Ruozhong Yang (ry8)

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 N_estimators form 1 to 100, report in-sample accuracies. We can get a list like this:

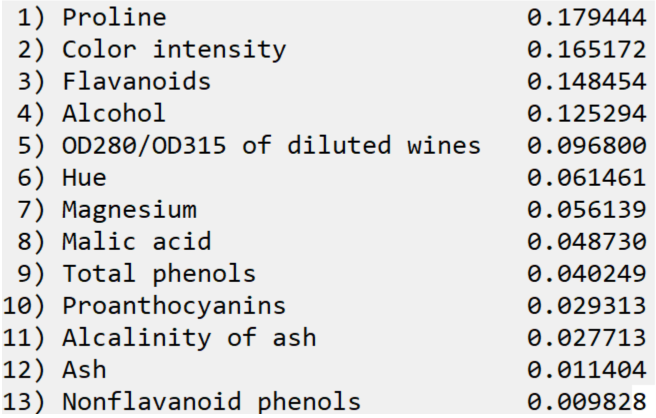
N estimators	in-sample accuracies	N estimators	in-sample accuracies	N estimators	in-sample accuracies	N estimators	in-sample accuracies
1	0.8541667	26	0.9748611	51	0.9748611	76	0.9748611
2	0.8575	27	0.9748611	52	0.9748611	77	0.9748611
3	0.94375	28	0.9681944	53	0.9748611	78	0.9748611
4	0.9304167	29	0.9748611	54	0.9748611	79	0.9748611
5	0.9308333	30	0.9748611	55	0.9748611	80	0.9748611
6	0.9308333	31	0.9748611	56	0.9748611	81	0.9748611
7	0.9375	32	0.9748611	57	0.9748611	82	0.9748611
8	0.9563889	33	0.9748611	58	0.9748611	83	0.9748611
9	0.9426389	34	0.9748611	59	0.9748611	84	0.9748611
10	0.9430556	35	0.9748611	60	0.9748611	85	0.9748611
11	0.9430556	36	0.9748611	61	0.9748611	86	0.9748611
12	0.9430556	37	0.9748611	62	0.9748611	87	0.9748611
13	0.9493056	38	0.9748611	63	0.9748611	88	0.9748611
14	0.9681944	39	0.9748611	64	0.9748611	89	0.9748611
15	0.9681944	40	0.9748611	65	0.9748611	90	0.9748611
16	0.9681944	41	0.9748611	66	0.9748611	91	0.9748611
17	0.9681944	42	0.9748611	67	0.9748611	92	0.9748611
18	0.9748611	43	0.9748611	68	0.9748611	93	0.9748611
19	0.9748611	44	0.9748611	69	0.9748611	94	0.9748611
20	0.9681944	45	0.9748611	70	0.9748611	95	0.9748611
21	0.9681944	46	0.9748611	71	0.9748611	96	0.9748611
22	0.9748611	47	0.9748611	72	0.9748611	97	0.9748611
23	0.9748611	48	0.9748611	73	0.9748611	98	0.9748611
24	0.9748611	49	0.9748611	74	0.9748611	99	0.9748611
25	0.9748611	50	0.9748611	75	0.9748611	100	0.9748611

With the rise of the n estimators the accuracies scores raised(more decision tree), however the time also increased. Since the result when n>=30 is almost the same, we make the N estimators equals to 30.

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_ }

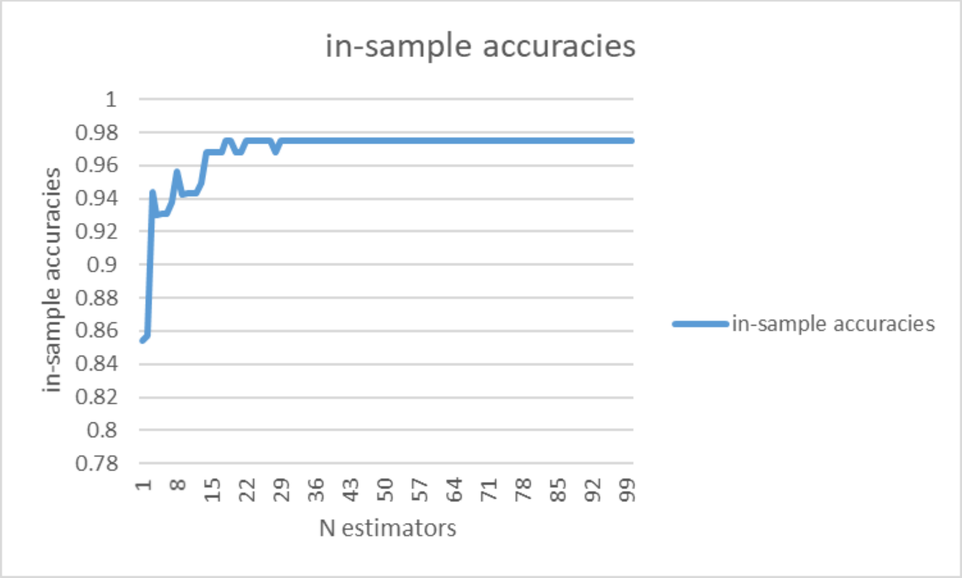
The result is as below.



```
 1) Proline                          0.179444
 2) Color intensity                  0.165172
 3) Flavanoids                       0.148454
 4) Alcohol                          0.125294
 5) OD280/OD315 of diluted wines     0.096800
 6) Hue                              0.061461
 7) Magnesium                        0.056139
 8) Malic acid                       0.048730
 9) Total phenols                    0.040249
10) Proanthocyanins                  0.029313
11) Alcalinity of ash                0.027713
12) Ash                              0.011404
13) Nonflavanoid phenols             0.009828
```

Part 3: Conclusions

When the n_estimator raise, the accuracy score raise until the n_estimators big enough(30). At the same time, the computation time also raise. The optimal number in this program is 30. The program calculates the variance and then calculates the mean of all trees' variance.



Part 4: Appendix

https://github.com/yrz437396236/IE598_F18_HW1/tree/master/IE598_F18-HW7