BBC_Classification

November 15, 2022

1 CSCC11 - Introduction to Machine Learning, Fall 2022, Assignment 2

1.1 Authors

Shawn Santhoshgeorge (1006094673)
Anaqi Amir Razif (1005813880)

1.2 Part 2: Programming Component

1.2.1 Setup

[1]:
```
from helper import *
from NB import NB
from GCC import GCC
from KNN import KNN

OUTPUT = 'latex' # or 'html'
```

1.2.2 Naive Bayes

[3]:
```
X_train, y_train, X_test, y_test, labels = load_data(mode=Mode.BINARY, train_size=0.7, test_size=0.3)

n_b = NB()
n_b.train(X_train, y_train, labels)

# Using Training Dataset
y_pred_train = n_b.predict(X_train)
label_acc_train, overall_acc_train = get_accuracy(labels, y_train, y_pred_train)

# Using Testing Dataset
y_pred_test = n_b.predict(X_test)
label_acc_test, overall_acc_test = get_accuracy(labels, y_test, y_pred_test)

create_report(label_acc_train, label_acc_test, overall_acc_train, overall_acc_test, output=OUTPUT)
```

Training Accuracy

Label	Accuracy
business	79.26 %
entertainment	96.24 %
politics	98.39 %
sport	98.89 %
tech	97.01 %

Testing Accuracy

Label	Accuracy
business	72.15 %
entertainment	96.67 %
politics	100.0 %
sport	98.67 %
tech	97.01 %

Training Overall Accuracy: 93.71 % Testing Overall Accuracy: 91.92 %

1.2.3 Gaussian Class Conditionals

[2]:
```
X_train, y_train, X_test, y_test, labels = load_data(mode=Mode.FREQ, train_size=0.7, test_size=0.3)

g_cc = GCC()
g_cc.train(X_train, y_train, labels)

# Using Training Dataset
y_pred_train = g_cc.predict(X_train)
label_acc_train, overall_acc_train = get_accuracy(labels, y_train, y_pred_train)

# Using Testing Dataset
y_pred_test = g_cc.predict(X_test)
label_acc_test, overall_acc_test = get_accuracy(labels, y_test, y_pred_test)

create_report(label_acc_train, label_acc_test, overall_acc_train, overall_acc_test, output=OUTPUT)
```

Training Accuracy

Label	Accuracy
business	100.0 %
entertainment	100.0 %
politics	100.0 %
sport	100.0 %
tech	100.0 %

Testing Accuracy

Label	Accuracy
business	85.44 %
entertainment	93.33 %
politics	92.45 %
sport	96.0 %
tech	91.79 %

Training Overall Accuracy: 100.0 % Testing Overall Accuracy: 91.62 %

1.2.4 K-Nearest Neighbours

[4]:
```
for mode in Mode:
    if OUTPUT == 'html':
        display_html(f'<center> <h3> With {mode.value} Data </h3> </center>', raw=True)
    elif OUTPUT == 'latex':
        display(Math(r'\text{With }' + r'\text{' f'{mode.value}' + r'}' + r'\text{ Data}'))
    X_train, y_train, X_test, y_test, labels = load_data(mode=mode, train_size=0.7, test_size=0.3)
    for k in [1, 3, 6]:
        if OUTPUT == 'html':
            display_html(f'<h4> For K = {k} </h4>', raw=True)
        elif OUTPUT == 'latex':
            display(Math(r'\text{For K = }' + f'{k}'))
        k_nn = KNN(k)
        k_nn.train(X_train, y_train)

        y_pred_train = k_nn.predict(X_train)
        label_acc_train, overall_acc_train = get_accuracy(labels, y_train, y_pred_train)

        y_pred_test = k_nn.predict(X_test)
        label_acc_test, overall_acc_test = get_accuracy(labels, y_test, y_pred_test)

        create_report(label_acc_train, label_acc_test, overall_acc_train, overall_acc_test)
```

With Frequency Data

For K = 1

Training Accuracy

Label	Accuracy
business	100.0 %
entertainment	100.0 %
politics	100.0 %
sport	100.0 %
tech	100.0 %

Testing Accuracy

Label	Accuracy
business	79.75 %
entertainment	60.0 %
politics	73.58 %
sport	100.0 %
tech	61.94 %

Training Overall Accuracy: 100.0 % Testing Overall Accuracy: 76.2 %

For K = 3

Training Accuracy

Label	Accuracy
business	82.1 %
entertainment	71.43 %
politics	71.38 %
sport	99.72 %
tech	74.53 %

Testing Accuracy

Label	Accuracy
business	69.62 %
entertainment	50.83 %
politics	62.26 %
sport	100.0 %
tech	45.52 %

Training Overall Accuracy: 80.92 % Testing Overall Accuracy: 67.07 %

For K = 6

Training Accuracy

Label	Accuracy
business	69.32 %
entertainment	53.76 %
politics	58.84 %
sport	100.0 %
tech	43.45 %

Testing Accuracy

Label	Accuracy
business	56.96 %
entertainment	44.17 %
politics	52.83 %
sport	100.0 %
tech	31.34 %

Training Overall Accuracy: 67.24 % Testing Overall Accuracy: 58.53 %

With Binary Data

For K = 1

Training Accuracy

Label	Accuracy
business	100.0 %
entertainment	100.0 %
politics	100.0 %
sport	100.0 %
tech	100.0 %

Training Overall Accuracy: 100.0 %

Testing Accuracy

Label	Accuracy
business	39.24 %
entertainment	34.17 %
politics	28.3 %
sport	41.33 %
tech	50.75 %

Testing Overall Accuracy: 39.37 %

For K = 3

Training Accuracy

Label	Accuracy
business	78.12 %
entertainment	58.65 %
politics	63.02 %
sport	68.7 %
tech	77.53 %

Training Overall Accuracy: 69.49 %

Testing Accuracy

Label	Accuracy
business	49.37 %
entertainment	33.33 %
politics	27.36 %
sport	38.0 %
tech	64.93 %

Testing Overall Accuracy: 43.56 %

For K = 6

Training Accuracy

Label	Accuracy
business	77.27 %
entertainment	56.39 %
politics	58.84 %
sport	52.35 %
tech	67.79 %

Training Overall Accuracy: 62.62 %

Testing Accuracy

Label	Accuracy
business	63.29 %
entertainment	35.83 %
politics	37.74 %
sport	28.67 %
tech	60.45 %

Testing Overall Accuracy: 45.96 %