

Assignment 3

May 31, 2020

You are currently looking at **version 1.2** of this notebook. To download notebooks and datafiles, as well as get help on Jupyter notebooks in the Coursera platform, visit the [Jupyter Notebook FAQ](#) course resource.

1 Assignment 3 - Evaluation

In this assignment you will train several models and evaluate how effectively they predict instances of fraud using data based on [this dataset from Kaggle](#). Each row in `fraud_data.csv` corresponds to a credit card transaction. Features include confidential variables `V1` through `V28` as well as `Amount` which is the amount of the transaction. The target is stored in the `class` column, where a value of 1 corresponds to an instance of fraud and 0 corresponds to an instance of not fraud.

```
In [1]: import numpy as np
import pandas as pd
```

1.0.1 Question 1

Import the data from `fraud_data.csv`. What percentage of the observations in the dataset are instances of fraud?

This function should return a float between 0 and 1.

```
In [15]: def answer_one():

    # Your code here
    data = pd.read_csv('fraud_data.csv')
    tgt_class = data.iloc[:, -1]
    return float(tgt_class[tgt_class.values == 1].count() / tgt_class.count())

answer_one()
```

```
Out[15]: 0.016410823768035772
```

```
In [16]: # Use X_train, X_test, y_train, y_test for all of the following questions
        from sklearn.model_selection import train_test_split

        df = pd.read_csv('fraud_data.csv')

        X = df.iloc[:, :-1]
        y = df.iloc[:, -1]

        X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
```

1.0.2 Question 2

Using `X_train`, `X_test`, `y_train`, and `y_test` (as defined above), train a dummy classifier that classifies everything as the majority class of the training data. What is the accuracy of this classifier? What is the recall?

This function should return a tuple with two floats, i.e. (accuracy score, recall score).

```
In [18]: def answer_two():
        from sklearn.dummy import DummyClassifier
        from sklearn.metrics import recall_score

        # Your code here
        clf = DummyClassifier(strategy='most_frequent').fit(X_train, y_train)
        y_pred = clf.predict(X_test)

        recall_score = recall_score(y_test, y_pred)
        accuracy_score = clf.score(X_train, y_train)

        return (accuracy_score, recall_score) # Return your answer

        answer_two()
```

```
Out[18]: (0.98303522035773561, 0.0)
```

1.0.3 Question 3

Using `X_train`, `X_test`, `y_train`, `y_test` (as defined above), train a SVC classifier using the default parameters. What is the accuracy, recall, and precision of this classifier?

This function should return a tuple with three floats, i.e. (accuracy score, recall score, precision score).

```
In [22]: def answer_three():
        from sklearn.metrics import recall_score, precision_score
        from sklearn.svm import SVC

        # Your code here
        clf = SVC().fit(X_train, y_train)

        y_pred = clf.predict(X_test)
```

```

acc_score = clf.score(X_test, y_test)
rec_score = recall_score(y_test, y_pred)
pre_score = precision_score(y_test, y_pred)

    return (acc_score, rec_score, pre_score) # Return your answer

answer_three()

```

Out[22]: (0.99078171091445433, 0.375, 1.0)

1.0.4 Question 4

Using the SVC classifier with parameters `{'C': 1e9, 'gamma': 1e-07}`, what is the confusion matrix when using a threshold of -220 on the decision function. Use `X_test` and `y_test`.

This function should return a confusion matrix, a 2x2 numpy array with 4 integers.

```

In [24]: def answer_four():
    from sklearn.metrics import confusion_matrix
    from sklearn.svm import SVC

    # Your code here
    clf = SVC(C=1e9, gamma=1e-07).fit(X_train, y_train)
    y_pred = clf.decision_function(X_test) > -220

    return confusion_matrix(y_test, y_pred) # Return your answer

answer_four()

```

Out[24]: array([[5320, 24],
[14, 66]])

1.0.5 Question 5

Train a logistic regression classifier with default parameters using `X_train` and `y_train`.

For the logistic regression classifier, create a precision recall curve and a roc curve using `y_test` and the probability estimates for `X_test` (probability it is fraud).

Looking at the precision recall curve, what is the recall when the precision is 0.75?

Looking at the roc curve, what is the true positive rate when the false positive rate is 0.16?

This function should return a tuple with two floats, i.e. (recall, true positive rate).

```

In [29]: def answer_five():

    # Your code here
    %matplotlib notebook
    import matplotlib.pyplot as plt
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import precision_recall_curve, roc_curve

```

```

y_scores_lr = LogisticRegression().fit(X_train, y_train).decision_function(X_test)
precision, recall, threshold = precision_recall_curve(y_test, y_scores_lr)
fpr_lr, tpr_lr, _ = roc_curve(y_test, y_scores_lr)

closest_zero_p = np.argmin(np.abs(precision-0.75))
closest_zero_r = recall[closest_zero_p]

#plt.figure()
#plt.plot(precision, recall)
#plt.plot(fpr_lr, tpr_lr)

closest_zero_fpr_lr = np.argmin(np.abs(fpr_lr - 0.16))
closest_zero_tpr_lr = recall[closest_zero_fpr_lr]

return (closest_zero_r, closest_zero_tpr_lr) # Return your answer

answer_five()

```

Out [29]: (0.82499999999999996, 0.98750000000000004)

1.0.6 Question 6

Perform a grid search over the parameters listed below for a Logistic Regression classifier, using recall for scoring and the default 3-fold cross validation.

```

'penalty': ['l1', 'l2']
'C':[0.01, 0.1, 1, 10, 100]

```

From `.cv_results_`, create an array of the mean test scores of each parameter combination. i.e.

	l1	l2
0.01	?	?
0.1	?	?
1	?	?
10	?	?
100	?	?

This function should return a 5 by 2 numpy array with 10 floats.

Note: do not return a DataFrame, just the values denoted by '?' above in a numpy array. You might need to reshape your raw result to meet the format we are looking for.

```

In [32]: def answer_six():
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression

# Your code here
grid_values={'penalty': ['l1', 'l2'], 'C':[0.01, 0.1, 1, 10, 100]}
lr = LogisticRegression()

```

```

# Number of folds included as a parameter in GridSearchCV. Default cv=
lr_custom = GridSearchCV(lr,param_grid=grid_values,scoring='recall',cv=
lr_custom.fit(X_train,y_train)

    return lr_custom.cv_results_['mean_test_score'].reshape(5,2) # Return

answer_six()

Out[32]: array([[ 0.66666667,  0.76086957],
 [ 0.80072464,  0.80434783],
 [ 0.8115942 ,  0.8115942 ],
 [ 0.80797101,  0.8115942 ],
 [ 0.80797101,  0.80797101]])

In [33]: # Use the following function to help visualize results from the grid search
def GridSearch_Heatmap(scores):
    %matplotlib notebook
    import seaborn as sns
    import matplotlib.pyplot as plt
    plt.figure()
    sns.heatmap(scores.reshape(5,2), xticklabels=['l1','l2'], yticklabels=
    plt.yticks(rotation=0);

#GridSearch_Heatmap(answer_six())

```