COMP219 - 2018 - First CA Assignment

Name: Hongyi Wu ID: 201376889

Document of implementation and evaluation of the result

Check list:

1. Loading Data
   - ✓ Successfully load the dataset
   - ✓ Display the dataset information
2. Training Data
   - ✓ Code for training and comments
   - ✓ Training can be done successfully and a simple test
3. Model Evaluation
   - ✓ Explain methods used and design of the evaluation experiment
   - ✓ Document the experiment results
4. Extra
   - ✓ Compare three models in evaluation phase

Loading Data

The dataset **Adult Data Set** is selected in UCI Machine Learning Repository.

It is a widely used dataset to predict whether income exceeds $50K/year according to provided census data.
Several steps to input the data:

1. Import pandas and use the function **read_csv** to input the dataset file "adult.data.txt".
2. Label the features of the file with Age, Workclass, fnlwgt, Education, Education_Num, Martial_Status, Occupation, Relationship, Race, Sex,

Capital_Gain, Capital_Loss, Hours_per_week, Country, Target.

3. Replace String data to numbers. From example, there are two strings in sex label, Female and Male, which should be replaced by 0 and 1.
4. Replace missing NaN values with rational numbers by fillna function.
5. Select list of features and targets by:
   X = data[features].values
   y = data["target"]
6. Display numbers of entries by:
   print('numbers of data entries: ',X.shape[0])

   The result should be:

```
In [135]: runfile('/Users/Hongyi.Wu/Desktop/adult.py', wdir='/Users/Hongyi.Wu/Desktop')
numbers of data entries:  32561
numbers of data features:  11
```

Training Data

First considered model is trained by logistic regression. (**model 1-3 in code**)
At the beginning, the basic regression is trained by this code:

lr3 = LogisticRegression(C = 1000.0, random_state = 0)
lr3.fit(X_train_std, y_train)

However, the result shows the Accuracy will be lower if we choose X_train_std rather than X_train which is conducted by transformation manually.

After that, there is an optimized algorithm by logistic regression:

```
def logistic_regression():

    regression = LogisticRegression()
    grid = {'penalty': ['l1', 'l2'], 'C': [0.01, 0.1, 1, 10, 100, 1000]}
#try to get best penalty from l1, l2 and C from 0.01, 0.1, 1, 10, 100, 1000
    lr = GridSearchCV(regression, param_grid=grid, scoring='accuracy')
    lr.fit(X_train, y_train)

     return lr.best_estimator_
```

The function chooses the best estimator from both of the penalty l1 and l2 and various C, then generate a grid to choose the best one.

Similarly, other models are trained by k-nearest neighbors.
(**model 3 in code**)

After that, it will print out the misclassified samples and Accuracy to validate the successful training, a predict result of one sample will also display.

```
Misclassified samples: 1518
Accuracy: 0.84
predict:  [1]
true:  [1]
```

Model Evaluation

As for the evaluation of the models, since algorithm of logistic regression preforms well in **GridSearchCV** function, it is chosen to compare with traditional logistic regression and k-nearest neighbor algorithm.
Firstly, three tables are generated to show the result of models.

	precision	**recall**	**f1-score**	**support**
Under 50k	0.87	0.93	0.90	7407
Over 50k	0.72	0.58	0.64	2362
avg / total	0.84	0.85	0.84	9769

Table 1: Classification result of linear regression with GridSearchCV function

	precision	**recall**	**f1-score**	**support**
Under 50k	0.81	0.97	0.88	7407
Over 50k	0.73	0.27	0.39	2362
avg / total	0.79	0.80	0.76	9769

Table 2: Classification result of linear regression with traditional function (C=1000)

	precision	**recall**	**f1-score**	**support**
Under 50k	0.81	0.92	0.86	7407
Over 50k	0.56	0.31	0.40	2362
avg / total	0.75	0.77	0.75	9769

Table 3：Classification result of k-nearest neighbor algorithm

The results show that the first model perform best in precision, recall, and f1-score.

To display the comparation visually, confusion matrix, ROC curve and Precision-Recall curve and a statistics diagram.

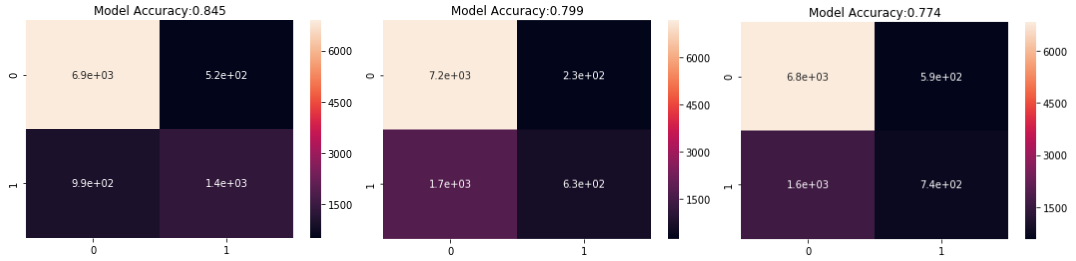

Figure 1: confusion matrix of three models

From confusion matrix we calculate the model accuracy which model one also perform best in the three models.

As for ROC curve, we can also easily see the difference of the three models.

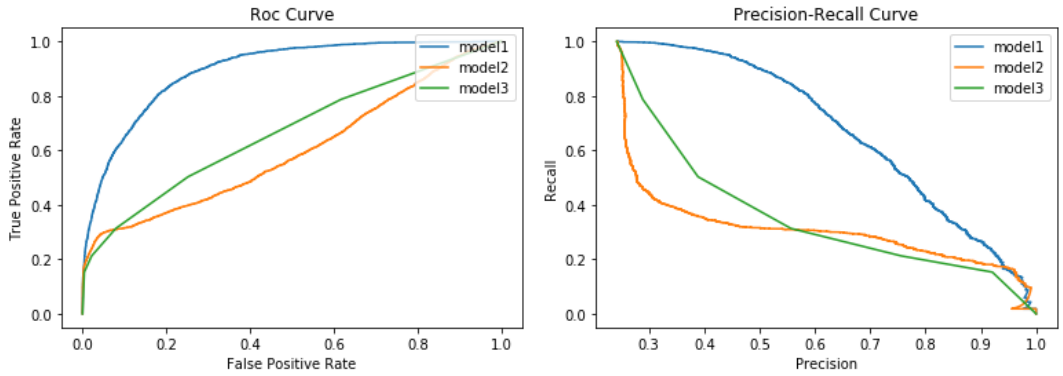

Figure 2: Roc Curve and Precision-Recall Curve of three models

The AUC (area under curve) of model one is obviously higher than other two models. As for Precision-Recall Curve, traditional logistic regression preforms obviously worst and optimized logistic regression performs well.

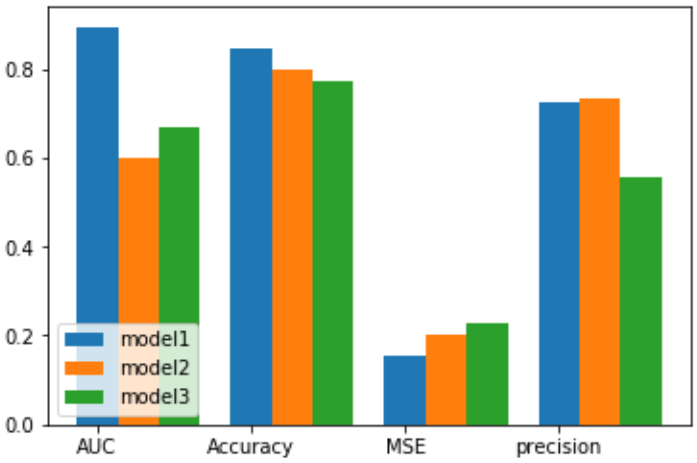

Figure 3: statics of the three models

As the figure above, the higher MSE (mean squared error) means the worse performance of misclassification samples. Therefore, model 1 is better in classification. We can conclude model with optimized logistic regression is the best among three models expect for precision, since model in traditional logistic regression is a bit higher than model 1 in this feature.