

ANALYSIS:

With digits dataset:

running Perceptron

Error score is: 0.937037037037037

Runtime in seconds: 0.07000565528869629

running SVM linear

Error score is: 0.9796296296296296

Runtime in seconds: 0.07772183418273926

running SVM with RBF kernel

Error score is: 0.7351851851851852

Runtime in seconds: 0.6229462623596191

running decision tree

Error score is: 0.8537037037037037

Runtime: 0.016900062561035156

running KNN

Error score is: 0.9722222222222222

Runtime: 0.1886587142944336

running logistic regression

Error score is: 0.9574074074074074

Runtime: 0.3323488235473633

With READLISP dataset:

running Perceptron

Error score is: 0.40177989611066633

Runtime in seconds: 3950.7089507579803

running SVM linear

Error score is: 0.7922399508480572

Runtime in seconds: 73923.22957277298

running SVM with RBF kernel

Error score is: 0.5936400364915939

Runtime in seconds: 76162.44222640991

running decision tree

Error score is: 0.902105713913351

Runtime: 1505.1740338802338

running KNN

Error score is: 0.7038781627599561

Runtime: 1344.4371984004974

running logistic regression

Error score is: 0.6782349010958327

Runtime: 24023.12098472181984

From these results, we see that the digits set runs significantly faster, and also has a better score overall (found using sklearn's score() method), with the exception of using DecisionTree. All dataset instances were standardized.

Understanding decision tree classifier:

Two methods used by the classifier to or pre-prune are min_samples_split() and min_weight_fraction_leaf(). These methods appear to help reduce the depth associated with the tree.

The post-prune method is found at line 87-100 of file tree.py

The pre-prune method is found at line 89, 238 of file tree.py