Module 6 Homework (Cross validation)

Zhicheng Tang zt17

Part 1: Random test train splits

In-sample mean	In-sample variance	Out-sample mean	Out-sample variance
1	0	0.7221	2.49e-05

Decision tree model can easily overfit the training data.

Part 2: Cross validation

In-sample cv mean	In-sample cv variance	Out-sample accuracy
0.7210	7.45e-05	0.7306

Conclusions

Two methods are similar I think, even though in random test train splits, the in-sample accuracies are 100%. When we are tuning the model, we actually don't care and inspect out-sample's result to iterate our model's parameters. The difference is that cross validation is kind of random test train splits without replacement.

I couldn't really tell which method provides the best estimate, because the accuracy on test data are actually similar. If I have to make a choice, I think cross validation method generally performs better.

For the efficiency, I think if they both run the same number of iterations, like in our quations both are 10 iterations, then there is no much difference. However, since cross validation is a split without replacement, it's more likely to require fewer iterations to let each part of the data serves as validation set. Thus, in general, random test train splits require more iteration to fully make sure every part of data have appeared in validation and takes more time.

Part 4: Appendix

Link to github repo