# Logistic Regression

December 13, 2019

[9]:
```
from sklearn.model_selection import train_test_split, cross_val_score, KFold,␣
↪StratifiedKFold
from sklearn.metrics import roc_auc_score, auc, roc_curve
# machine learning
from sklearn.linear_model import LogisticRegression
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn import metrics
import numpy as np
```

[2]:
```
train_df = pd.read_csv('train.csv')
test_df = pd.read_csv('test.csv')
```

[3]:
```
act_test_df = pd.read_csv('act_test.csv', dtype={'people_id': np.str,␣
↪'activity_id': np.str},
                          parse_dates=['date'])
```

[4]:
```
test_id = act_test_df.activity_id
```

[5]:
```
X_train = train_df.drop(['outcome'], axis=1)
Y_train = train_df['outcome']
```

[6]:
```
# train, validation set split
```

[7]:
```
x_train, x_val, y_train, y_val = train_test_split(X_train, Y_train, test_size =␣
↪0.5, random_state=1)
x_train.shape, x_val.shape, y_train.shape, y_val.shape
```

[7]:
```
((1098645, 59), (1098646, 59), (1098645,), (1098646,))
```

[10]:
```
logreg = LogisticRegression()
logreg.fit(x_train, y_train)
```

```
/Users/xigao/anaconda3/lib/python3.7/site-
packages/sklearn/linear_model/logistic.py:432: FutureWarning: Default solver
will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
```

[10]:
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='warn', tol=0.0001, verbose=0,
                   warm_start=False)
```

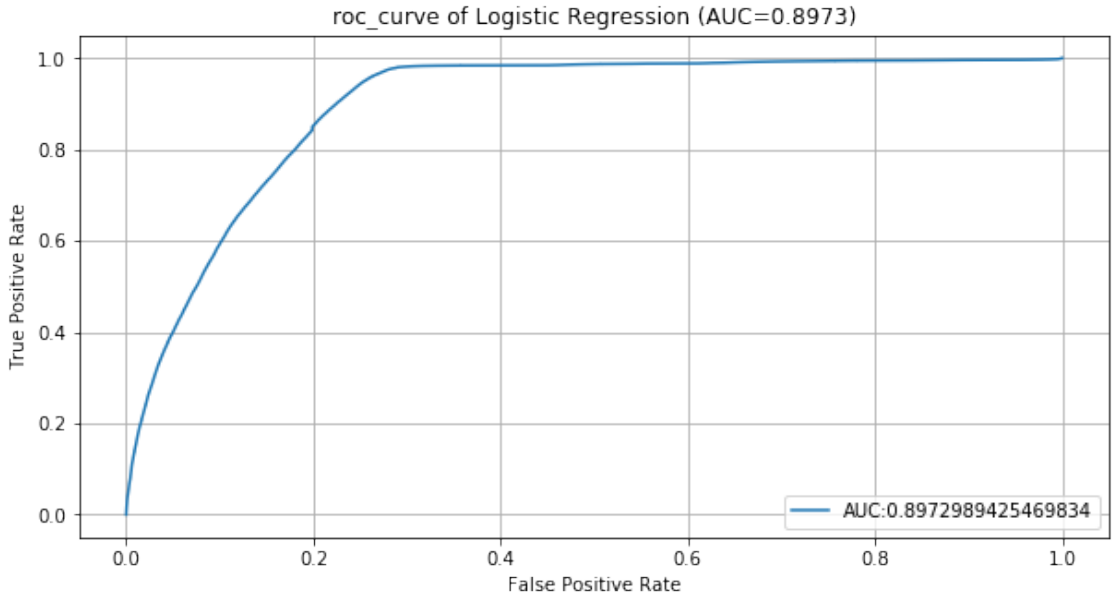
[11]:
```
acc_log = round(logreg.score(x_val, y_val) * 100, 2)
acc_log
```

[11]:
```
82.71
```

[13]:
```
lr_predictions = logreg.predict_proba(x_val)[:,1]
fpr, tpr, thresholds = metrics.roc_curve(y_val,lr_predictions)
lr_roc = pd.DataFrame()
lr_roc['fpr'] = fpr
lr_roc['threshold'] = thresholds
auc = metrics.roc_auc_score(y_val,lr_predictions)
auc
```

[13]:
```
0.8972989425469834
```

[14]:
```
plt.figure(figsize=(10,5))
plt.plot(fpr,tpr,label='AUC:'+str(auc))
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('roc_curve of Logistic Regression (AUC=%.4f)' %(auc))
plt.legend(loc=4)
plt.grid()
```




[15]:
```
Y_pred_lr= logreg.predict(test_df)
```

[16]:
```
submission_lr = pd.DataFrame({'activity_id' : test_id, 'outcome': Y_pred_lr})
submission_lr.to_csv('submission_lr.csv', index = False)
```

[ ]:
```
# kaggle score of Linear Regression: 0.81687
```