Problem Set 10

Samuel Hunt

April 15, 2025

1 Classification Models using the Machine Learning Tribes

For this problem set we attempted to classify each person in UC Irvine's Income dataset as high-income or not based on a variety of other variables. We used 4 of the 5 machine learning tribes and 5 total models. The first model was the Logit model. We first found an optimal parameter of 1e-10, and then an out-of-sample accuracy of .8526. The next model was a tree model, and we found optimal parameter values of 0.001 for cost complexity, 15 for tree depth, and 10 for minimum split sample size. This model had an out-of-sample accuracy of 0.8684. Third, we used a neural net model and tuned to get optimal parameter values of 1 for the penalty and 4 for the number of hidden units. This resulted in an accuracy of 0.8567. Finally, under the analogizers branch, we ran a kNN and SVM model. The kNN model had optimal parameters of 30 neighbors and an accuracy of 0.8434. The SVM model had optimal parameters of 2^0 for the cost, 2^-2 for the rbf sigma, and an accuracy of 0.8640. All of these results are contained in the table below. Overall, the tree model had the highest accuracy although each of the models was fairly similar at around 85 percent accuracy.

penalty	.estimate	alg	cost_complexity	tree_depth	min_n	hidden_units	neighbors	cost	rbf_sigma
0.0000	0.8526	logit							
	0.8684	tree	0.0010	15.0000	10.0000				
1.0000	0.8567	nnet				4.0000			
	0.8434	knn					30.0000		
	0.8640	svm						1.0000	0.2500