Jingxia Zhu(jingxia6)

IE598 MLF F18

Module 6 Homework (Cross validation)

Using the Iris dataset, with 90% for training and 10% for test and the decision tree model that you submitted for Module 2:

Part 1: Random test train splits

Run in-sample and out-of-sample accuracy for 10 different samples by changing random_state from 1 to 10 in sequence.

Display the individual scores, calculate the mean and standard deviation of the set. Report in a table format.

	Random state	In-sample accuracy scores	Out-of-sample accuracy scores
	1	1.0	1.0
	2	1.0	0.9333333333333333
	3	1.0	1.0
	4	1.0	0.9333333333333333
	5	1.0	0.8
	6	1.0	1.0
	7	1.0	0.8666666666666667
	8	1.0	0.8666666666666667
	9	1.0	1.0
	10	1.0	1.0
Mean		1.0	0.9400000000000001
Standard deviation		0.0	0.06960204339273698

Part 2: Cross validation

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

Report the individual fold accuracy scores, the mean CV score and the standard deviation of the folds. Now run the out-of-sample accuracy score. Report in a table format.

	Accuracy fold 1	0.93333
	Accuracy fold 2	0.93333
	Accuracy fold 3	1.0
	Accuracy fold 4	1.0
	Accuracy fold 5	0.92857
	Accuracy fold 6	1.0

	Accuracy fold 7	0.91667
	Accuracy fold 8	0.91667
	Accuracy fold 9	0.91667
	Accuracy fold 10	0.91667
Mean		0.946190476190476
Standard deviation		0.03579673023743385
Out-of-sample accuracy	1.0	

Part 3: Conclusions

Write a short paragraph summarizing your findings. Which method of measuring accuracy provides the best estimate of how a model will do against unseen data? Which one is more efficient to run?

Random test train split is to split a data set into test and training set based on a certain percentage. Cross validation splits the data into k equal sized folds and repeats k times. Every time there is one fold is used as test set and the rest k-1 folds as the training data. The k-fold cross validation provides a better estimate for how a model will do for an unseen data set. That's because it sets the weight equally k times and, in this way, we can avoid the possibility that some data is sampled too often, and some are rarely sampled. And we can see from the above result tables, cross validation has a higher mean and lower variation, and also reports an out-of-sample accuracy of 1.0. Random test rain split may be more computationally efficient because it is only split once and usually it gives us a result that is not bad.

Part 4: Appendix

Link to github repo: https://github.com/jzhuuhzj/IE598_F18_HW6.git