Assignment 2 - Group38

Tutors: Kelvin, Anthony, Harrison, Aswani
Group members: Heng Guo (hguo2558)
Shen Wang (swan2892)
Haoxing Wu (hawu0737)

Abstract

In this assignment, choosing UCI Adult Data Set which includes 32561 people's background information and their annual salary as our data set. The aim of this task is using related background information (such as education level and occupation) to predict if a person's annual salary exceeds 50,000 dollars. This study could give the government a better understanding of social conditions and provides a basis of the decision in education and economy fields. It is also beneficial to enterprises and individuals. Such as select target customers and choice of occupation.

Trying PCA, standardization, several means to deal with the missing values and feature selection methods in pre-processing step to make the data more convenient for processing. For acquire more accurate results efficiently, different algorithm methods (Naïve Bayes, Logistic Regression, K-Nearest neighbours, Random Forest, Adaptive Boosting, Support vector machines) as classifiers were used and compared. The result shows that Adaptive Boosting (AdaBoost) algorithm has the highest accuracy 85.63% in 14.98 seconds of running time which is the best algorithm for this task. Naïve Bayes is the fastest algorithm, only spend 0.16 seconds for processing and got a 79.36% accuracy.

Contents

1. Introduction

1.1. Data set

Name: Adult Data Set
Download from: https://archive.ics.uci.edu/ml/datasets/adult
The dataset was extracted from the 1994 Census database by Barry Becker. There are 32561 records and each record contains 14 attributes. Attributes include two type of data which are categorical and integer. The characteristic of data set is multivariate, and some values are unknown.

Table 1.1: result of each algorithm

Age	continuous
Workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked, fnlwgt: continuous
Education	Bachelors, Some-college, 11th, HS-grad, Pro-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Education-num	continuous
Marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Female, Male
Capital-gain	Continuous
Capital-loss	Continuous
Hours-per-week	Continuous
Native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
Income	>50K, <=50K

1.2. Aim

The main purpose of this assignment is providing a solution for a real problem by using machine learning and data mining methods. Specifically, the task is to determine if a person's annual salary greater than 50 thousand dollars

based on background information. The reason for choosing this dataset is that this problem has a close connection with real life. For the government, this prediction methods could be used in the improvement of education and making economic plans and provide a deep understanding of social conditions. It is also could help individuals for major choosing and occupation selection.
This assignment is an opportunity for us to learn and apply different machine learning and data mining methods. We will get a better understanding and find the weakness of knowledge in this process.

2. Previous work

There are a plenty of articles discussing with the Adult data set. In *Learning and evaluating classifiers under sample selection bias.*, they applied Naive Bayes, logistic regression, C4.5 and SVMLight to verify the effects of sample selection bias. In *Saharon Rosset. Model selection via the AUC.*, they used only the first ten variables in the data set to make a large scale experiment feasible. In *An Empirical Evaluation of Supervised Learning for ROC Area.*, they attempted SVMs, Boosted stumps and plain decision tree, and found that the boosted stumps are best on the Adult data. In *A Fast Dual Algorithm for Kernel Logistic Regression.*, they tested the cost of SMO algorithm compared with BFGS algorithm. After reading these literature and analyzing the data, we decided to choose the following algorithms for the classification.

* Naive Bayes: suitable when the dimensionality of the inputs is high
* Logistic Regression: usually taken to apply to a binary dependent variable
* K-nearest neighbors: a non-parametric and lazy learning method.
* SVM: could use in both classification and regression
* Random Forest: an ensemble learning method for classification, regression
* Adaboost: best used to boost the performance of decision trees on binary classification problems.

3. Method

3.1. Naive Bayes

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes models are commonly used as an alternative to decision trees for classification problems. When building a Naïve Bayes classifier, every row in the training dataset that contains at least one NA will be skipped completely. If the test dataset has missing values, then those predictors are omitted in the probability calculation during prediction. Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from P(c), P(x), and $P(x|c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

3.2. Logistic Regression

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

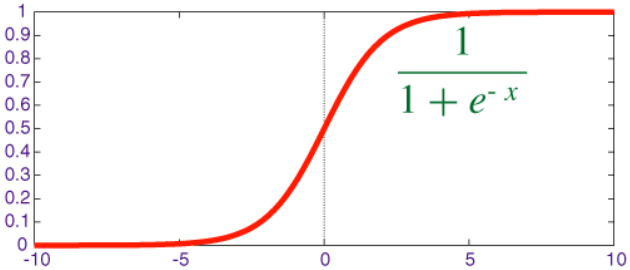

Figure 3.1: sigmoid function

3.3. K-nearest neighbors

K-nearest neighbors algorithm (k-NN) is a non-parametric and lazy learning method. K-NN can be used in classification and regression. The k-NN algorithm is the simplest algorithm in machine learning.
Both for classification and regression, a useful technique can be to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.
The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

3.4. Random Forest

Random forest is a classifier which includes many decision-trees. The number of output types is depended on the mode of individual trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called ”feature bagging”. The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated. Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way.

3.5. SVM

Support vector machines (SVM) could be used in machine learning for both classification and regression analysis. It’s a supervised learning model. In machine learning process, using support vector machine to train dataset provide an efficient way to get a base-line trainer. Support vectors classification (SVC) is one type of SVM, the main function is to classify.

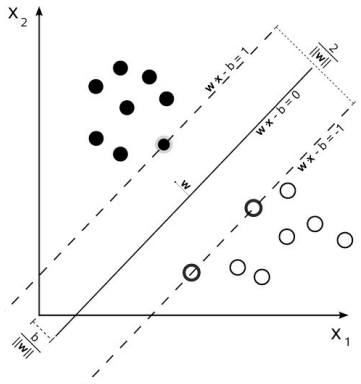

Figure 3.2: SVM

3.6. Adaboost

AdaBoost is best used to boost the performance of decision trees on binary classification problems.AdaBoost was originally called AdaBoost.M1 by the authors of the technique Freund and Schapire. More recently it may be

referred to as discrete AdaBoost because it is used for classification rather than regression.
AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem.
The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level. Because these trees are so short and only contain one decision for classification, they are often called decision stumps.
Each instance in the training dataset is weighted. The initial weight is set to:

$$weight(x_i) = \frac{1}{n}$$

Where x_i is the i'th training instance and n is the number of training instances.

4. Experiments and Discussion

All algorithm using the 10-fold cross-validation to calculate the averaged accuracy and try to get the best performance for each algorithm.
To avoid overfitting, change the n neighbors for K-NN and gamma for SVC to get the appropriate values. Change the K for KNN in range 1 to 50, when the K = 25, the accuracy is highest. Using the validation_curve to calculate the neg_mean_squared_error. When the value of gamma greater than 0.1, the loss of testing increase, therefore, choosing 0.1 to be the gamma for SVC.

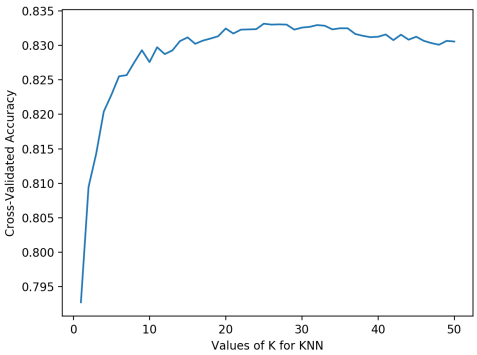

Figure 4.1: KNN

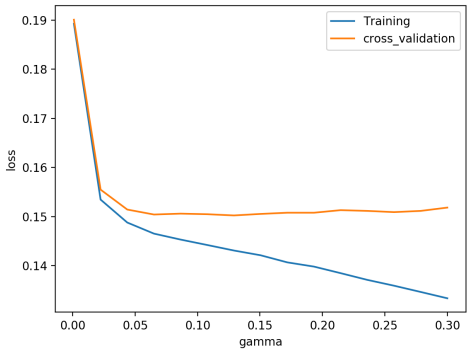

Figure 4.2: SVC

4.1. Pre-processing

4.1.1. Get data

For this assignment, we choose the Adult dataset for the classification. The data is downloaded from https://archive.ics.uc
This data is using 14 features such as age, workclass, sex to predict whether the income exceeds $50K per year.

4.1.2. Manage missing value

After loading the data set, there are some missing value in it. The location of them are as follows.
These missing values are all in three dimensions (Workclass, Occupation, Country), and the amount of missing value is not large for the whole data set(2.96%, 2.97%,0.84%).
There are several ways to deal with the missing values. Using the accuracy of each algorithm to compare different means.

```
Age                963
Workclass          963
fnlwgt               0
Education            0
Education-Num        0
Martial Status       0
Occupation         966
Relationship         0
Race                 0
Sex                  0
Capital Gain         0
Capital Loss         0
Hours per week       0
Country            274
Target               0
dtype: int64
```

Figure 4.3: Missing value number for each feature

Table 4.1: different means for missing value

	Naïve Bayers	K-NN	LR	Random Forest	SVC	Adaboost
remove rows contain ?	0.7975	0.8330	0.8204	0.8534	0.8375	0.8572
keep ? as a feature	0.8036	0.8376	0.8248	0.8566	0.8424	0.8596
Change the ? to mean/most frequent value	0.7975	0.8330	0.8204	0.8522	0.8375	0.8572

The result as table 4.1 shows. Therefore, choosing to regard all missing value as one kind of characteristic in order to get the best performance.

4.1.3. Analyze data

Before using these data, it is very important to analyze them to manage data. Firstly, use several pictures to plot the distribution for each column in the dataset.

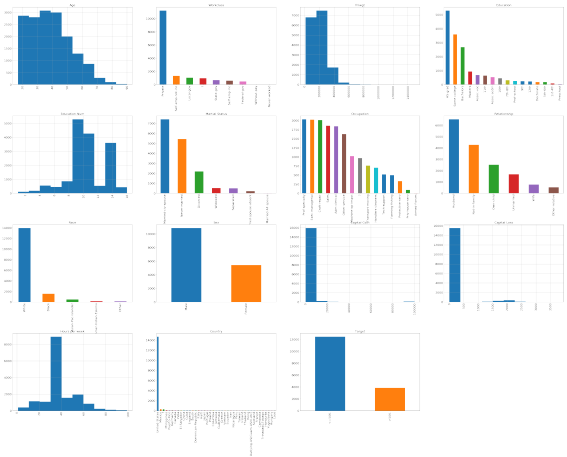

Figure 4.4: Data allocation

There is a connection between education and education number, each education is representing one education

number, such as 13 represents Bachelors, 14 represents Masters. Most of these data are from United states (which is 89.58% of the total data). Most of the data in capital gain(which is 91.69%) and loss(95.33%) is zero. Moreover, there is a relationship between workclass and occupation because workclass indicates the characteristics of the work and occupation means the actual work. Some features have inner link. For instance, sex, martial status and relastionship.

4.1.4. Encode data and Standardize dataset

Since there are many categories likely columns, using the sklearn.preprocessing.LabelEncoder to encode the data. After deleting corresponding columns, we found out that there are significant differences between different columns. For example, in fnlwgt, the data is between 0 to 1484705 which is between 1 to 16 in education number. Therefore, we use scale function to standardize the dataset. After that, the number in each column is in a same scale.

	Age	Workclass	fnlwgt	Education	Education-Num	Martial Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
0	25	4	226802	1	7	4	7	3	2	1	0	0	40	38	0
1	38	4	89814	11	9	2	5	0	4	1	0	0	50	38	0
2	28	2	336951	7	12	2	11	0	4	1	0	0	40	38	1
3	44	4	160323	15	10	2	7	0	2	1	7688	0	40	38	1

	Age	Workclass	fnlwgt	Education	Education-Num	Marital Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country
0	-0.994129	0.085414	0.353474	-2.387116	-1.196864	0.905239	0.097403	0.981172	-1.990286	0.706521	-0.142662	-0.218062	-0.031432	0.286402
1	-0.055417	0.085414	-0.942391	0.188304	-0.417886	-0.418769	-0.374987	-0.902239	0.389812	0.706521	-0.142662	-0.218062	0.769918	0.286402
2	-0.777503	-1.265356	1.395450	-0.841864	0.750582	-0.418769	1.042181	-0.902239	0.389812	0.706521	-0.142662	-0.218062	-0.031432	0.286402
3	0.377835	0.085414	-0.275397	1.218472	-0.028397	-0.418769	0.097403	-0.902239	-1.990286	0.706521	0.871091	-0.218062	-0.031432	0.286402

4.1.5. Decomposition

For decomposition, try two different means: PCA and feature select.

1 PCA

The bars show the accuracy and lines show the runtime of using and not using the PCA. For this task, there is only 14 features and the the runtime is acceptable. To compare different performance of different algorithms, accuracy is more important measure. Therefore, PCA is a not suitable for this task.

2 Feature Select

After analyzing the data in each column, there are 4 columns which is less valuable for classification. The Country, capital gain and loss columns will not affect the classification much because there is a highly concentrated data in each column. The education can be replaced by education number. Therefore, we delete these 4 columns to reduce the number of the dimensions without affecting the result much.

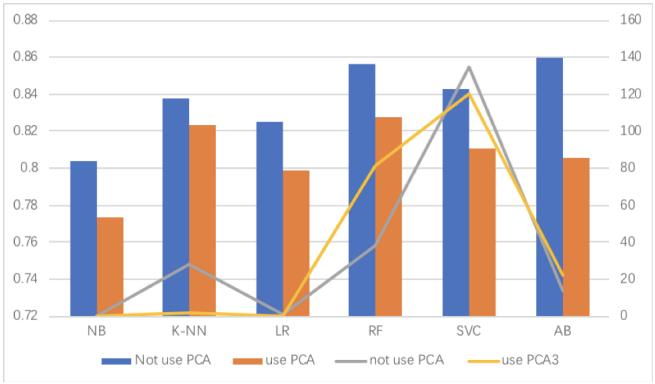

Figure 4.5: result of PCA

4.2. Performance measure

Accuracy, precision, recall and f1 score are important indicator while evaluating machine learning algorithm.

Table 4.2: result of each algorithm

	Accuracy	Precision	Recall	F1 score	Run time
Naïve Bayes	0.7936	0.7595	0.6337	0.6558	0.1584
K-NN	0.8375	0.7929	0.7599	0.7737	16.0681
Logistic Regression	0.8206	0.7709	0.6949	0.7178	0.7745
Random Forest	0.8458	0.8110	0.7713	0.7874	39.3031
SVM	0.8375	0.8087	0.7292	0.7550	127.1243
AdaBoost	0.8563	0.8224	0.7706	0.7906	14.9806

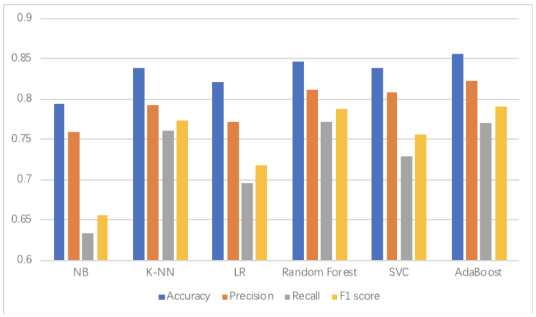

Figure 4.6: Accuracy,precision,recall and F1 scores

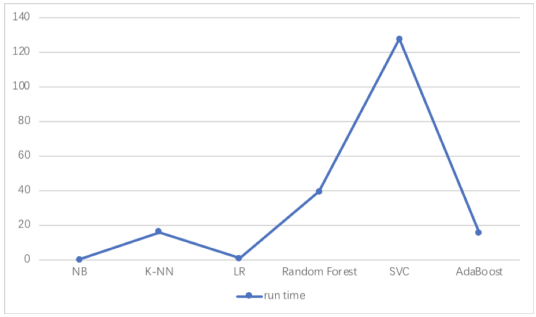

Figure 4.7: runtime

For accuracy, the differences between different algorithms are not too big. The result of all algorithms are between 0.78 and 0.87. The AdaBoost has the highest accuracy and Naïve Bayers has the lowest accuracy. This is because that NB assumes all data are independent but in fact it is not ture. For example there are relationship between ”martial status” and ’sex’. As for running time, there is a huge gap between different algorithms. Naïve Bayers is the fastest one with 0.1434 and the SCV is the slowest one.

ROC curve, as receiver operating characteristic curve, indicates the relationship between true positive rate and false positive rate of a classifier, it is useful to compare the relatively performance between different classifiers.

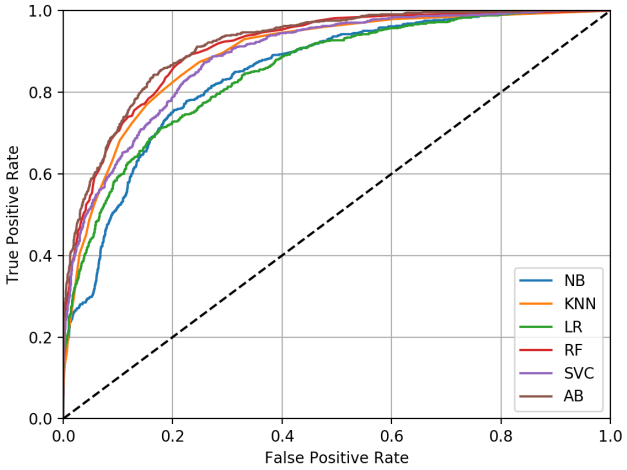

Figure 4.8: ROC curve

The bigger AUC (area under curve) indicates the better classifier. As the figure shows, the AdaBoost is relatively good for all situation. The Logistic regression and Random forest are in the second level but every close to Adaboot. SVC algorithm is in the middle level. The K-NN and Naïve bayers are in third level. The Naïve bayers is a little bit better than K-NN.

5. Conclusion

We have used six different methods to deal with this task and used 10 fold cross-validation to get the results. AdaBoost is the most accuracy algorithm for this task. The main reason is that AdaBoost algorithm using several different weak classifiers for training data, then combine these weak classifiers as a powerful classifier. This algorithm could change the data distribution. It could determine the weight of each sample base on the accuracy of last classification for every sample. The new data set modified by the weights is given to the lower class classifier for training. Finally, the classifier is fused as the final decision classifier. The accuracy of Naïve Bayes is the lowest one. The reason of low accuracy for NB is that in this algorithm every attribute is thought independent. However, in this project, some attributes are obviously relevant. Such as relationship and sex and marital-status. However, Naïve Bayes is the fastest method. Because the principle is the most simple and there is no iteration in this algorithm. For running time, NB is the fastest and SVC is the slowest algorithm. Therefore, cosidering the accuracy, F1 score and running time, AdaBoost is the most suitable classifier for this task.

For future work, we are planning to apply and compare other different algorithms to deal with this problem. In addition, we found several efficient algorithms for this task, such as AdaBoost and Random Forest. We could try to use these methods for different data set and evaluate the performance. This process may help us to find a universal pattern for certain type data set. Because the limitation of time, there are something to make the task better. For example, when training the different classifier, using parallel computing can reduce the running time. Although we consider the overfitting, for some algorithm, it is the best performance. In the future work, could use cross validation to test if all classifier get the best performance.

6. Reference

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185.
Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18.
Mihov, V. (2017). Adult Income Data Set Analysis with IPython. [online] Valentin Mihov's Blog. Available at: https://www.valentinmihov.com/2015/04/17/adult-income-data-set/.HUSTLX (2016).
Example of data mining using weka and clementine. [online] Cnblogs.com. Available at: http://www.cnblogs.com/hustl
Schapire, R. E. (2013). Explaining adaboost. In Empirical inference (pp. 37-52). Springer, Berlin, Heidelberg.
Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. ICML. 2004. Saharon Rosset. Model selection via the AUC. ICML. 2004.
Rich Caruana and Alexandru Niculescu-Mizil. An Empirical Evaluation of Supervised Learning for ROC Area.
S. Sathiya Keerthi and Kaibo Duan and Shirish Krishnaj Shevade and Aun Neow Poo. A Fast Dual Algorithm for Kernel Logistic Regression. ICML. 2002.

7. Appendix

.1. Contribution of group member

Heng Guo: Analysis data, Coding, Pre-processing, Experiments and Discussion, Use LaTeX for typesetting
Shen Wang: Analysis data, Previous work, Pre-processing, Discussion
Haoxing Wu: Abstract, Introduction, Classifier algorithm methods analysis, Conclusion

.2. operating environment

software environment: python3.6.5
hardware environment: Apple MacBook Pro, Intel Core i5, 8 GB Memory

.3. How to get the data and external lib

For this assignment, we choose the Adult dataset for the classification. The data is downloaded from https://archive.ics.uci.edu/ml/datasets/adult.
We use sklearn as the third part lib. It can be download use command 'pip install -u scikit-learn', or from http://scikit-learn.org/stable/install.html.

.4. How to run the code

Download the adult.data.csv from https://archive.ics.uci.edu/ml/datasets/adult and put the adult.data.csv(if it is a txt file, chage the suffix to .csv) file in the same folder of the python file 'assignment2.py'
Then input command 'python3 assignment2.py' to run the program.

python3 assignment2.py

The preprocessipynb is a python notebook file that we use to analyze the data such as the allocation of the data. Using jupyter notebook to open it.