Machine learning Assignment-2

Nishant Gupta(2018SIY7502)

Question-1: Naive Bayes:

part-A:

```
train accuracy: 62.1293212981628831
    recall=0.61634612312105
    precision=0.5477166299122

    f-score=0.580008286556

test accuracy: 59.110217023886094

    recall=0.5868003454665505
    precision=0.5546856584327722

    f-score:0.570291242962
```

part-B:

```
test accuracy using random class: 20.00329050688763
    recall=0.20097152553287373
    precision=0.20058101439606935
    f-score=0.200776080078

test accuracy using majority class: 43.9895900327555
    recall=0.20003532986893652
    precision=0.24797738608915446

    f-score=0.221441206731
```

part-C:

```
[[14180  4736  1009   150    94]
 [ 2731  4250  3313   434   110]
 [ 1385  2497  6655  3607   387]
 [ 1361  1436  5026 17246  4289]
 [ 4676  1119  1881 18967 32179]]
```

part-D:

```
test accuracy  : 57.1314258364618
    recall = 0.5382725032269992
    precision = 0.514217248324144

    f-score=0.525969977475
```

part-E:

'lemmatization,stopwords and bigram'

```
test accuracy: 58.76172243078718

recall=0.477228625286

precision=0.495788644046

f1 score: 0.486331621212
```

'stopwords and bigram'

```
test accuracy using: 63.69823060470543

recall=0.51620996533

precision=0.54576829329

f1 score: 0.530577776844
```

part-F: if the data is biased(the data has too many examples for one class and very less for others then f-score gives better results than accuracy)

Question 2: MNIST Handwritten digit Classification using Support vector Machines

1. Binary Classification:

A) Accuracy using linear kernel and CVXOPT package: 99.7%

B) Accuracy using gaussian kernel and CVXOPT package: 99.17%

C) Accuracy using linear kernel and LIBSVM package: 98.99%
Accuracy using gaussian kernel and LIBSVM package: 99.14m
Libsvm takes very less time as compared to the cvxopt package

2. **Multiclass classification:**

A) Train data Accuracy: 98.56%
Test data Accuracy: 92.4%
Time taken: 7 hours

B) Train data Accuracy: 99.91%
Test data Accuracy: 97.23%
Time taken: 25 minutes

C) **Confusion matrix**:

```
        0      1    2    3    4    5    6     7    8     9
0    [[969   0    1    0    0    3    4    1    2    0]
1    [0    1121   3   2    1   2  2    0   3   1]
2    [4   0   1000   4   2   0   1   6   15   0]
3    [0   0   8   985   0   4   0   6    5   2]
4    [0   0   4   0   962   0   6   0   2   8]
5    [2   0   3   6   1   866   7   1   5   1]
6    [6   3   0   0   4   4   939   0   2   0]
7    [1   4   19  2   4   0   0   987   2   9]
8    [4   0   3   10   1   5  3   3   942   3]
9    [4   4   3   6   13   4   0   9   12   952]]
```

Most of the digits are classisified correctly as can be seen on the diagonal entries. The missclassification error between two digits is because of there resemblance. Like 7 and 1 hae many pixels with same value so they have some misclassification.

D)

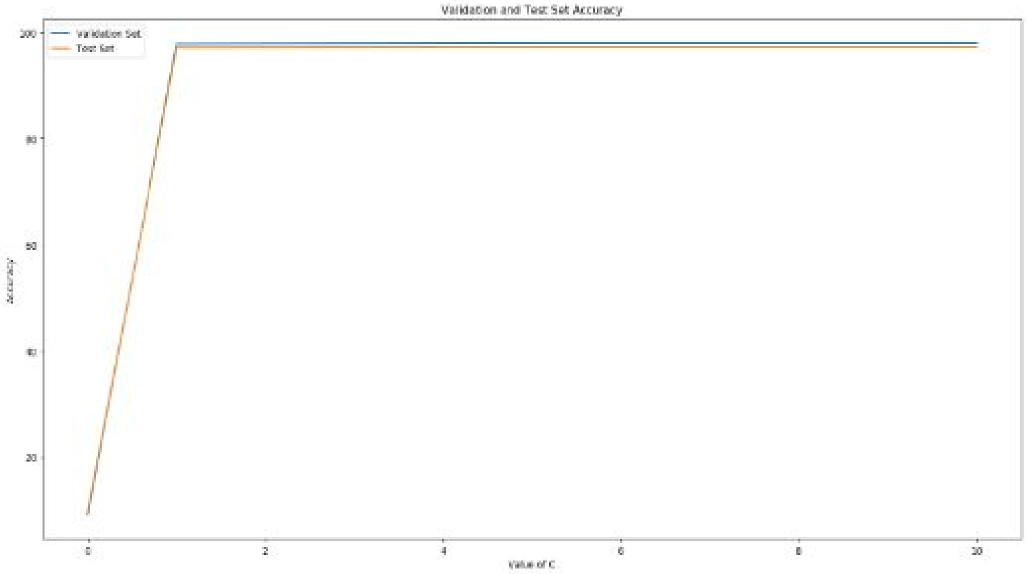
c=[0.00001,0.001,1,5,10]

Accuracy

validation set: [9.45, 9.45, 97.15, 97.35, 97.35]

test set : [72.1,72.1,97.23,97.29,97.29]

c=5 and 10 gives the best accuracy. From 5 to 10 there is no change in test set accuracy. This means that increasing the values of c will not change the accuracy.


Validation and Test Set Accuracy
Validation Set
Test Set
Accuracy
Value of C