CS 334 - Homework 5

Alex Welsh

November 13th 2020

1 Question 1 - PCA

a. An unregularized logistic regression model was trained on data normalized with Standard Scaler. The test AUC was 0.88. The code can be found in **pca.py**.

b. PCA was run on the normalized data. 9 components were needed to capture 97.9% of the variance. For Component 0, fixed acidity (.49), citric acid (.46), and pH (-.43) had large variance values. For Component 1, free sulfur dioxide (.50), total sulfur dioxide (.58), and alcohol (-.42) had large variance values. For Component 2, volatile acidity (-.45), free sulfur dioxide (.46), and alcohol (.46) had large variance values. The code can be found in **pca.py**.

c. The two models were trained and the ROC curves generated. They both had similar AUC scores of .88 so that means the models perform about the same. The ROC curves can be seen in Figure 1 and the code can be found in **pca.py**.

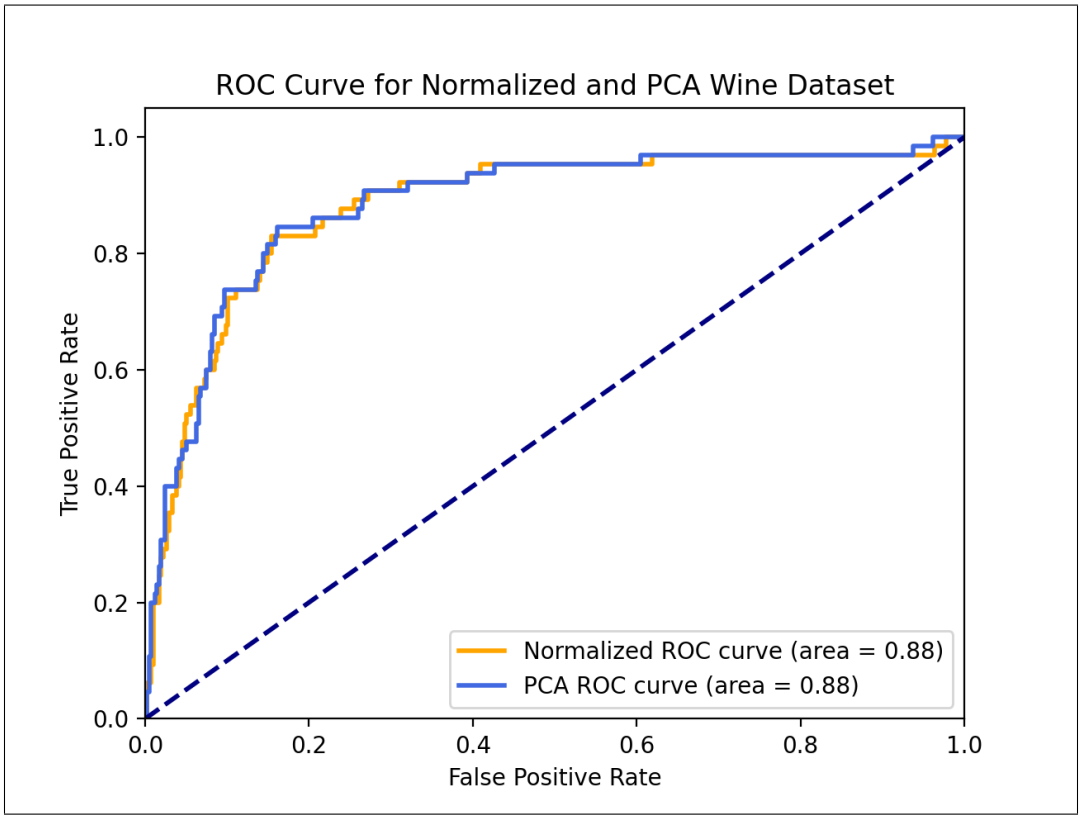

Figure 1. ROC curves for logistic regression and PCA.

2 Question 2 - Almost Random Forest

a. The random forest model was created. The OOB error is about .11. The code can be found in **rf.py**.

b. I used my optimal parameters from homework 2 for the decision tree which was entropy criteria, min leaf sample of 4 and max depth of 8 as a start point. I first used those variables as the initial parameters and then did a grid search by varying the nest ([2, 5, 7, 10, 15, 20, 25, 30]) and max features ([3, 4, 5, 6, 7, 8, 9]). I found that the best model was with nest of 15 and max feature of 6. It had a classification error of 0.114583. A 3D graph of the search space can be see in Figure 2. I then made those constant and searched by varying min leaf sample ([2, 5, 7, 10, 15, 20, 25, 30]) and max depth ([3, 4, 5, 6, 7, 8, 9]). The ideal parameters were min leaf sample of 2 and max depth of 9. It had a classification error of 0.104167. A 3D graph of the search space can be see in Figure 3. Thus, the final model with the best hyper-parameters is nest=15, max features=6, min leaf sample=2, and max depth=9. The final The code can be found in **rf.py**.

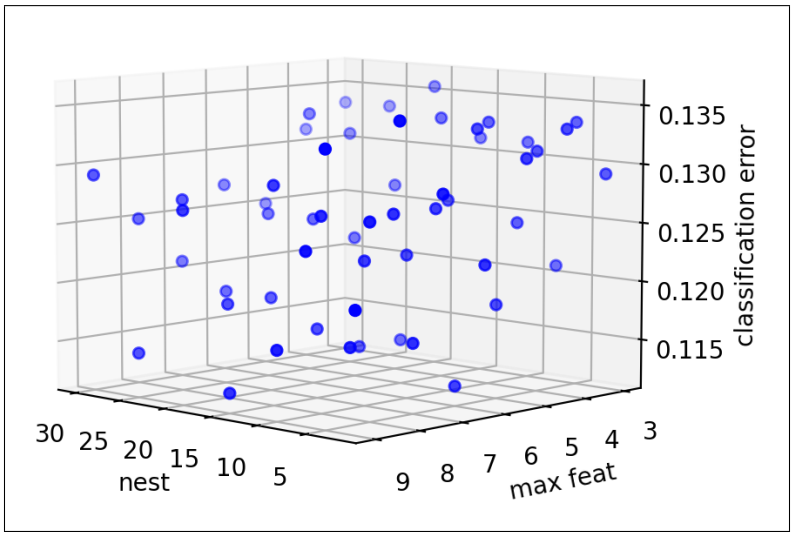

Figure 2. Classification error of varying nest and max feature values.

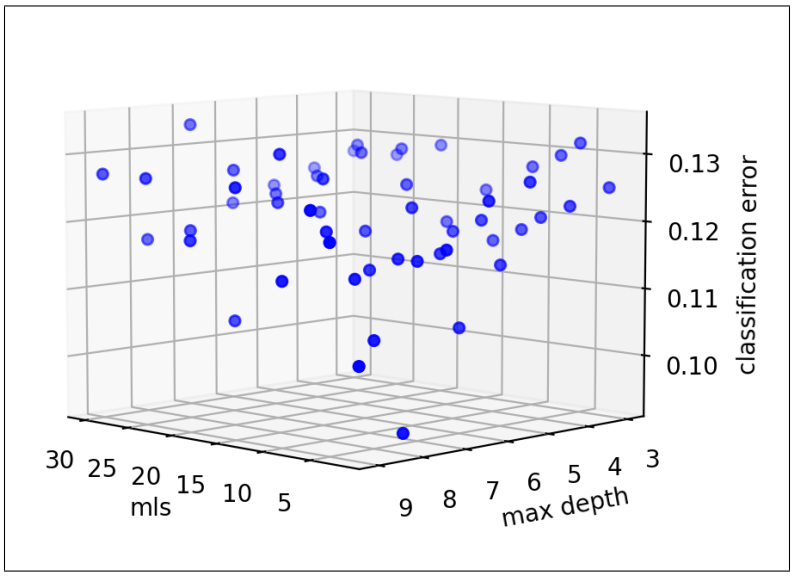

Figure 3. Classification error of varying min leaf sample and max depth values.

c. For the model created with the optimal parameters in b. The average OOB error is 0.127. The classification error is 0.0895 does better than the OOB error. The code can be found in **rf.py**.