for details. |

We choose to use Random Forest as the model for classification

**Q3. Produce the output**

We use Random Forest on 30% of 2000 images and we have the following result

|  |  |
| --- | --- |
|  | Figure: the output result on the 30% dataset.  See output-enduser.html for details |