Joseph Loss (loss2)

IE598 MLF F18

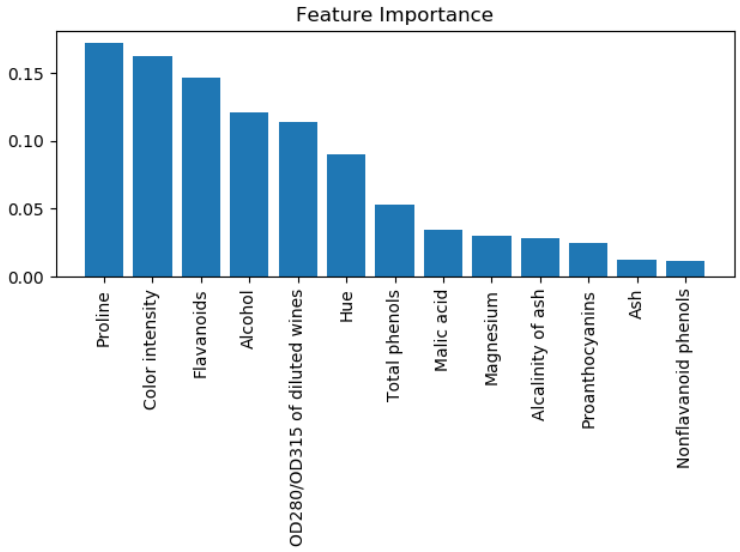
Module 7 Homework (Random Forest)

Part 1: Random forest estimators

Forest	n_estimators	CV Score (in-sample)	CV Score (out-of-sample)
1	25	0.975	0.8333
2	100	0.975	0.8333
3	500	0.982	0.9500
4	1,000	0.982	0.9500
5	10,000	0.975	0.9500

Part 2: Random Forest - Feature Importance
Best Model: CV4, n_estimators = 1000

1)	Proline	17.20%
2)	Color intensity	16.27%
3)	Flavanoids	14.62%
4)	Alcohol	12.10%
5)	OD280/OD315 of diluted wines	11.34%
6)	Hue	8.97%
7)	Total phenols	5.33%
8)	Malic acid	3.47%
9)	Magnesium	3.04%
10)	Alcalinity of ash	2.81%
11)	Proanthocyanins	2.44%
12)	Ash	1.25%
13)	Nonflavanoid phenols	1.16%



Part 3: Conclusions

As the n_estimators parameter increases, the in-sample CV accuracy and the computation time both increase as well. The optimal number of estimators seemed to be 1,000. This number led to the best scoring with the highest computational efficiency.

Feature importance is the average impurity decrease of the dataset. This is computed from all decision trees in the forest and does not make/need any assumptions about whether the data is linearly separable or not. As can be seen in Forest 4 (my optimal forest): the top 3 most-important features are Proline (17.2%), Color intensity (16.3%), and Flavanoids (14.6%).

Part 4: Appendix

[loss2_IE598_F18_HW7 - Github Repo]