REPORT

ASSIGNMENT 2

Sudeep Agarwal(2015CS50295)

1. Naive Bayes

In this part we have applied a multinomial naive bayes model to predict ratings of movie reviews.

a)

```
Processing Input Time =  26.75474715232849s
Training time =  6.53382682800293s
Training set prediction time =  30.10230827331543s
Training accuracy =  65.872%
Test set prediction time =  40.59362292289734s
Test accuracy =  38.904%
```

b)

```
Random guess prediction accuracy =  12.468%
Most occuring class prediction accuracy =  20.4%
```

c)

	Pred: 1	Pred: 2	Pred: 3	Pred: 4	Pred: 7	Pred: 8	Pred: 9	Pred: 10
Actual: 1	18.4	0.04	0.16	0.508	0.088	0.18	0.012	0.7
Actual: 2	7.44	0.048	0.216	0.724	0.096	0.136	0	0.548
Actual: 3	6.928	0.04	0.316	1.4	0.22	0.452	0.004	0.804
Actual: 4	5.528	0.016	0.248	2.12	0.508	0.832	0.028	1.26
Actual: 7	2.164	0.012	0.088	0.616	1.012	1.904	0.048	3.384
Actual: 8	2.012	0	0.08	0.392	0.604	2.332	0.084	5.896
Actual: 9	1.424	0	0.008	0.12	0.276	1.424	0.088	6.036
Actual: 10	3.244	0.016	0.032	0.168	0.24	1.552	0.156	14.588

In the above confusion matrix, each entry is percentage of test cases for which prediction = column and actual = row. The diagonal entries denotes the cases for which prediction is true. Significant values are

```
Prediction = 1 and Actual = 1
Prediction = 10 and Actual = 10
Prediction = 1 and Actual = 2
Prediction = 1 and Actual = 3
Prediction = 10 and Actual = 9
Prediction = 10 and Actual = 8
```

The predictions for 1 and 10 outweighs others due to class imbalance in training data.

d)

```
Processing Training Input Time =  140.55864596366882s
Training time =  4.162050008773804s
Processing Test Input Time =  143.16893196105957s
Training prediction time =  22.801979303359985s
Training accuracy =  66.18%
Test prediction time =  19.1402530670166s
Test accuracy =  39.532%
Random guess prediction accuracy =  12.495%
Most occuring class prediction accuracy =  20.4%
```

There's a slight increase in accuracy in training and test data as expected. This is because we are removing redundant words and words of same type.

e)

In this part I have used the positive and negative words lexicon form nltk's opinion_lexicon. For ratings >= 7, if a word is positive then the I add a very small weight the log answer -

```
ans += (e1 + (label - 7) * e2) * occInLabel
```

Similar treatment for ratings <= 4

```
e1 = 0.0006
e2 = 1.5e-05
Processing training input time =  140.55864596366882s
Training time =  4.162050008773804s
Processing test input time =  143.16893196105957s
Training prediction time =  22.801979303359985s
Training accuracy =  65.11%
Test prediction time =  19.1402530670166s
Test accuracy =  39.728%
Random guess prediction accuracy =  12.495%
Most occuring class prediction accuracy =  20.4%
```

There's a slight increase in test set accuracy but a slight decrease in training accuracy.

2. Support Vector Machine

In this part the support vector machines are trained for recognizing hand - written digits.

b)

```
c = 1.0
No. of iterations = 1000
Batch size = 100
Test Accuracy =  92.12
Train Accuracy =  95.62
```

c)

```
gamma = 0.05
c = 1.0

Linear Kernel -

      Train Accuracy = 98.795% (19759/20000) (classification)
      Test Accuracy = 92.76% (9276/10000) (classification)

Gaussian Kernel -

      Train Accuracy = 99.92% (19984/20000) (classification)
      Test Accuracy = 97.23% (9723/10000) (classification)
```

d)

Various cross - validation accuracies are -

```
c = 10^-5 Cross Validation Accuracy = 72.645%
c = 10^-5 Cross Validation Accuracy = 72.645%
c = 1.0   Cross Validation Accuracy = 97.485%
c = 5.0   Cross Validation Accuracy = 97.575%
c = 10.0  Cross Validation Accuracy = 97.575%
```

We select c for which cross validation accuracy is highest. We select c = 5.0 and then train gaussian model.

e)

	Pred: 0	Pred: 1	Pred: 2	Pred: 3	Pred: 4	Pred: 5	Pred: 6	Pred: 7	Pred: 8	Pred: 9
Actual: 0	9.69	0	0.01	0	0	0.03	0.04	0.01	0.02	0
Actual: 1	0	11.22	0.03	0.02	0.01	0.02	0.02	0	0.02	0.01
Actual: 2	0.04	0	10	0.04	0.02	0	0.01	0.06	0.15	0
Actual: 3	0	0	0.08	9.85	0	0.04	0	0.07	0.05	0.01
Actual: 4	0.01	0	0.04	0	9.62	0	0.05	0	0.02	0.08
Actual: 5	0.02	0	0.03	0.06	0.01	8.66	0.07	0.01	0.05	0.01
Actual: 6	0.05	0.04	0	0	0.03	0.04	9.4	0	0.02	0
Actual: 7	0.01	0.04	0.2	0.02	0.03	0	0	9.86	0.02	0.1
Actual: 8	0.04	0	0.03	0.1	0.01	0.05	0.03	0.03	9.42	0.03
Actual: 9	0.04	0.04	0.03	0.08	0.09	0.04	0	0.09	0.11	9.57

It is high for all diagonal elements indicating higher accuracy. However, it is significant for the following cases -

```
Actual = 3 Predicted = 2
Actual = 4 Predicted = 9
Actual = 9 Predicted = 3
Actual = 9 Predicted = 4
```

Images of some misclassified examples -

Actual = 4 Predicted = 6

Actual = 6 Predicted = 0

Actual = 9 Predicted = 4