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

Programming Assignment

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

- Logistic Regression,
- Naïve Bayes,
- Decision Tree and
- Nearest Neighbor

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'

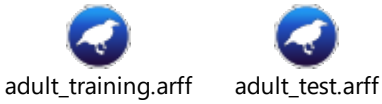
Test dataset = 'http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test'

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 (Continuous)	This feature used as is.
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.

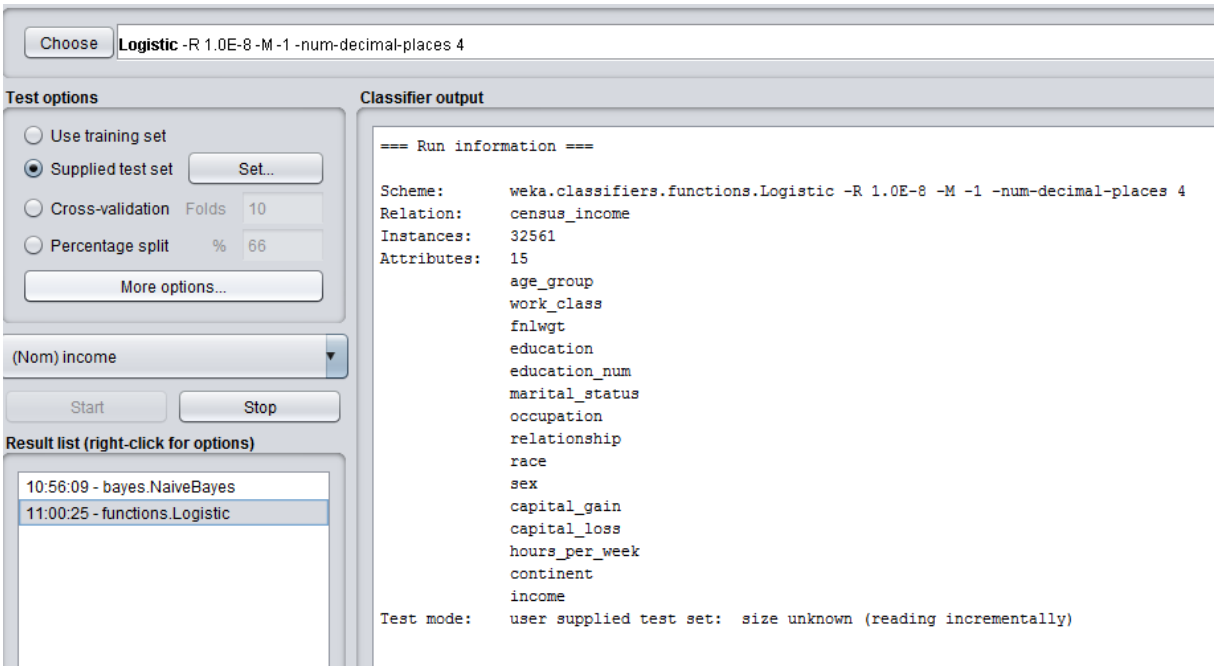


The rest of the analysis will be held in Weka.

2) Weka Analysis

a) Logistic Regression:

The method:



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

Correct classification rate = 85.60%
Incorrect classification rate = 14.39%

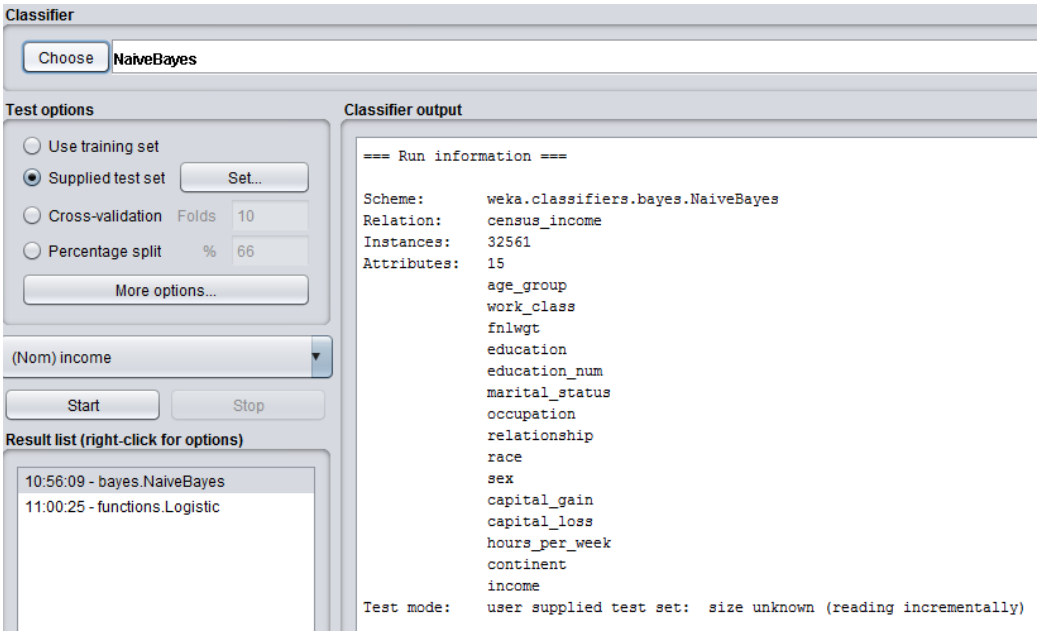
Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,933	0,394	0,884	0,933	0,908	0,579	0,908	0,968	<=50K
	0,606	0,067	0,738	0,606	0,665	0,579	0,908	0,770	>50K
Weighted Avg.	0,856	0,317	0,850	0,856	0,851	0,579	0,908	0,921	

CONFUSION MATRIX		
a	b	classified
11608	827	a = <=50K
1516	2330	b >50K

b) Naïve Bayes:

The method:



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

Correct classification rate = 83.21%
Incorrect classification rate = 16.78%

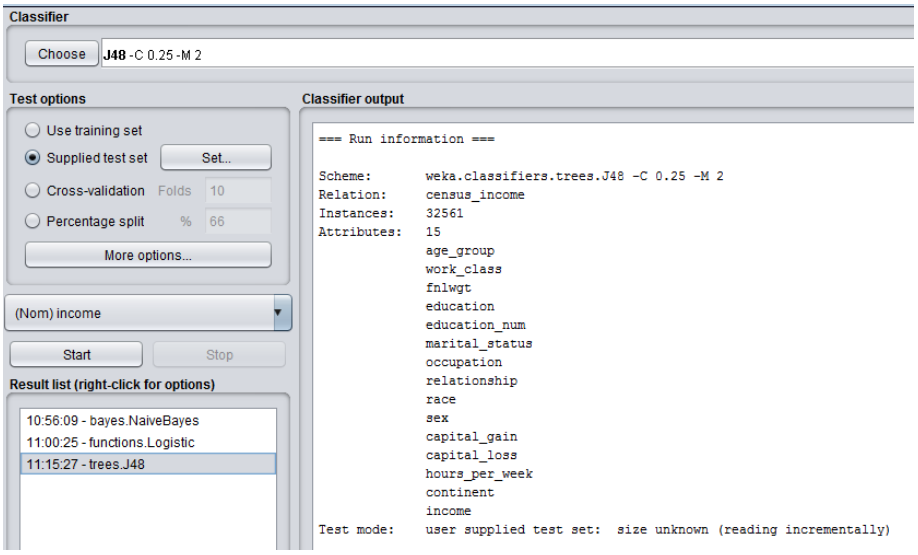
Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,929	0,480	0,862	0,929	0,894	0,499	0,892	0,965	<=50K
	0,520	0,071	0,693	0,520	0,594	0,499	0,892	0,719	>50K
Weighted Avg.	0,832	0,383	0,822	0,832	0,823	0,499	0,892	0,906	

CONFUSION MATRIX		
a	b	classified
11548	887	a = <=50K
1845	2001	b >50K

c) Decision Tree

The method:



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

Correct classification rate = 85.78%
Incorrect classification rate = 14.21%

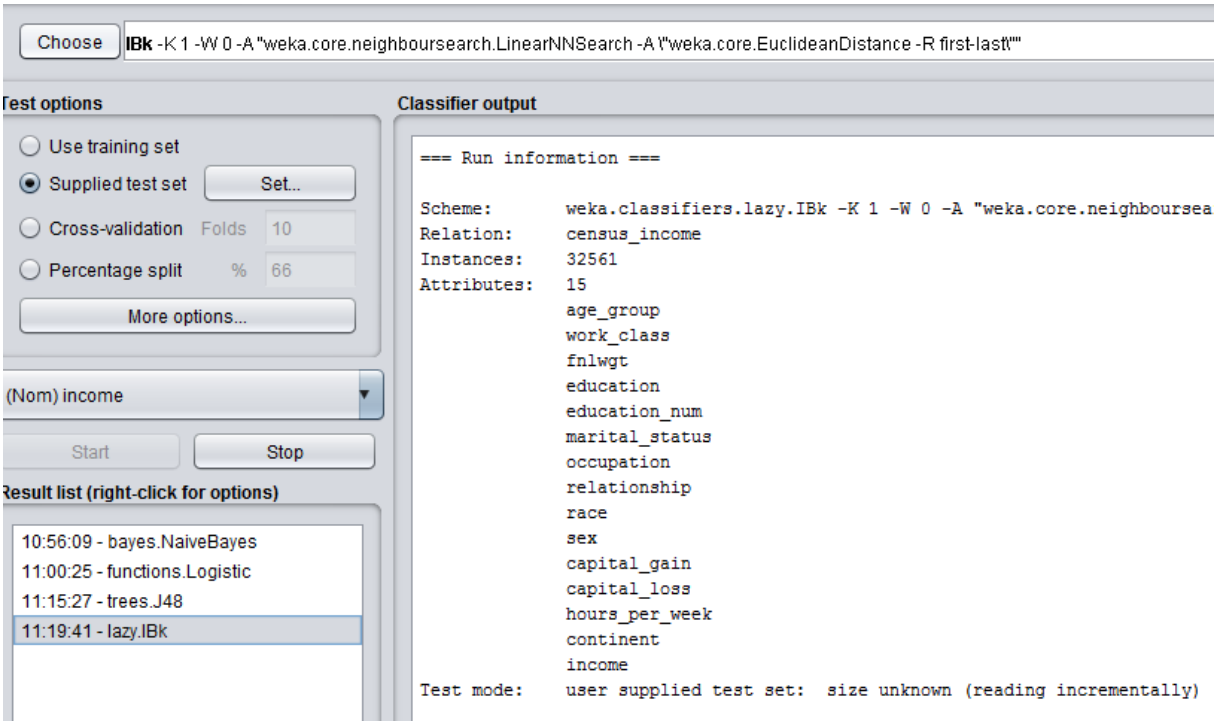
Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,942	0,413	0,881	0,942	0,910	0,580	0,867	0,935	<=50K
	0,587	0,058	0,757	0,587	0,661	0,580	0,867	0,722	>50K
Weighted Avg.	0,858	0,329	0,851	0,858	0,851	0,580	0,867	0,885	

CONFUSION MATRIX		
a	b	classified
11710	725	a = <=50K
1589	2257	b >50K

d) K-Nearest Neighbors:

The method (K=1 by default):



In this method, we can change the value of k to evaluate the performance of the model.

	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11
Correct classification	79%	80.9%	81.8%	82.0%	82.5%	82.8%	82.9%	83.0%	83.2%	83.2%	83.2%
Incorrect classification	21%	19.1%	18.1%	18.0%	17.5%	17.2%	17.0%	16.9%	16.7%	16.7%	16.7%

After k=9, the correct classification rate gets monotone. Therefore, I'll evaluate results by choosing k=9.

Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,913	0,430	0,873	0,913	0,893	0,513	0,865	0,940	<=50K
k=9	0,570	0,087	0,670	0,570	0,616	0,513	0,865	0,647	>50K
Weighted Avg.	0,832	0,349	0,825	0,832	0,827	0,513	0,865	0,871	

CONFUSION MATRIX		
a	b	classified
11355	1080	a = <=50K
1652	2194	b >50K

3) Results and Comparison

The compared results of 4 methods:

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
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

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.

Naïve Bayes and kNN have low 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
Decision Tree	85.8%	14.2%	0,867	0,858	0,329	0,851
k-Nearest Neighbors	83.2%	16.7%	0,865	0,832	0,349	0,825

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

Decision Tree and kNN both have “B” performance, but slightly lower than Naïve Bayes.

c) Results:

The highest performance by Accuracy and ROC: “Logistic Regression”

The lowest performance by Accuracy and ROC: “k-Nearest Neighbors”