Lab 4-Binary Classifier
Stat 215A, Fall 2014

PLEASE WRITE YXiang (Lisha) Li

November 7, 2014

1 Introduction

Blah blah...

2 EDA

2.1 Densities of the three most important features as selected by paper

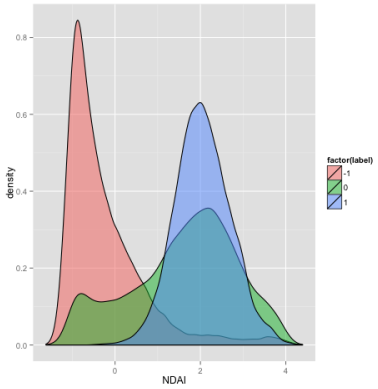

Figure 1: NDAI density Image 1

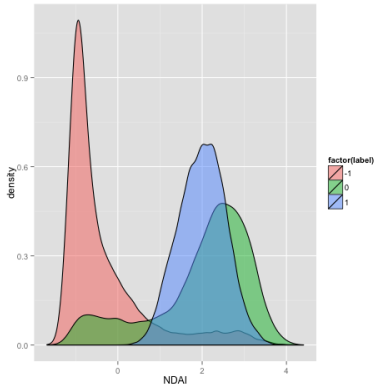

Figure 2: NDAI density Image 2

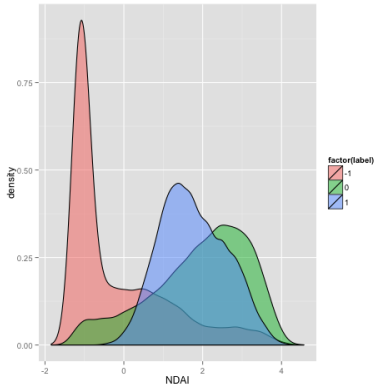

Figure 3: NDAI density Image 3

3 Modeling

3.1 LDA

3.2 QDA

Cross-validation for QDA revealed that while the method works extremely well in some cases, producing AUC scores of almost 1, it sometimes fails to perform better than even the theoretical random classifier.

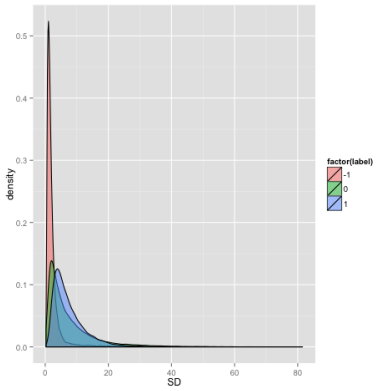

Figure 4: SD density Image 1

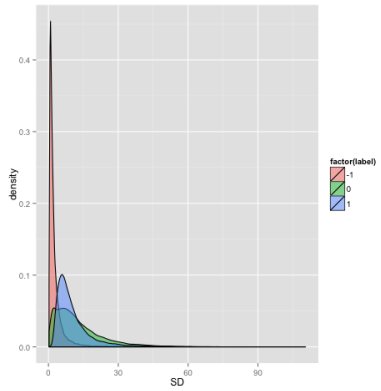

Figure 5: SD density Image 2

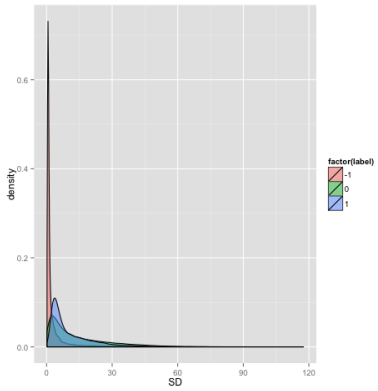

Figure 6: SD density Image 3

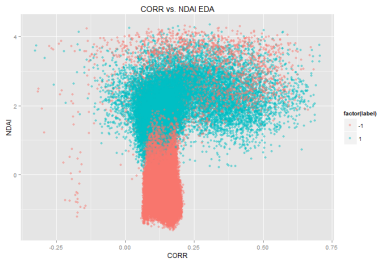

Figure 7: CORR vs. NDAI Plot of Image 1

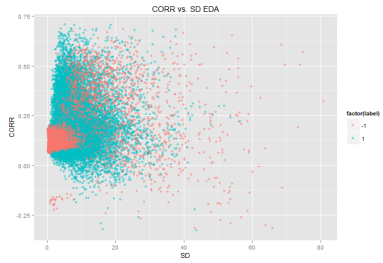

Figure 8: CORR vs. SD Plot of Image 1

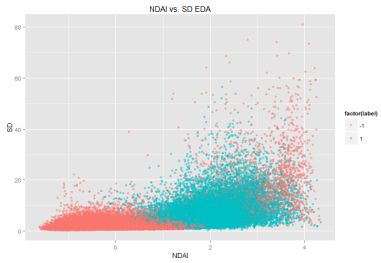

Figure 9: NDAI vs. SD Plot of Image1

3.3 Logit/Probit

3.4 Random Forest

For random forest, as in all the other classifiers, we divided the three images into equal sized quadrants (2X2) rectangles in order to do 12 fold validation on the dataset. That is, for each iteration of the validation, we dropped one of the quadrants as a test set, and trained on the remaining 11 quadrants. Keeping the images segments disjoint and continuous ensured that our models were picking up on 'higher' level structure of the dataset, and not the continuous variation of neighbouring pixels. We also trained on each image and tested on the remaining two. To test convergence, one of the things we did was increase the training set from including 1 quadrant, to including 2 quadrants, up to including 11 quadrants (using the complement as the test set). For each of the 3 aforementioned classes of training, we trained on a range of forest sizes, from 2 trees to 50. Here are the results:

Finally, this entire set of training models was done with all the features, and then restricted to only SD, CORR and NDAI. These three were particularly chosen because their GiniImportance was consistently ranked at least 2fold above the next highest in all the cross validations. As we also saw with the random forest model that just used SD, CORR, NDAI, it did not fare poorly compared to training the forests on all 9 predictors.

Cross Validation ROC curves

ROC curves for cross validation between images

4 Reproducibility

How we organized out code and github repo

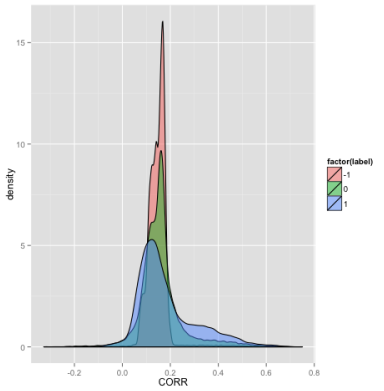

Figure 10: CORR density Image 1

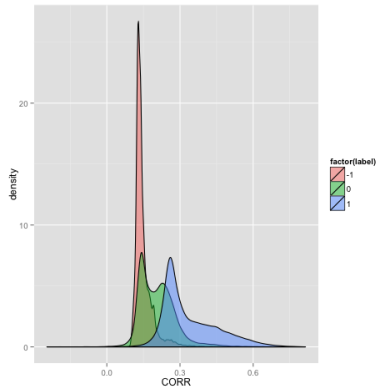

Figure 11: CORR density Image 2

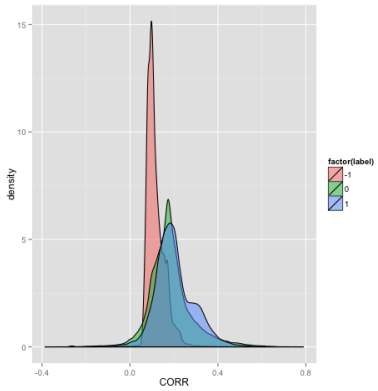

Figure 12: CORR density Image 3

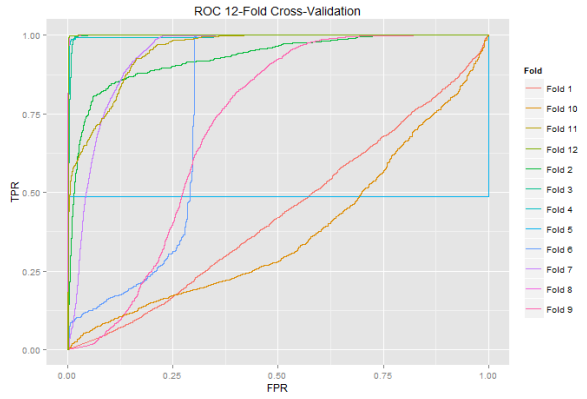

Figure 13: Response Operator Curve for 12-fold Cross-valudation of QDA

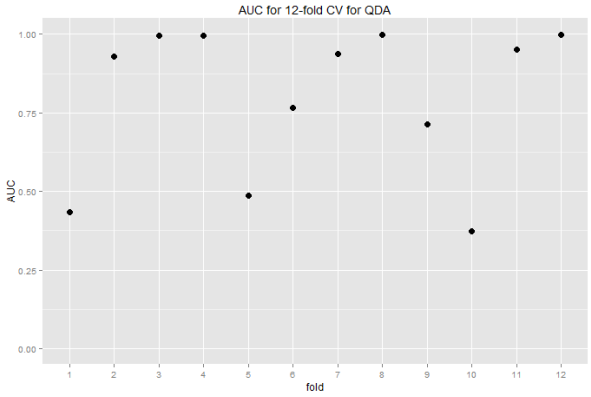

Figure 14: AUC for QDA Classifiers

References

[1] Ben-Hur, A., Elisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In: Pacific Symposium on Biocomputing, pp. 617.

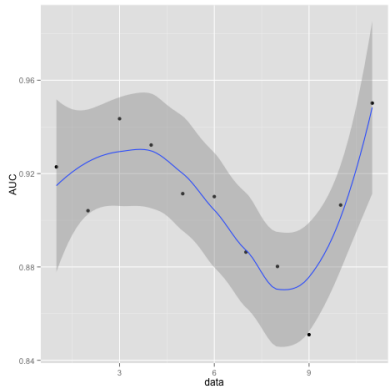

Figure 15: Smoothed convergence of AUC for growing training set 50 trees and 3 features

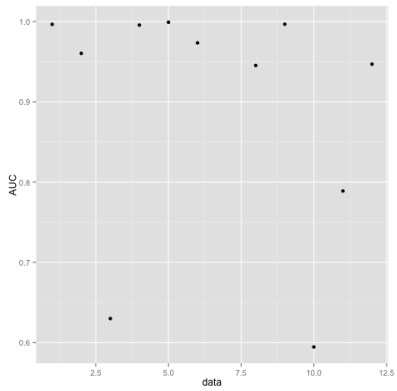

Figure 16: AUC of test set in each fold with 50 trees and 3 features

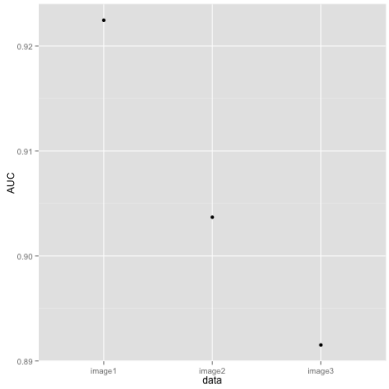

Figure 17: AUC of the three images with 50 trees and 3 features

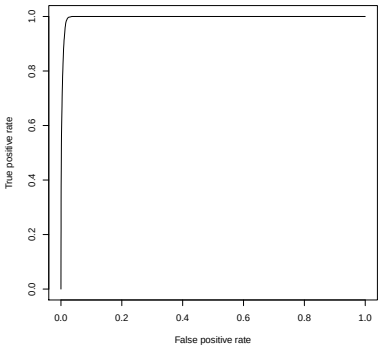

Figure 18: ROC fold 1

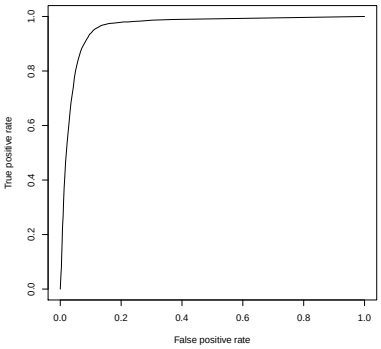

Figure 19: ROC fold 2

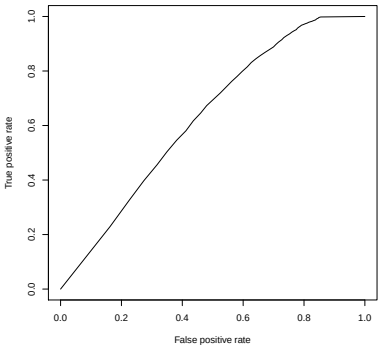

Figure 20: ROC fold 3

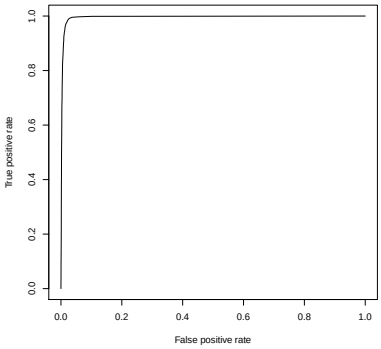

Figure 21: ROC fold 4

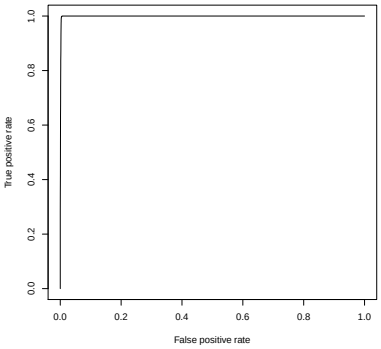

Figure 22: ROC fold 5

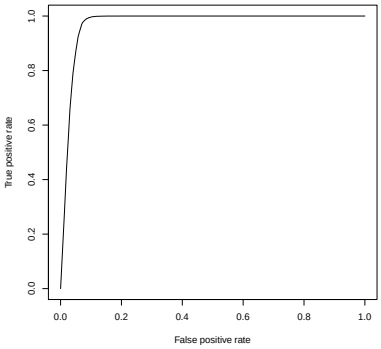

Figure 23: ROC fold 6

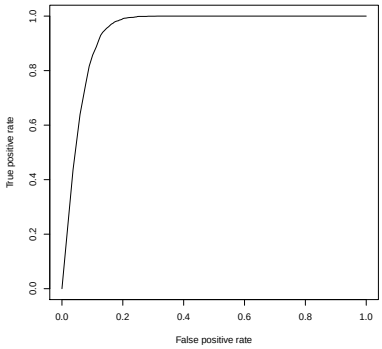

Figure 24: ROC fold 8

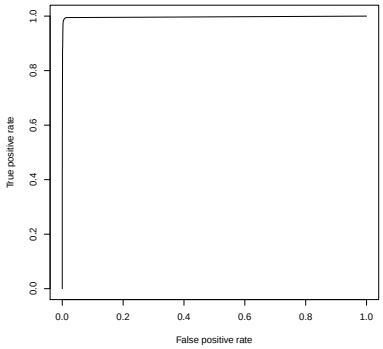

Figure 25: ROC fold 9

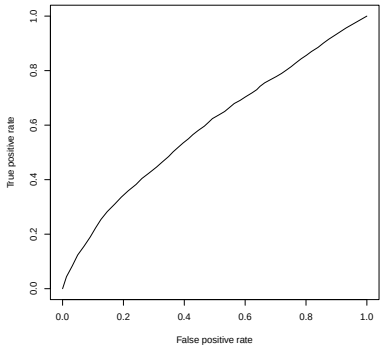

Figure 26: ROC fold 10

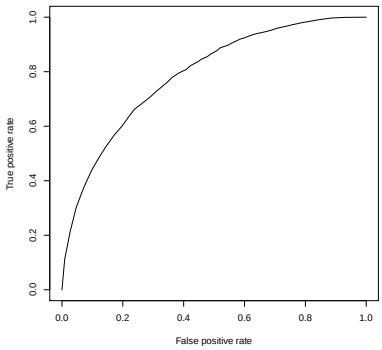

Figure 27: ROC fold 11

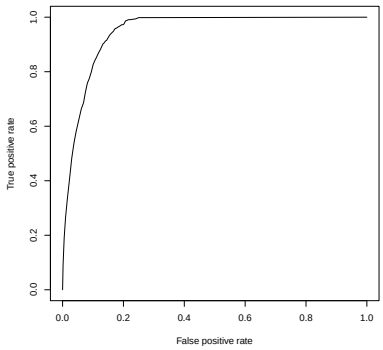

Figure 28: ROC fold 12

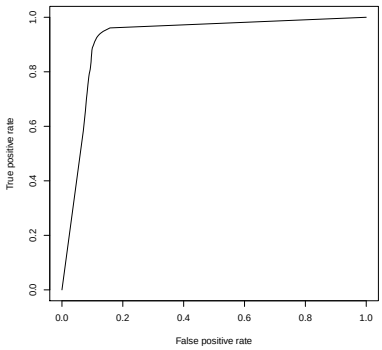

Figure 29: Trained on image1

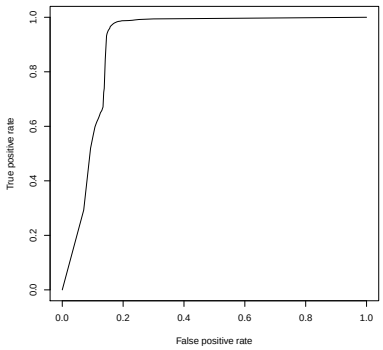

Figure 30: Trained on image2

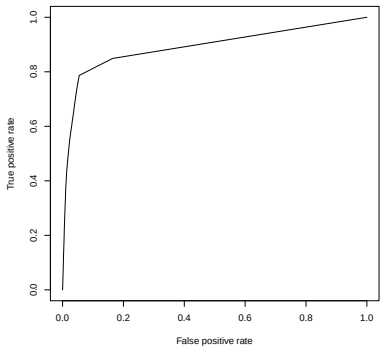

Figure 31: Trained on image3