Junle Lu

Professor Charles Isbell

CS 4641: Machine Learning

Fall 2018

Supervised Learning

Abstract

In this pager, we performed detailed analyses of five different supervised learning algorithms: Decision Tree, AdaBoosting, k-nearest Neighbors, Neural Network, Support Vector Macchine. Two data sets are used for the analysis: Diabetic Retinopathy Debrecen and Gender Recognition by Voice. The model complexity and learn curve are compared between different algorithms. The. The algorithm implementation are used from python library “scikit-learn” with version number 3.7 and 0.19.2 respectively.

Dataset Introduction

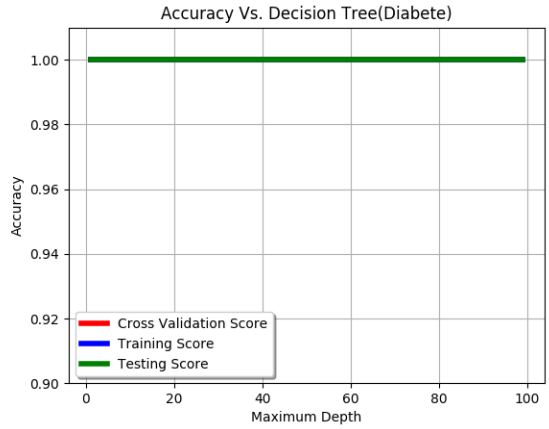
Diabetic Retinopathy Debrecen Data Set: this dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. All features represent either a detected lesion, a descriptive feature of a anatomical part or an image-level descriptor. It has 1151 instance and 20 features.

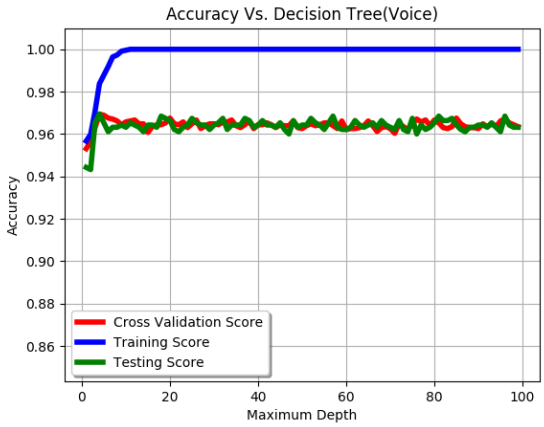
Gender Recognition by Voice: this database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. The dataset consists of 3,168 recorded voice samples, collected from male and female speakers. The voice samples are pre-processed by acoustic analysis in R using the seewave and tuneR packages, with an analyzed frequency range of 0hz-280hz. It has 3168 instance and 20 features.

Training and Test Datasets

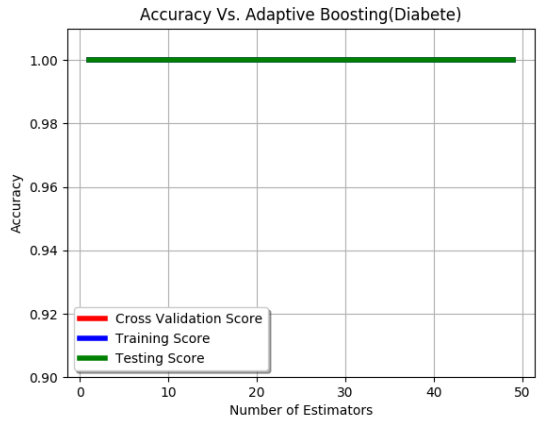
Each data set is split into training and testing data sets. 70% of the dataset will be the training set and the other 30% is the testing set. All the mentioned five algorithms are applied on the training set first and then the testing set. Hyperparameters are tested and tuned with 10-fold cross validation method on the training set for each classifier. The testing set is only used to test the accuracy of the trained models.

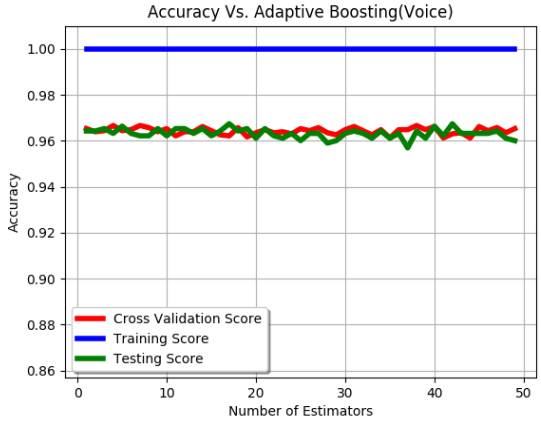
Decision trees:



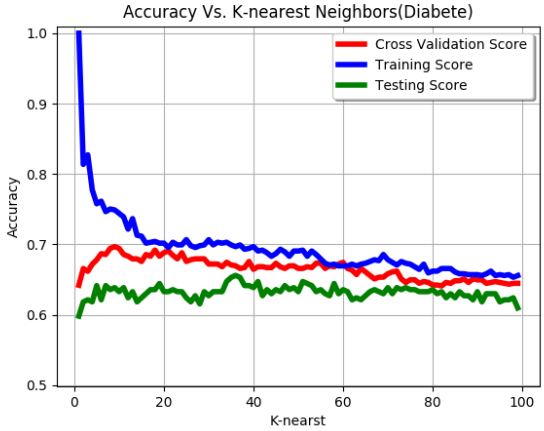


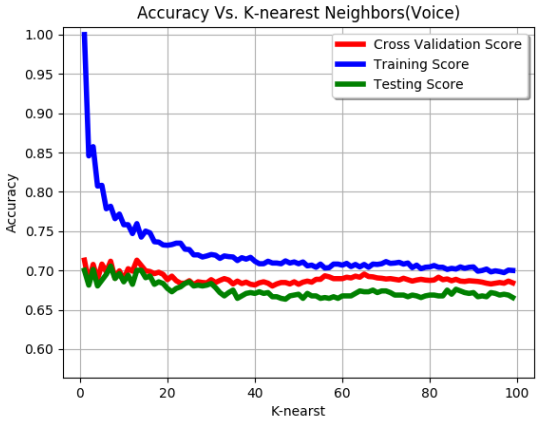
Adaptive boosting:



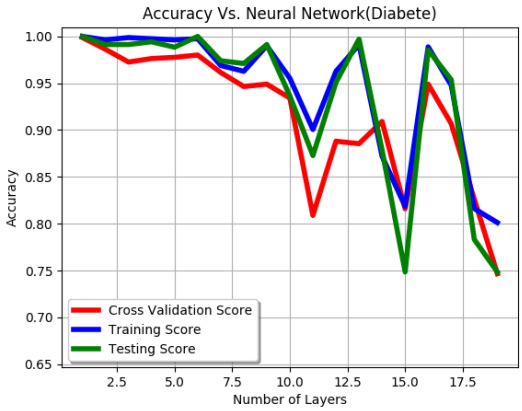


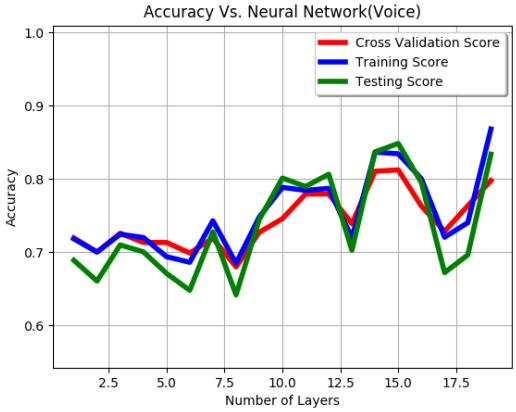
k-nearest neighbors:



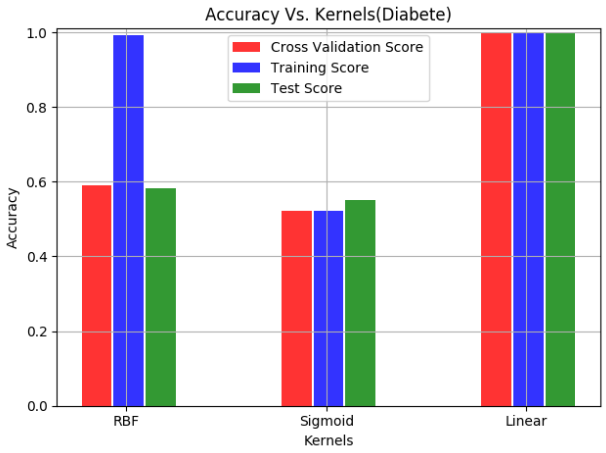


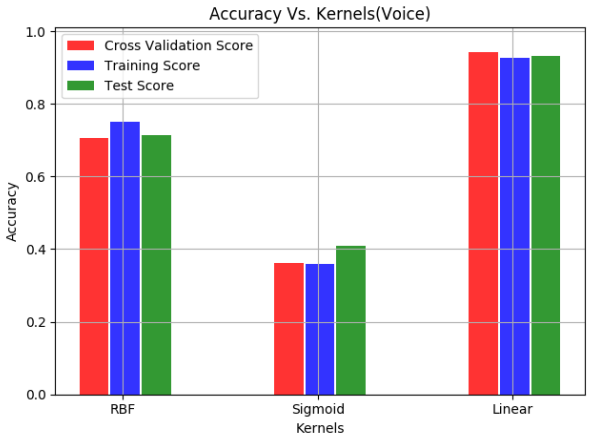
Neural networks:





Support vector machines:





Learning curve: