My Name (yuhangm2)

IE598 MLF F18

Module 7 Homework (Random Forest)

Using the Wine dataset, described in Raschka chapter 4 and 10 fold cross validation;

Part 1: Random forest estimators

Fit a random forest model, try several different values for N_estimators, report in-sample accuracies.

|	0	1	2	3	4	5	6	7	8	9	10
n_estimators	1.0	2.0	5.0	10.0	50.0	100.0	500.0	1000.0	2500.0	5000.0	10000.0
in_sample_accuracy(%)	86.3	87.9	91.1	96.8	96.8	97.6	97.6	97.6	97.6	97.6	97.6
sys_time(ms)	0.2	101.8	102.8	104.1	104.3	104.1	104.6	104.7	307.0	508.9	944.8

Part 2: Random forest feature importance

Display the individual feature importance of your best model in Part 1 above using the code presented in Chapter 4 on page 136. {importances=forest.feature_importances_ }

```
{'n_estimators': 100}

best score: 0.981481

 1) Alcohol                        0.194819
 2) Malic acid                     0.164348
 3) Ash                            0.145228
 4) Alcalinity of ash              0.119075
 5) Magnesium                      0.088747
 6) Total phenols                  0.080266
 7) Flavanoids                     0.061579
 8) Nonflavanoid phenols           0.039190
 9) Proanthocyanins                0.029782
10) Color intensity                0.023107
11) Hue                            0.022962
12) OD280/OD315 of diluted wines   0.016638
13) Proline                        0.014258
```
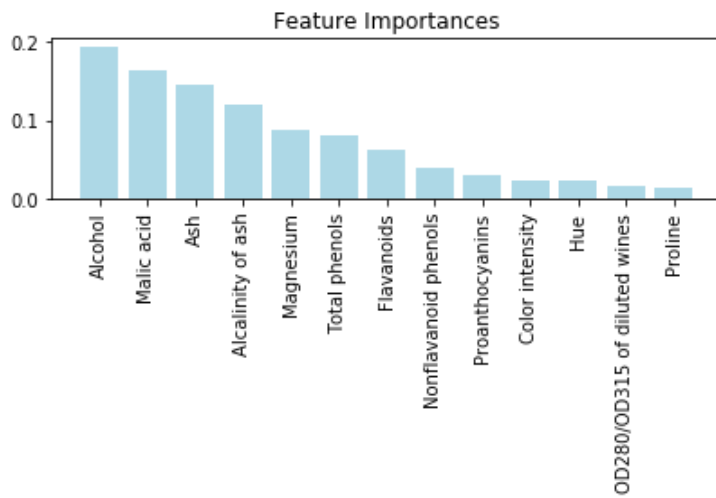
Part 3: Conclusions

Write a short paragraph summarizing your findings.

- What is the relationship between n_estimators, in-sample CV accuracy and computation time?
- What is the optimal number of estimators for your forest?
- Which features contribute the most importance in your model according to scikit-learn function?
- What is feature importance and how is it calculated? (If you are not sure, refer to the Scikit-Learn.org documentation.)

1.As table shown on Part 1, with the increasing of n_estimators, both in-sample CV accuracy and computation time rising.

2.The optimal number of estimators is 100.

3.'Alcohol' is the most important feature contributed in my model.

4.The feature importance is the mean decrease impurity(here is gini importance) for all decision trees. The results are shown after normalization, so their sum are 1.

Part 4: Appendix

Link to github repohttps://github.com/YHM3/IE598_F18/tree/master/IE%20598_F18_HW7
(https://github.com/YHM3/IE598_F18/tree/master/IE%20598_F18_HW7)

Toggle Code