Report - Main Steps

Arnold Jiadong Yu

March 8, 2019

1 Data Visualization

Data Preprocessing started in this step. It is very important to gain an initial understanding of how the correlation is between each independent variable and the dependent variable.

1. Import data into Tableau. Gain insight to missing values and data imbalanced.

2. Used Tableau to create histograms to visualize the relationship between each attribute and the class label.

3. Discretizeing continuous attributes and visualize the relationship between new attribute and the class label.

4. Recorded details of insights in the folder 1.Data Visualization.

2 Handling Missing Value

Data Preprocessing continues in this step. It is very important to handle missing value, remove duplicates items and generates new data files for analysis.

1. Import data.

2. Add Header to dataset.

3. Drop duplicate attribute which is Education-Num.

4. Remove all rows contain ’ ?’. Or replace all entry contain ’ ?’ with most frequency item in the attribute.

5. Output new data files for future analysis.

6. Recorded details and instruction of compiling and running in the folder 2.Handling missing value

3 Naive Bayesian Classifier

Data Proprocessing continues in this step before applying the classifier. In order to use Naive Bayesian Classifer with all categorical data, we need to discretize all continuous attributes. Recorded details and instruction of compiling and running in the folder 3.NaiveBayesion/dataset.

1. Write Naive Bayesian Classifier class for all categorical data and with Gaussian distribution.

2. Import data and run Naive Bayesian Classifier on them.

3. Calculate confusion matrix, accuracy, precision, recall, f1 and mcc.

4. Recorded result and instruction of compiling and running in the folder 3.NaiveBayesion

4 K-Fold Cross Validation

1. Write kFoldValidation class to validate out Naive Bayesian Classifier.

2. Import data and run k-fold cross validation on them.

3. Calculate all 10 accuracy and average accuracy.

4. Recorded result, evaluation, instruction of compiling and running in the folder 4.k-fold cross validation.