CSCI3320 Fundamentals of Machine Learning

Project Report

WU Shishang
1155062020

GUO Jialiang
1155095825

Contents

Raw Data Preprocessing

Data Analysis

By counting the NaNs in each row, we found that the attribute votes nearly represents the number of valid values in a sample. Therefore, we dropped the samples whose votes are no greater than 30, obtaining a training set of 3176 samples.

By further observation, we found that almost every feature suffered a loss of approximately 1000 values, except that feature Q124742 has a loss of more than 2000 values. Hence, we dropped feature Q124742.

We then mapped all the values into numerical value. Finally, we used OneHotCode in Scikit-Learn to map these features to several binary sub-features which only take value 0 or 1. In summary, our training set is of size (3176,233), with test data of size (924,233).

Missing Data Filling

We have developed a new way to impute missing values. For a given feature, we first divide the training examples to two parts basing on whether this example has a missing value in this feature. Those with a value was treated as the training set, and we use these samples to build an random forest regressor, and then predict missing values in the testing set.

Unfortunately, this algorithm involves too much computation, and is therefore time-consuming. For demonstration purpose, we have chosen 6 'important' features to apply this algorithm, where NaNs in other features were imputed by the mode. That is, missing values of feature ['YOB', 'Gender', 'Income', 'HouseholdStatus' ,'EducationalLevel', 'Q101162'] were determined by machine learning algorithm, while others were replaced by the mode of the feature.

Classification

Logistic Regression

Our implementation is much slower than that of scikit-learn with lower train accuracy.

Naive Bayes

Our implementation is much slower than that of scikit-learn with lower train accuracy.

SVM

rbf is chosen as the kernel function by experience.

Random Forest

General Questions

- Q: What are the characteristics of each of the four classifiers?
  Logistic regression can describe the relationship between a categorical outcome (response variable) and a set of covariates (predictor variables). The categorical outcome may be binary or ordinal. The predictor variable(s) may be continuous or categorical.

  Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

  support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

  Random forests or random decision forests are an ensemble learning-method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests

correct for decision trees' habit of overfitting to their training set.

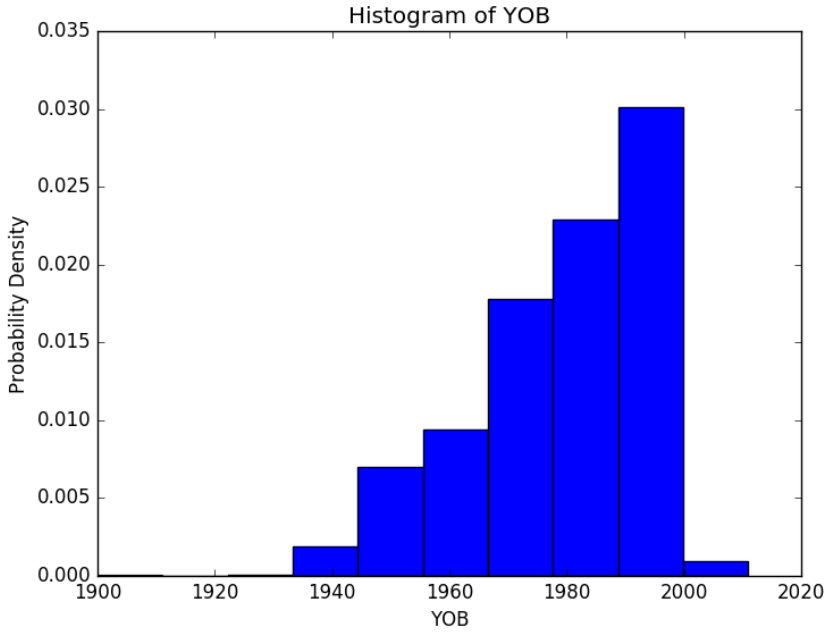
- Q: Different classification models can be used in different scenarios. How do you choose classification models for different classification problems? Please provide some examples.
  Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations where a small number of training data is required to estimate the parameters necessary for classification and the features are strongly independent.

- Q: How do the cross validation techniques help in avoiding overfitting?
  10-fold cross validation is used in this project to train the models.

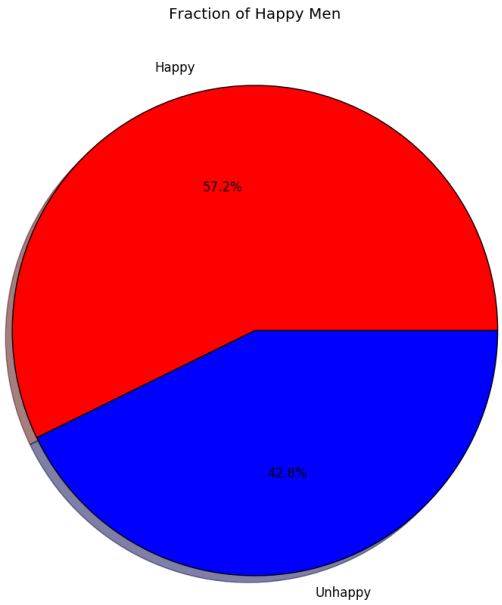
Visualization

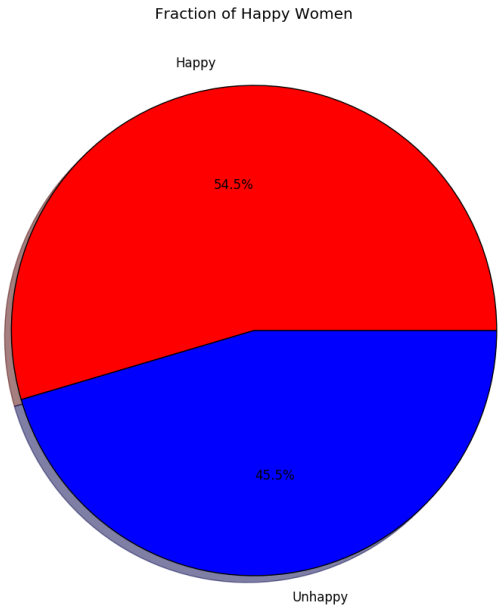
Basic Visualization Methods

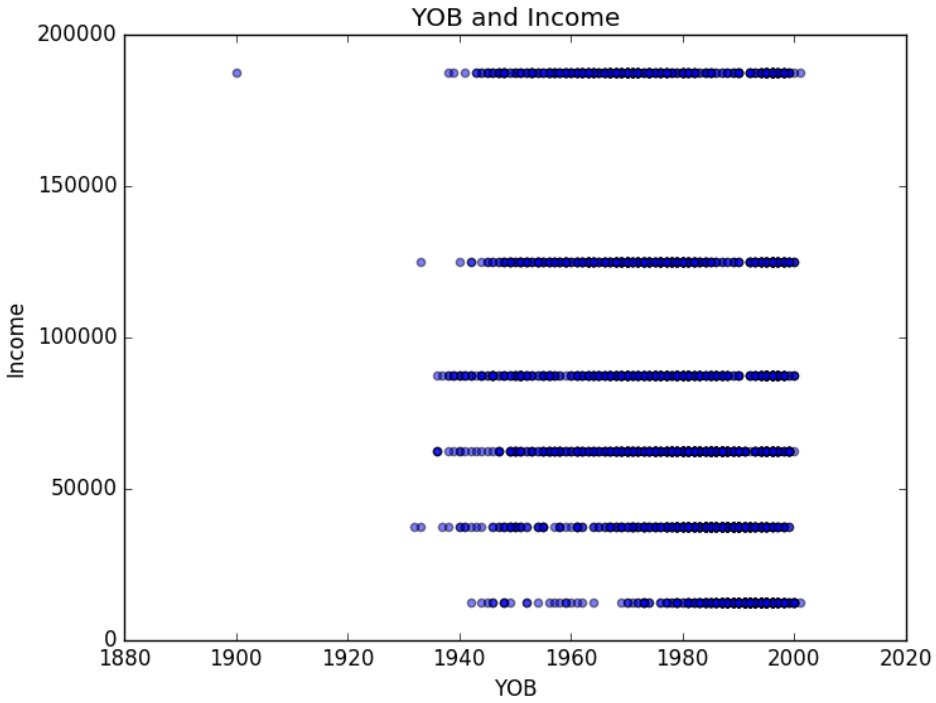
It is clearly shown that most people's YOB is between 1980 and 2000, in fact showing a upward trend.



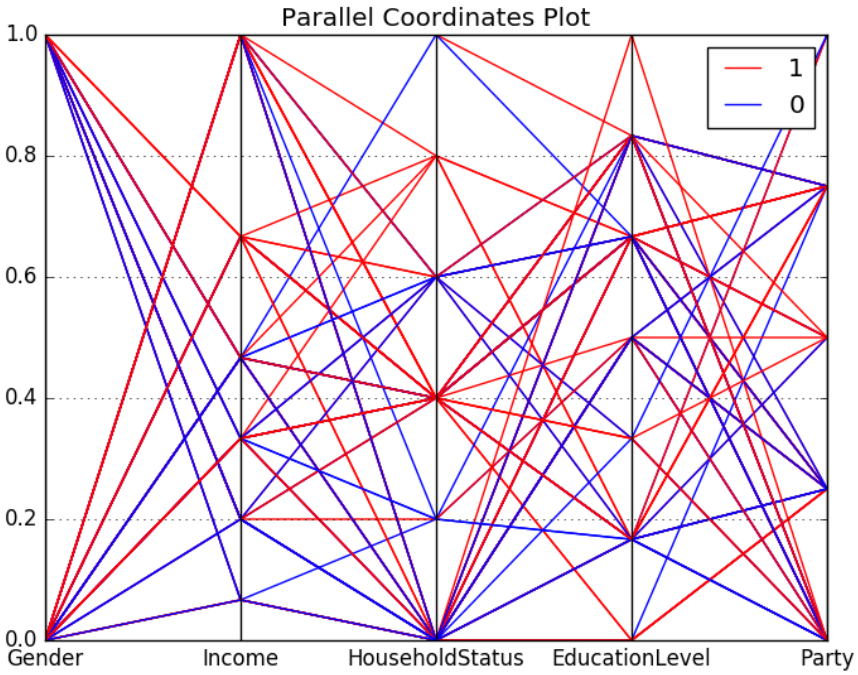
It can be concluded that happiness is not clearly related to gender

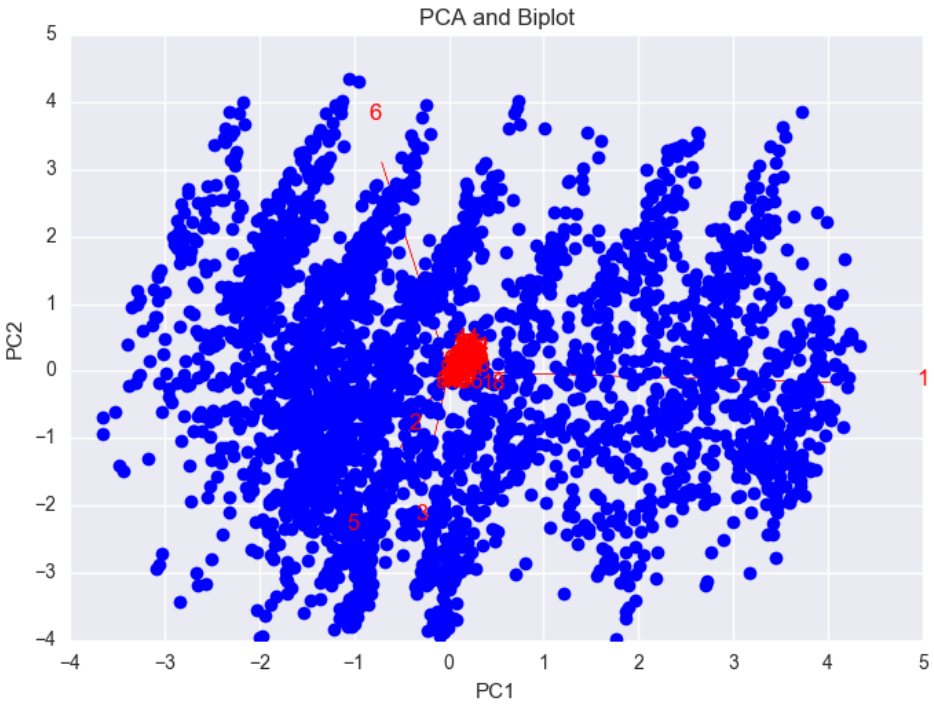




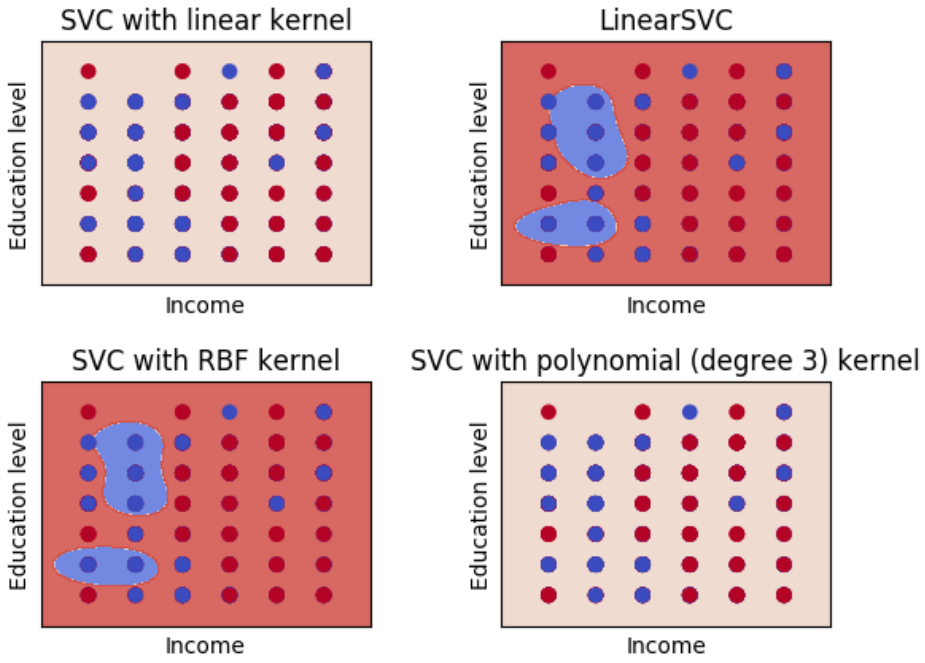


It can be concluded that income is not clearly related to YOB.


Parallel Coordinates Plot
1.0
0.8
0.6
0.4
0.2
0.0
1
0
Gender
Income
HouseholdStatus
EducationLevel
Party


PCA and Biplot
PC1
PC2
6
1
2
3
5
4
5
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4
5

Visualizing Classification Result



We actually used 503 samples to train, but data points covered each other because of discrete value of the two features. From the plot, we can see that SVC with RBF kernel and LinearSVC have the best performance.