My Name (myNetID)

IE598 MLF F18

Module 6 Homework (Cross validation)

Part 1: Random test train splits

In this part, we use the Iris dataset, with 90% for training and 10% for test and the decision tree model. Then, we run in-sample and out-of-sample accuracy for 10 different samples by changing random_state from 1 to 10 in sequence and get the table below.

random state	score(train)	mean(train)	std(train)	score(test)	mean(test)	std(test)
1	0.992593	0.99407	0.00444	1	0.94	0.0696
2	1			0.933333		
3	0.992593			1		
4	1			0.933333		
5	0.992593			0.8		
6	0.992593			1		
7	0.985185			0.866667		
8	1			0.866667		
9	0.992593			1		
10	0.992593			1		

Part 2: Cross validation

Now rerun the model using cross_val_scores with k-fold CV (k=10).

cross val score	mean	std	out-of-sample
1	0.95333	0.05364	0.866666667
0.93333333			
1			
1			
0.93333333			
1			
0.91666667			
1			
0.91666667			
0.83333333			

Part 3: Conclusions

According to our result, we can find that random test train splits bring better test score with bigger standard deviation, which means more “accurate” and less “precise”. The cross validation is more more “precise” and less “accurate”.

Part 4: Appendix

https://github.com/yrz437396236/IE598_F18_HW1/tree/master/IE598_F18-HW6