COL774 Assignment – 2

**Anirudha Kulkarni**

**2019CS50421**

Code running instructions:

1. All codes are arranged in respective folders

2. Each question contains a asset folder where all plots/GIFs will be stored

3. Each code is in a python file and the parameters are written in the very first line of each function which can be changed

4. Data sets need to be put in a directory in submission directory with name “data” which can be passed as an argument if not

5. python files can be run with `python q1.py`

**Q2: Naïve Bayes**

**a) Naïve Bayes with Laplace smoothing and logarithms:**

Train accuracy: 51.038

Test accuracy: 0.6595714285714286

**b) Random predictions, Majority Predictions:**

Random Train: 0.20118

Random Test: 0.1985

Majority Train: 0.51864

Majority Test: 0.6607857142857143

**c) Confusion matrix**

**d) Removal of stopwords, stemming**

**e) Feature engineering:**

**i. Bingram**

**ii. Lemmatization**

**f) F1- score**

**g) With summary removal of stopwords, stemming**

**Q2: SVM**

**a) Binary classification:**

Dual problem with noise:

Text, letter

Description automatically generated

CVXOPT Format:

Text, letter

Description automatically generated

Max can be converted to min using negative sign,

Text, letter

Description automatically generated

**i) Linear Kernel:**

**Finding P,q,G,h,A,b:**

q­­ T = [-1, -1, -1, …, -1]

A = YT

b = [0]

Pi, j = y(i) y(j)x(j)Tx(j)

h T = [0, 0, …, 0, C, C, …, C]

G = A picture containing table

Description automatically generated

The condition is modelled using G. First m rows cover the part and second m rows cover

Parameters:

d=1, threshold = 0.0001, C=1.0

**Support vectors:** Indices of support vectors are saved in a file *support\_vector\_linear.txt*.

Weight = stored in the file

bias = 0.795

nSV = 95

Validation accuracy: 100%

Test accuracy: 98.89%

Computational cost: 8.87 seconds

Confusion Matrix Train:

Chart

Description automatically generated

2000 0

0 2000

Confusion Matrix test:

Chart

Description automatically generated

1012 20

5 1130

**ii) Gaussian Kernel:**

Everything same as linear kernel except P

Pi, j = y(i) y(j) K(x(j), x(j))

W is given by,

Text, letter

Description automatically generated

b is given by,

A picture containing text, watch

Description automatically generated

Parameters:

d=1, threshold = 0.0001, Gamma = 0.05, C=1.0

**Support vectors:** Indices of support vectors are saved in a file *support\_vector\_linear.txt*.

nSV = 867

Validation accuracy: 99.825%

Test accuracy: 95.846%

Computational cost: 28.100 sec

**iii) LIBSVM:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear | | Gaussian | |
|  | CVXPOT | LIBSVM | CVXPOT | LIBSVM |
| Validation accuracy | 100% | 100% | 99.825% | 99.975% |
| Test accuracy | 98.84% | 99.03% | 95.846% | 99.58% |
| nSV | 158 | 158 | 845 | 847 |
| Bias | 1.377 | -1.219 | -- | 0.890 |
| Computational cost | 37.28 | 0.50 | 28.100 | 1.988 |

Chart

Description automatically generated Chart

Description automatically generatedChart

Description automatically generated Chart

Description automatically generated

**a) Multi-class classification:**

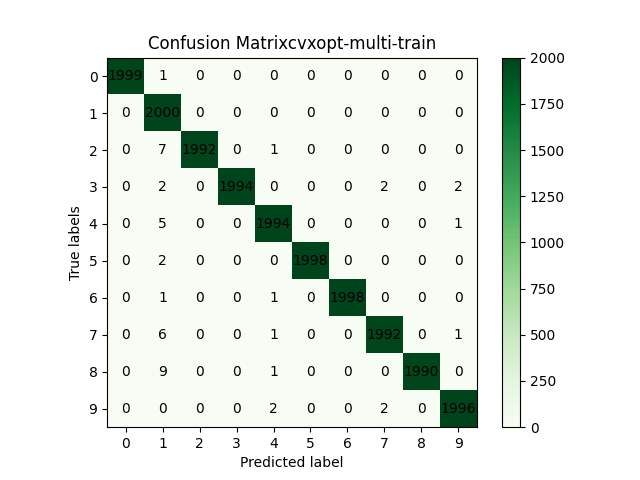
**i) CVXOPT:**

Total training time: 1198.8615338802338

Multiclass LIBSVM Training accuracy: 0.99765

Multiclass LIBSVM Test accuracy: 0.9494

Total nSV: 9866

Chart, scatter chart

Description automatically generated

**ii) LIBSVM:**

Training time: 170.986590385437

Total nSV: 10493

Train accuracy: 99.92% (19984/20000) (classification)

Test Accuracy: 97.23% (9723/10000) (classification)

**iii) Confusion matrices:**

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Missed train examples:



Missed test examples:



Miss-classified digits:

Most miss-classified are (7,2), (2,7), (8,3), (2,8), (9,4), (9,8), (9,4), (5,6), (6,0), (7,9)

Most of the results are intuitive as these are confusing to humans too.