Applied Machine Learning

Term Project Report

Shweta Bhartia, sbhartia@umail.iu.edu

Pramod Sripada, ksripada@umail.iu.edu

The project involves analyzing two classification datasets, both the datasets have been taken from UC Irvine Machine Learning repository.

Letter Recognition Dataset: The letter recognition dataset contains 20,000 records that are extracted raster scan images of the English letters. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a dataset of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes which were then scaled to fit into a range of integer values from 0 through 15.

The objective is to classify each of the given data attributes as one of the 26 capital letters of the English alphabet. The dataset consists of 20,000 records, the class distribution is approximately equal among the 26 alphabets in the English language. There are no missing values in the dataset. The training of the classifiers has been done with the first 16,000 records and the rest 4,000 records have been used as the testing data.

Adult Dataset: The adult dataset contains 48,842 records from the 1994 census database. The dataset contains various features of an adult such as its personal information, demographic information etc.

The dataset contains a mixture of categorical and integer attributes.7% of the records in the dataset contains missing values. For categorical attributes, the missing values have been filled by the mode of the attribute, while for numerical attributes, the missing values have been filled by the mean of the attribute. The objective of the analysis is to classify whether a person makes over 50K dollars a year.

Decision Trees: Decision trees are a non-parametric supervised learning algorithm that is used for classification of datasets. The goal of a Decision tree classifier is to create a model that predicts the value of a target variable by learning simple decision rules inferred from data features.

Decision trees are simple to understand and infer. Decision trees can handle both categorical and numerical attributes. Decision trees can learn complex models that memorize the training data completely or can also build naïve models that don't capture the training data. A good decision tree model should never be too simple or too complex. Decision trees are unstable learners in the sense that subtle changes in data can change the whole decision tree classifier.

Random Forest Classifiers: Random forest classifier is an ensemble learning model that is built using decision trees. Each tree in the ensemble is built from a sample drawn with the replacement of the training set. Also during the construction of the tree, when the data has to be split, the split is chosen from a random subset of features. Due to the randomness of the model even though the bias of an individual decision trees increases since we average out the variance of a model as whole decreases.

Support Vector Machines: Support Vector Machines is a supervised learning method used for classification. SVM classifier is formally defined by the separating hyperplane. The algorithm outputs a line or hyperplane that gives the largest minimum distance to the training examples. The training examples close to the margin are known as Support Vectors.

Since most of the datasets won't be linearly separable in the current space, SVM raises the dimensionality of the examples. Since raising the dimensions is a costly task, the kernel function is specified which is a function that implicitly computes the product of two points in the higher dimension without the need of projecting the data point in a higher dimension. There are various kernels such as linear kernel, polynomial kernel, RBF kernel etc.

K-Nearest-Neighbors Classifier: KNN classifier is a supervised learning method used for classification. The principal behind nearest neighbors classifier is to find a predefined set (K) of training examples closest to a new data point and predict the label of the class from the label from the K nearest neighbors. The Euclidean distance measure is generally used in calculating the distance between the data points. K-Nearest-Neighbors classifier suffers from high bias and high variance

AdaBoost Classifier: Adaboost classifier is an ensemble classifier that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. In the initial iterations Adaboost classifier reduces bias and in the later iterations, it reduces variance.

Analysis of Adult Dataset

The adult dataset contains 48,842 records and 17 features. The dataset contains missing values in 7% of the dataset. The missing data for the categorical columns has been filled by the mode of the values and for numerical columns, it has been filled by the mean of the values. The objective is to find whether an adult's income will be greater than 50,000 dollars or less than 50,000 dollars.

5 classification algorithms- Decision Trees, Random Forest, K-Nearest-Neighbor, AdaBoost Classifier, and Support Vector Machines have been applied on the adult dataset. The class distribution in the dataset is 23.93% for the >50K and 76.07% for <=50K. Since the data is not evenly distributed Area under the Curve metric is used for evaluating the classifiers.

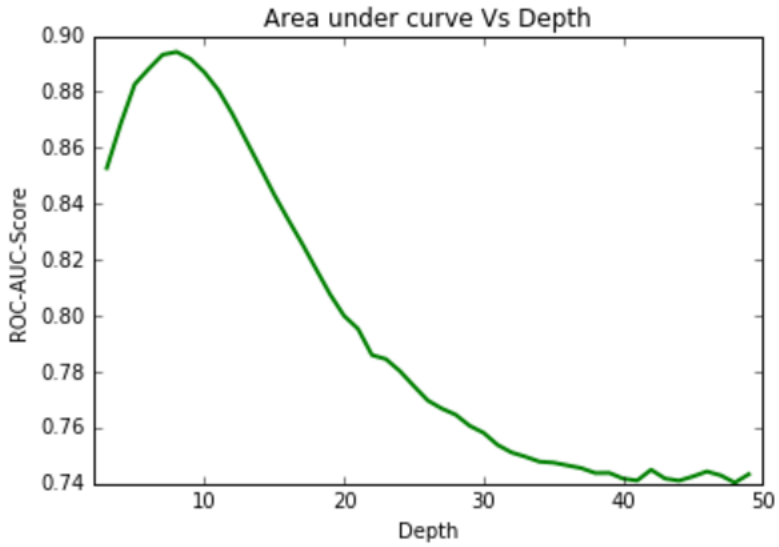
Cross-validation has been performed to tune the parameters particular to a classifier. Once the best parameters have been tuned the model has been applied to the test dataset.

Decision Trees:

Decision trees under fit when the depth of the tree is very less and over fit when the depth of the tree is huge. Also the splitting criterion and the minimum number of samples required to split effect the area under the curve.

The tuned parameters are:

- Depth: Starting from depth 1 the depth has been increased, the area under the curve has steadily increased as the tree is learning, but after depth 8 the AUC has decreased as the decision tree over fitted the data.
- Splitting criterion: Gini performed better than information gain



Applying the model to the testing data with the tuned parameters gives a ROC AUC of 0.75

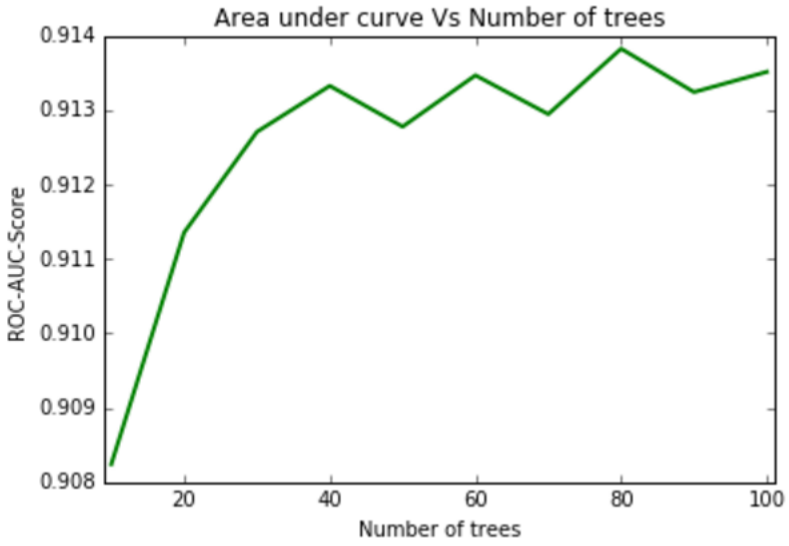
Random Forest Classifier:

Random Forest Classifiers tend to show more bias if the number of ensemble classifiers is less and overfits if the number of ensemble classifiers is more. The splitting criterion if it is

The tuned parameters:

- A number of trees: The number of trees in a random forest classifier, the best number of trees performing are 80. The accuracy has increased as the number of trees have increased from 10 to 80 and the accuracy seems to slightly decline after increasing the number of trees as the model is slightly over fitting.
- Splitting criterion: Entropy as a splitting criterion has performed better than the gini index.

- Max depth of the trees: The max depth is set to 14 as the accuracy being is produced by the trees is maximal.



Applying the model to the testing data with the tuned parameters gives a ROC AUC of 0.913

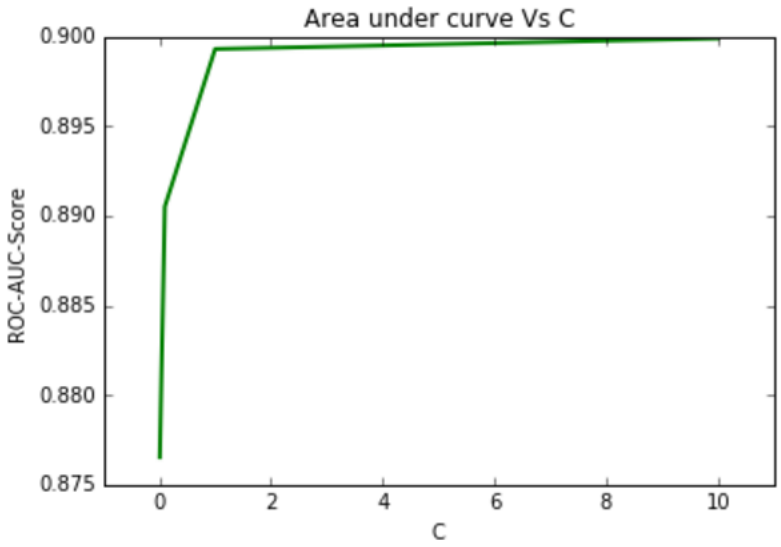
Support Vector Machines:

Support Vector Machines depends on the type of kernel function used. The different kernels available are linear, polynomial kernel, RBF and sigmoid kernels. The parameter C tells how much SVM wants to avoid the misclassification of training examples. Higher the values of C, SVM tends to under fit as the margin is larger. Lower the values of C, SVM tends to overfit. Depending upon the data separability, various kernels work best.

Parameters tuned:

C: The C values have been taken as 0.1,1,10. The highest accuracy has been produced with a C value of 10 it

Kernel: The kernels that have been tested are Linear, Polynomial, and RBF. The polynomial kernel is providing bad results since the data is linearly separable. The linear kernel has a much better performance than the RBF kernel.



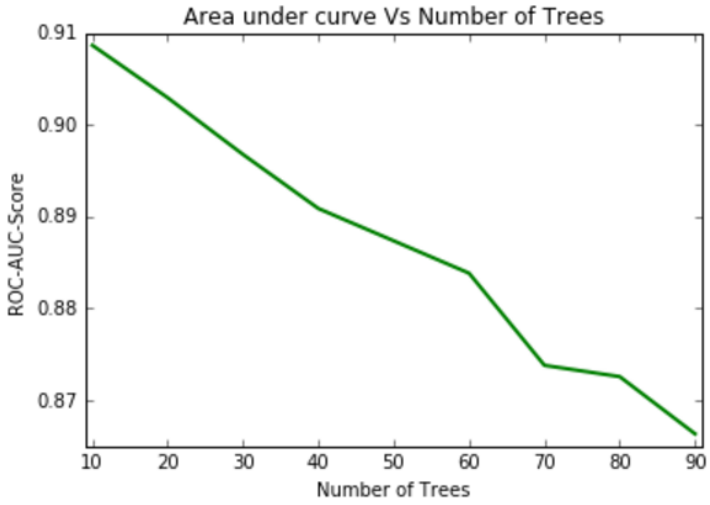
Applying the model to the testing data with the tuned parameters gives a ROC AUC of 0.8998

Ada-boost classifier:

Ada-boost classifier is an ensemble technique, we have implemented with a base learning algorithm of a decision tree. The important parameters are the number of trees in the learning algorithm and the depth of the underlying decision trees.

Parameters tuned:

- A number of trees: The number of trees has been taken in multiples of 10 and the best number of trees is 10, after that the accuracy decreases.
- The base depth of the decision tree: With various depth and parameters tried for the base decision tree classifier the depth of 5 was providing a simple tree that was boosting up the performance of the ensemble learner.



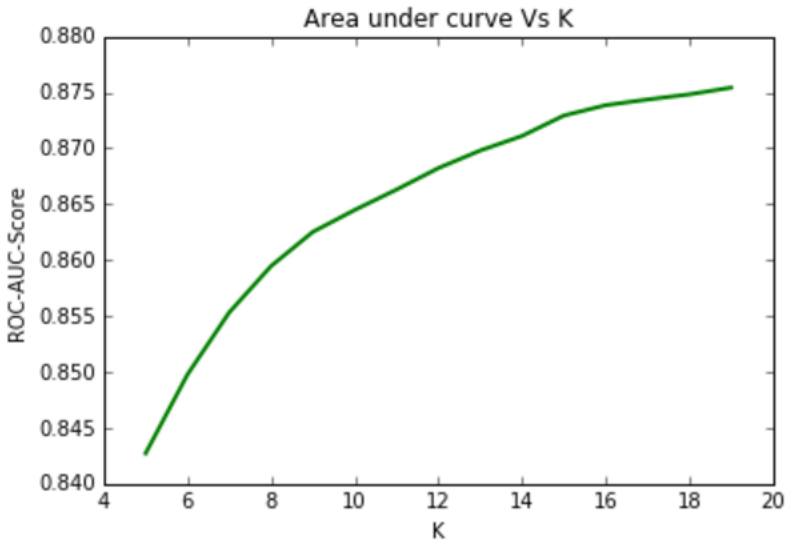
Applying the model to the testing data with the tuned parameters gives a ROC AUC of 0.908

K-Nearest-Neighbors:

The algorithm depends upon the number of nearest neighbors chosen. If the value of k is less the data tends to under fit and produce less accuracy, but as the k-value increase the accuracies increase until a point where the data tends to over fit and accuracies decrease

Parameters tuned:

- K: The number of neighbors used for classifying a point, the accuracy increase till k = 24 and goes for a downward for k values after 3 as the data tends to over fit.



Applying the model to the testing data with the tuned parameters gives a ROC AUC of 0.875

Analysis of Letter Recognition dataset:

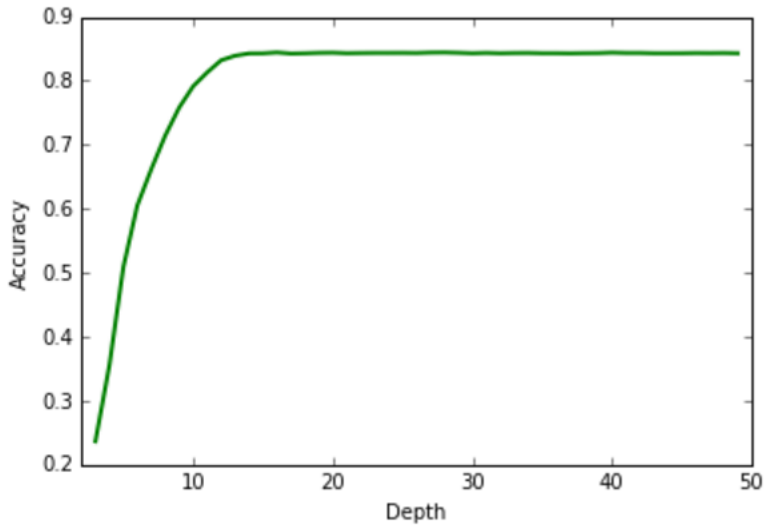
The Letter recognition dataset consists of 20,000 records and 16 attributes. There are no missing values in the dataset. The objective is to classify the letters in English alphabet.

5 classification algorithms- Decision Trees, Random Forest, K-Nearest-Neighbor, AdaBoost Classifier, and Support Vector Machines have been applied on the adult dataset. Accuracy has been used as a metric for evaluation. Cross-validation has been performed to tune the parameters particular to a classifier. Once the best parameters have been tuned the model has been applied to the test dataset.

Decision Trees:

The tuned parameters are:

- Depth: Starting from depth 1 the depth has been increased, the accuracy has steadily increased as the tree is learning, but after depth 18 the accuracy has slightly decreased as the decision tree over fitted the data.
- Splitting criterion: Entropy as a splitting criterion has performed better than Gini index.
- Minimum leaves at the node: The number of samples required to be at leaf node. With the minimum leaves at node at 5, the accuracy has improved by 0.3%.



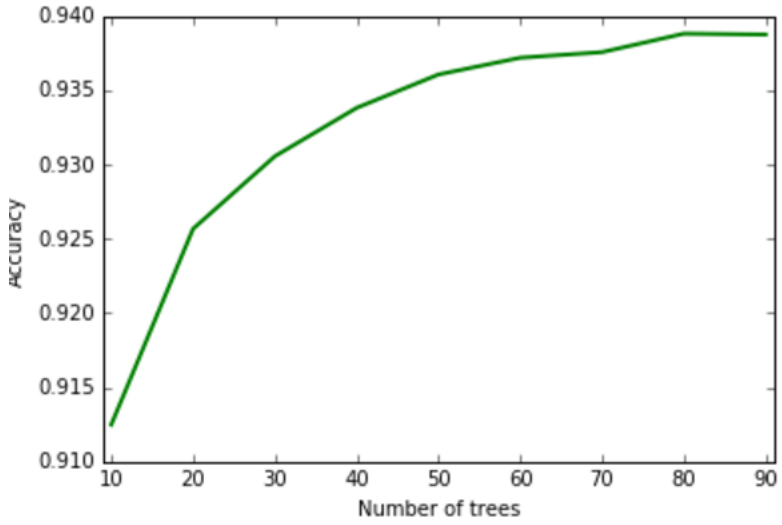
The final accuracy of the test data for a depth 18 decision tree with the tuned parameter settings is 83.8%

Random Forest Classifier:

The tuned parameters:

- A number of trees: The number of trees in a random forest classifier, the best number of trees performing are 80. The accuracy has increased as the number of trees have increased from 10 to 80 and the accuracy seems to slightly decline after increasing the number of trees as the model is slightly over fitting.
- Splitting criterion: Entropy as a splitting criterion has performed better than the gini index.
- Bootstrap samples: While building the tree bootstrap samples are used in building the tree.
- Max depth of the trees: The max depth is set to 18 as the accuracy being is produced by the trees is maximal.

The accuracy produced by the model on the test dataset with the optimal parameters is 93.775%.



Support Vector Machines:

Parameters tuned:

C: The C values have been taken as 0.1,1,10. The highest accuracy has been produced with a C value of 10 it

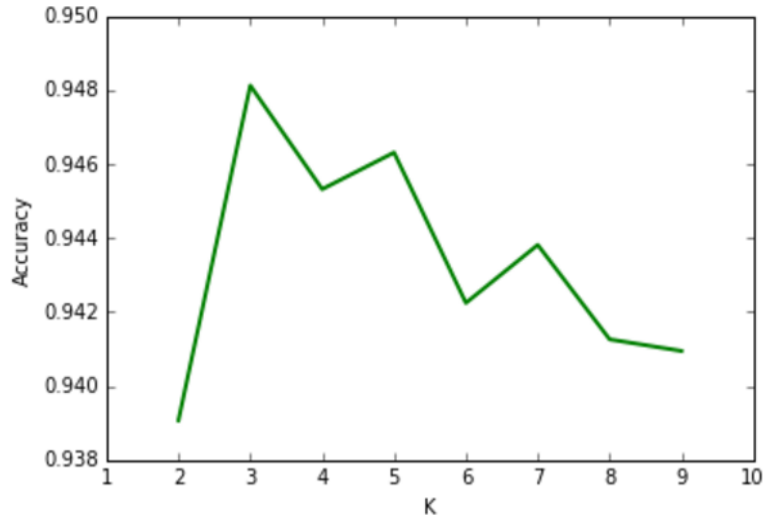
Kernel: The kernels that have been tested are Linear, Polynomial and RBF. The polynomial kernel is providing bad results since the data is linearly separable. The linear kernel has a much better performance than the RBF kernel

The SVM model built with the best-tuned parameters when applied to the test data has produced an accuracy of 83.5%.

K-Nearest-Neighbors:

Parameters tuned:

- K: The number of neighbors used for classifying a point, the accuracy increase till k = 3 and goes for a downward for k values after 3 as the data tends to over fit.

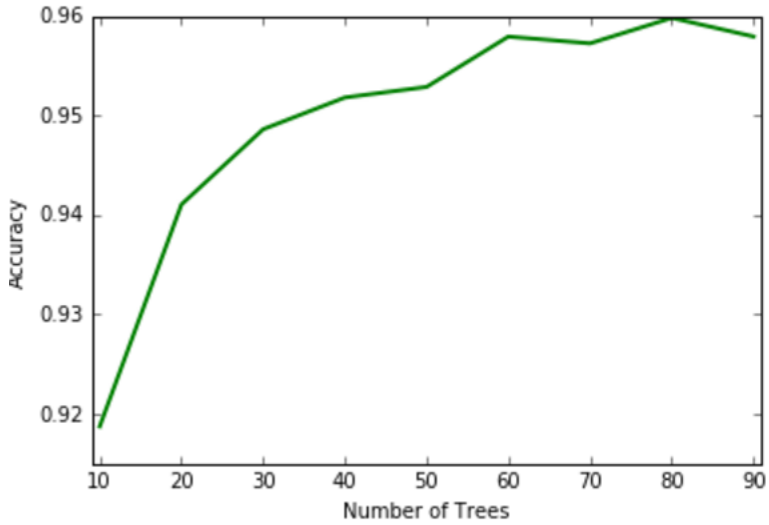


The accuracy provided by k value of 3 on the dataset is 94.925%

Ada boost classifier:

Parameters tuned:

- Number of trees: The number of trees have been taken in multiples of 10 and the best number of trees is 80, after that the accuracy decreases.
- Base depth of the decision tree: With various depth and parameters tried for the base decision tree classifier the depth of 10 was providing a simple tree that was boosting up the performance of the ensemble learner.



The final accuracy on the test dataset with the tuned parameters of 80 base estimators and a base decision tree of depth 10 is 95.65%.

Conclusion:

On both the datasets Ensemble learners outperform simple models. The two ensemble learners which we implemented – Random Forest and Ada boost classifier perform better than a single classifier.

In the Letter recognition dataset kMeans provided higher accuracy than SVM and Decision trees since the data is not in high dimensions.

In the Adult dataset SVM performs the data, since the number of dimensions are very high and SVM is able to draw a line in higher dimension.