Homework 1: CS 498 AML (ONL)

Ramya Narayanaswamy : rpn2

Problem 1:

Part 1A:

Code : prob1a.R
Results: Accuracy of evaluation data = 75.03%
Analysis: Cross-validation over 10 runs and accuracy was averaged over 10 runs of evaluation data
References: Professor's sample code in course website

Part 1B:

Code : prob1b.R
Results: Accuracy of evaluation data = 74.64%
Analysis:Cross-validation over 10 runs and accuracy was averaged over 10 runs of evaluation data
References: Professor's sample code in course website

Part 1C:

Code : prob1c.R
Results: Accuracy of evaluation data = 76.47 %
References: Professor's sample code in course website

Part 1D:

Code : prob1d.R
Results: Accuracy of evaluation data = 74.51 %
References: Professor's sample code in course website

Problem 2

Part 2A

Code : prob2aOriginal.R and prob2aCropped.R
Results:

Accuracy (%)	Gaussian	Bernoulli
Untouched images	61.37	84.54
Stretched bounding box	82.94	81.42

Analysis: naive_bayes function from naivebayes library was used for both Gaussian and Bernoulli model. The thresholding was set to 35 as the accuracy was marginally better with this value. Threshold was applied for both Gaussian and Bernoulli model. Numeric features were used for Gaussian and categorical features were used for Bernoulli.
Model Comparison :

For untouched images, Bernoulli is better from the above results. This is because the features take only 2 values for a binarized image and it fits naturally for a bernoulli model for discrete variables. Gaussian model fits well for continuous variables and the model from binary data set is not accurate. Moreover, there is a lot of white space or empty space in images (value 0) in untouched images. These empty spaces skew the normal distribution model to a certain degree, lowering the overall accuracy.

For stretched images, the accuracies of both Gaussian and Bernoulli are close enough. This is because, not so useful pixels from untouched images have been removed due to resizing. This has caused an increase in accuracy of Gaussian model and is marginally better than Bernoulli.

References: Relevant Piazza posts : @129, @83, @131

Part 2B

Code : prob2bOriginalh20.R and prob2bCroppedh20.R
Results:
Untouched Images:

Accuracy (%)	Depth = 4	Depth = 8	Depth = 16
#trees = 10	83.12	92.85	96.06
#trees = 20	85.45	93.35	96.45
#trees = 30	86.16	93.36	96.69

Stretched Bounding Box Images:

Accuracy (%)	Depth = 4	Depth = 8	Depth = 16
#trees = 10	83.23	93.47	96.23
#trees = 20	84.73	94.23	96.83
#trees = 30	85.6	94.27	97.04

Analysis:

randomForest function from H2O library was used. The accuracy of both untouched and stretched image is similar based on the above results. This is because an ensemble classifier like Random Forest is capable of averaging out feature abnormalities or irrelevant features (like white spaces in original image) over multiple decision trees. Hence, the accuracy results are similar.

However, the accuracy increased when the depth of the trees increased. Increase in number of trees for same depth did not produce significant improvement of accuracy. For this data set, depth was a dominant factor in determining accuracy. The main reason is the large number of features (784 and 400). With large number of features that are not redundant, increase in depth makes the classifier to use as many features as possible within a single tree and eventually leading to improved accuracy in a single tree. Overall, the average accuracy of an ensemble of trees is improved too. Moreover, feature sampling will have higher probability of choosing more unseen features if the depth of trees is increased.

References: Relevant Piazza posts : @129, @83, @131, @85