Aydan Alyanak
I526 Applied Machine Learning | Programming Assignment 2
aalyanak@indiana.edu

Programming Assignment #2

For this programming/Weka assignment, I compared the performance of

- SVM,
- Random Forest and
- Adaboost

on the Adult data set from the UCI repository (http://archive.ics.uci.edu/ml/datasets/Adult).

The prediction task associated with this data set is to predict whether or not a person makes more than $50K a year using census data.

1) Downloading the data via Python and Creating Weka Files

First, I used "requests" library of Python to connect to the website and download it, "pandas" to manipulate the data and "csv" library to read and write csv & arff files.

Training dataset = 'http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data'

Test dataset = 'http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test'

The features:

AGE_GROUP (Discrete)	This is a categorized value, from a continuous age value to discrete. Values: {'< 25', '25-35', '35-45', '45-55', '55-70', '> 70'}
WORK_CLASS (Discrete)	This feature used as is. **NOTE: There were unknown (?) values, 70% of the known values are classified as "Private", therefore the unknown work_classes are also considered as "Private".**
FNLWGT (Continuous)	Final weight (This feature used as is)
EDUCATION (Discrete)	This feature used as is.
EDUCATION_NUM (Discrete)	Classes of the education (This feature used as is)
MARITAL_STATUS (Discrete)	This feature used as is.
OCCUPATION (Discrete)	This feature used as is. **NOTE: There were unknown (?) values, the unknown work_classes are considered as "Other-service".**
RELATIONSHIP (Discrete)	This feature used as is.
RACE (Discrete)	This feature used as is.
SEX (Discrete)	This feature used as is.
CAPITAL_GAIN (Continuous)	This feature used as is.
CAPITAL_LOSS	This feature used as is.

(Continuous)	
HOURS_PER_WEEK (Continuous)	This feature used as is.
CONTINENT (Discrete)	The native countries are remapped as continents. **NOTE: There were unknown (?) values, 90% of the known values are classified as "United-States", therefore the unknown work_classes are also considered as "United-States".** VALUES: {'Asia', 'Europe', 'North America', 'South America'}
INCOME (Class)	The class value.

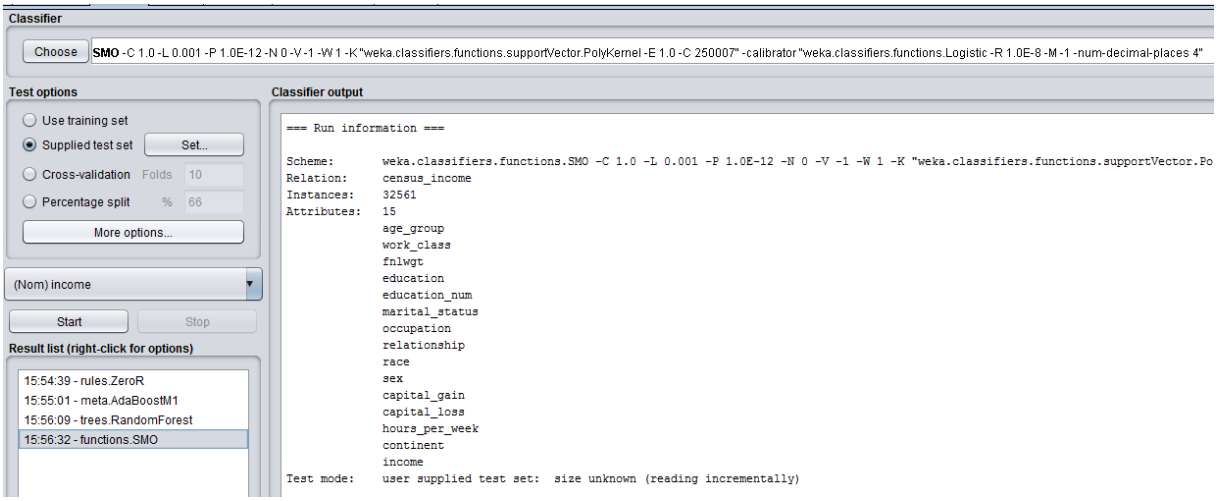
2 arff files are created by Python, and ready to use.

The rest of the analysis will be held in Weka.

2) Weka Analysis

a) SVM:

The method:



The model is learned on training set, and evaluated on test set.

Correct classification rate = 85.68%
Incorrect classification rate = 14.32%

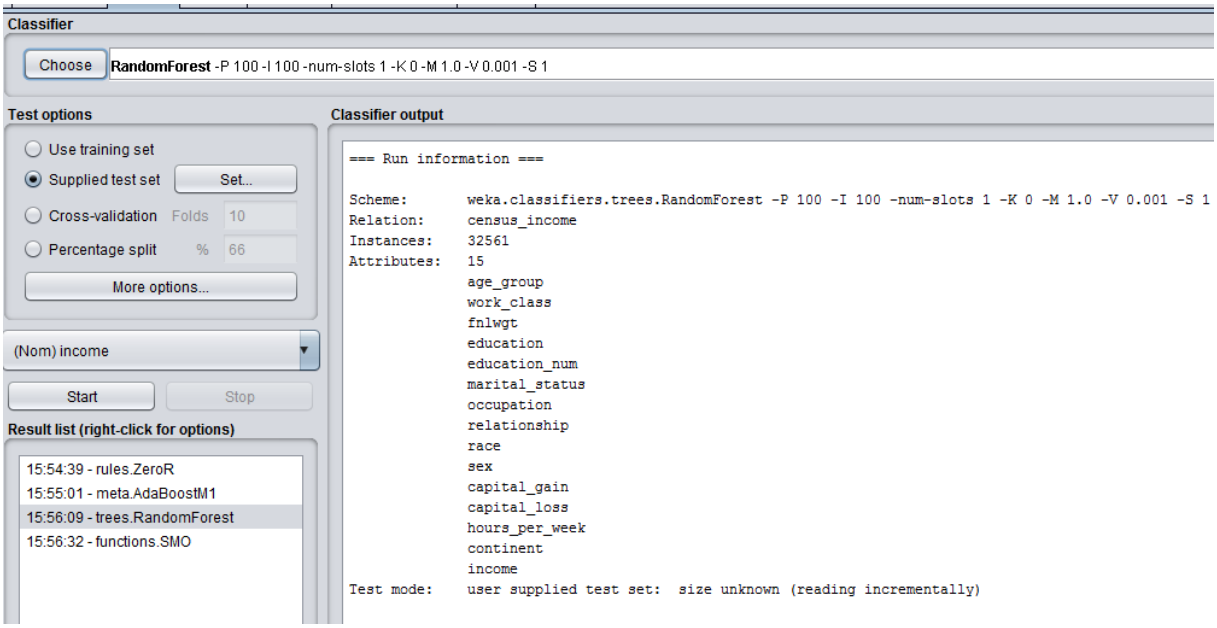
Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,938	0,407	0,882	0,938	0,909	0,579	0,766	0,875	<=50K
	0,593	0,062	0,748	0,593	0,662	0,579	0,766	0,540	>50K
Weighted Avg.	0,857	0,325	0,850	0,857	0,851	0,579	0,766	0,796	

CONFUSION MATRIX		
a	b	classified
11668	767	a = <=50K
1564	2282	b >50K

b) Random Forest:

The method:



The model is learned on training set, and evaluated on test set.

Correct classification rate = 83.6%
Incorrect classification rate = 16.4%

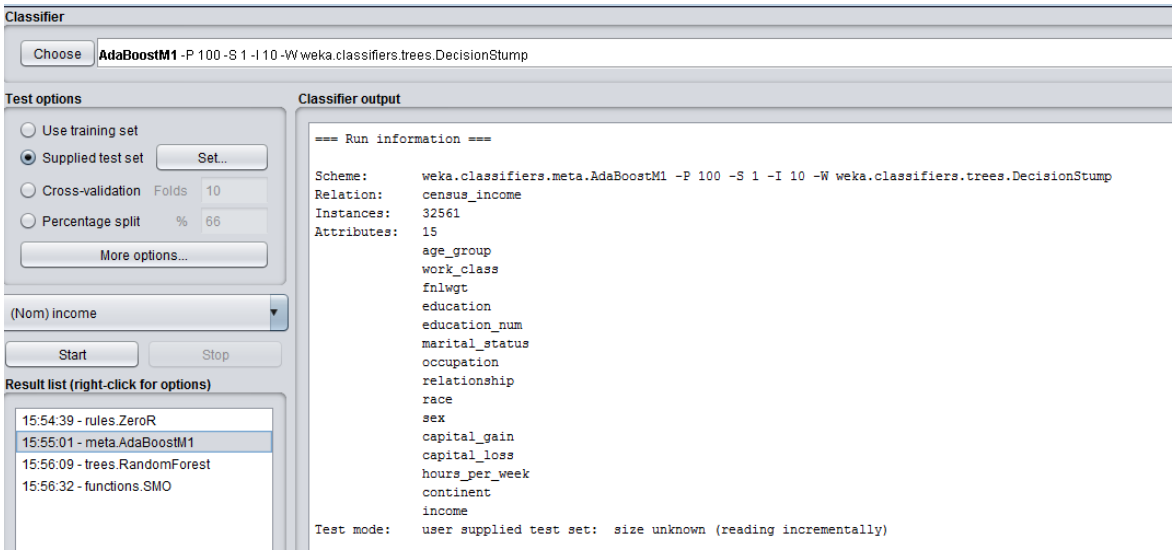
Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,906	0,391	0,882	0,906	0,894	0,532	0,879	0,955	<=50K
	0,609	0,094	0,668	0,609	0,637	0,532	0,879	0,710	>50K
Weighted Avg.	0,836	0,321	0,832	0,836	0,833	0,532	0,879	0,897	

CONFUSION MATRIX		
a	b	classified
11268	1167	a = <=50K
1503	2343	b >50K

c) Adaboost

The method:



The model is learned on training set, and evaluated on test set.

Correct classification rate = 82.4%
Incorrect classification rate = 17.6%

Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,978	0,673	0,824	0,978	0,895	0,444	0,866	0,947	<=50K
	0,327	0,022	0,822	0,327	0,467	0,444	0,866	0,646	>50K
Weighted Avg.	0,824	0,520	0,824	0,824	0,794	0,444	0,866	0,876	

CONFUSION MATRIX		
a	b	classified
12163	272	a = <=50K
2590	1256	b >50K

3) Results and Comparison

The previous assignment (Programming Assignment #1) also had the same training & test set with different evaluation methods. I preferred to include all 7 methods into the comparison, the new 3 methods for assignment 2 are highlighted:

a) By Correct Classification (TP&TN Rate) and Precision:

	Correct Classification	Incorrect Classification	ROC Area	TP Rate	FP Rate	Precision
Decision Tree	85.8%	14.2%	0,867	0,858	0,329	0,851
Logistic Regression	85.6%	14.4%	0,908	0,856	0,317	0,850
SVM	**85.6%**	**14.3%**	**0,766**	**0,938**	**0,407**	**0,882**
Random Forest	**83.6%**	**16.4%**	**0,879**	**0,906**	**0,391**	**0,882**
Naïve Bayes	83.2%	16.8%	0,892	0,832	0,383	0,822
k-Nearest Neighbors	83.2%	16.7%	0,865	0,832	0,349	0,825
Adaboost	**82.4%**	**17.6%**	**0,866**	**0,978**	**0,673**	**0,824**

Using accuracy as a classifier, the most successful method in this dataset is “Decision Tree”. It rated as a 85.8% correct classification, and 0.851 precision.

Followed by logistic regression, almost has the same results.

The new methods for this assignment (SVM and Random Forest) have average results.

Adaboost has the lowest performance on accuracy.

b) By ROC Area:

The graph at right shows three ROC curves representing excellent, good, and worthless tests plotted on the same graph. The accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question. Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of .5 represents a worthless test. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

	Correct Classification	Incorrect Classification	ROC Area	TP Rate	FP Rate	Precision
Logistic Regression	85.6%	14.4%	0,908	0,856	0,317	0,850
Naïve Bayes	83.2%	16.8%	0,892	0,832	0,383	0,822
Random Forest	**83.6%**	**16.4%**	**0,879**	**0,906**	**0,391**	**0,882**
Decision Tree	85.8%	14.2%	0,867	0,858	0,329	0,851
Adaboost	**82.4%**	**17.6%**	**0,866**	**0,978**	**0,673**	**0,824**
k-Nearest Neighbors	83.2%	16.7%	0,865	0,832	0,349	0,825
SVM	**85.6%**	**14.3%**	**0,766**	**0,938**	**0,407**	**0,882**

By ROC area, “Logistic Regression” has an “A” (excellent) performance, Naïve Bayes has a B (good) but very close to “A” performance.

The new methods, Random Forest and Adaboost both have “B” performance, which is also not comparable with Logistic Regression.

SVM has the lowest performance (Fair – C) for ROC area.

c) Results:

If we only evaluate the 3 methods which are used for assignment #2:

The highest performance by Accuracy: “SVM”

The lowest performance by Accuracy: “Adaboost”

While SVM has a very high performance (85% - the highest score of all methods) on overall accuracy, it has the lowest ROC area. Which is possible, the area under the curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example. It measures the classifiers skill in ranking a set of patterns according to the degree to which they belong to the positive class, but without actually assigning patterns to classes.

The overall accuracy also depends on the ability of the classifier to rank patterns, but also on its ability to select a threshold in the ranking used to assign patterns to the positive class if above the threshold and to the negative class if below.