Student(s) Name: Engincan Varan 25050

CS412 Machine Learning
HW 3 – Text Classification: Logistic Regression and Naive Bayesian
100pts

- **Please TYPE your answer.**
- **Use this document to type in your answers** (rather than writing on a separate sheet of paper), to keep questions, answers and grades together so as to facilitate grading.
- **SHOW all your work for partial/full credit.**

Goal:

1. By using gaussian distributed artificial dataset with two cluster, makes the decision boundary and conditional independence assumption clearer.
2. The dataset contains around 200k news headlines from the year 2012 to 2018 obtained from HuffPost, make a classification of 5 hot topics by Naive Bayesian and Logistic Regression.

Grading: The algorithmic parts needs to be supported by discussions. In both parts of the homework, it is very important to discuss Naive Bayesian and Logistic Regression differences. The aim here is to make sure that you can follow a good ML experimental methodology (as taught in HW1); know the weaknesses/strengths and requirements of each classifier for a given problem and that you are able to assess and report your results clearly and concisely.

Data:

1. It is expected to generate two artificial datasets. In each of the data points, they are drawn from Gaussian distributions with different standard deviations.
2. This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from HuffPost. Politics, Wellness, Entertainment and Travel topics are selected for processing. Split in two subsets: one for training (or development) and the other one for testing (or for performance evaluation). The split between the train and test set is based upon a messages posted before and after a specific date.

Software: You may find the necessary function references here:

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

Submission: Fill and submit this document with a link to your Colab notebook (make sure to include the link obtained from the **share link on top right**)

Please follow the instructions of the notebook:

https://colab.research.google.com/drive/1tkKUs1MmR0sMW3OXnfD-3B3upMZ61zJD

Question 1) 25pts – Use an artificial dataset to clarify decision boundary and conditional independence assumption.

a) 10pts - What is the test set performance for Naive Bayesian and Logistic Regression with different standard deviation? Print the confusion matrix, classification report.

Since we are using an artificial dataset, the performance of both Naïve Bayesian and Logistic Regression approaches are nearly the same in terms of time and accuracy. We have a slight decrease when the standard deviation is 5 since we add some bias and we have the variance*(sqrt(standard deviation))*-bias trade-off. However, this decrease is not too much. You can see the performances and the results of both approaches.

Standard Deviation = 1

Gaussian Naïve Bayesian Reports

Classification Report:

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        53
           1       1.00      1.00      1.00        47

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```

Confusion Matrix:

```
[[53  0]
 [ 0 47]]
```

Logistic Regression Reports

Classification Report:

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        53
           1       1.00      1.00      1.00        47

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```

Confusion Matrix:

```
[[53  0]
 [ 0 47]]
```

Standard Deviation = 5

Gaussian Naïve Bayesian Reports

Classification Report:

```
              precision    recall  f1-score   support

           0       0.98      0.91      0.95        57
           1       0.89      0.98      0.93        43

    accuracy                           0.94       100
   macro avg       0.94      0.94      0.94       100
weighted avg       0.94      0.94      0.94       100
```

Confusion Matrix:

```
[[52  5]
 [ 1 42]]
```

Logistic Regression Reports

Classification Report:

```
              precision    recall  f1-score   support

           0       0.96      0.93      0.94        55
           1       0.91      0.96      0.93        45

    accuracy                           0.94       100
   macro avg       0.94      0.94      0.94       100
weighted avg       0.94      0.94      0.94       100
```

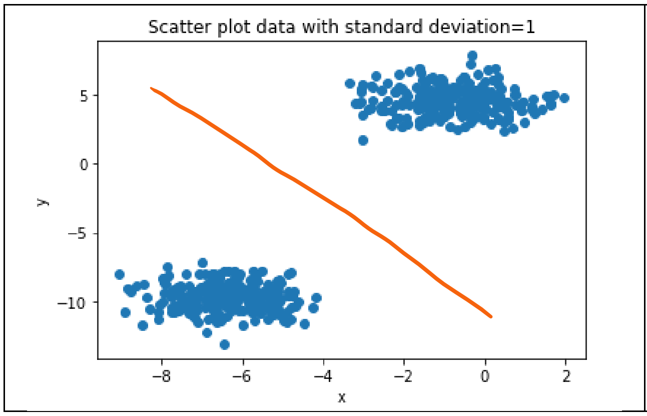
Confusion Matrix:

```
[[51  4]
 [ 2 43]]
```

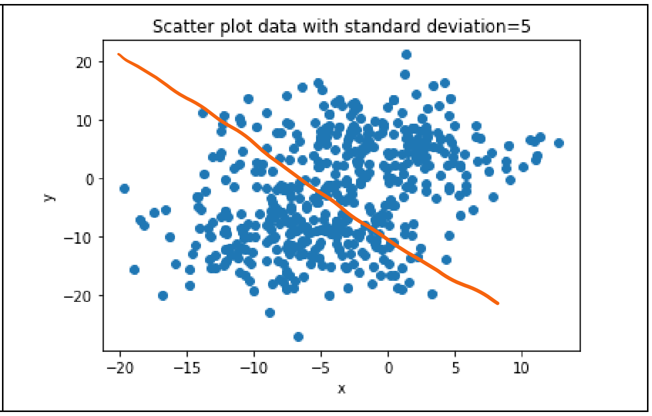
b) 10pts - Discuss the reason behind why Gaussian Naive Bayesian works better for artificial dataset with the concept of conditional independence.

We created this data artificially by using a method called make_blobs which creates a dataset using Gaussian Distribution. Since this dataset is an artificial dataset, we are sure that our data is conditionally independent. So, Gaussian Naïve Bayesian classifier works a bit better than the Logistic Regression classifier, however, in real life there usually is conditional dependency.

c) 5pts - Draw the perfect decision boundary for the dataset on the scatter plots.



Scatter plot data with standard deviation=1



Scatter plot data with standard deviation=5

Question 2) 20pts – Use a Gaussian Naive Bayesian

Import Kaggle dataset and filter 4 principle topics, Politics, Wellness, Entertainment and Travel. Sample 50000 rows from the data. The occurrences of the topics,

POLITICS	8246
WELLNESS	4352
ENTERTAINMENT	3951
TRAVEL	2426

Merge the `short description and headline` cells of the corresponding row to use as text to process.

a) 15pts - What is the best test set performance you obtained by Gaussian Naive Bayesian?

b) 5pts – Print the confusion matrix, classification report.

As you can see from the report, the accuracy of the Gaussian Naïve Bayesian classifier is **%71**. However, the time it required was pretty quick compared to logistic regression, which is what we expect.

Gaussian Naïve Bayesian Report

Classification Report:

```
              precision    recall  f1-score    support

           0      0.76      0.77      0.77      1576
           1      0.69      0.69      0.69       881
           2      0.68      0.68      0.68       822
           3      0.61      0.57      0.59       525

    accuracy                          0.71      3804
   macro avg      0.69      0.68      0.68      3804
weighted avg      0.71      0.71      0.71      3804
```

Confusion Matrix:

```
[ [1221  141  129   85]
  [ 134  608   64   75]
  [ 173   58  561   30]
  [  81   78   67  299] ]
```

Question 2) 20pts – Use a Logistic Regression

Import Kaggle dataset and filter 4 principle topics, Politics, Wellness, Entertainment and Travel. Sample 50000 rows from the data. The occurrences of the topics,

POLITICS	8246
WELLNESS	4352
ENTERTAINMENT	3951
TRAVEL	2426

Merge the `short description and headline` cells of the corresponding row to use as text to process.

a) 15pts - What is the best test set performance you obtained by Logistic Regression?

b) 5pts – Print the confusion matrix, classification report.

As you can see from the report, the accuracy of the Logistic Regression classifier is **%89**. However, the time it required is huge since we have a lot data and logistic regression takes some time, which is what we expect.

Logistic Regression Report

Classification Report:

```
              precision    recall  f1-score   support

           0       0.96      0.88      0.92      1747
           1       0.91      0.89      0.90       900
           2       0.82      0.89      0.85       759
           3       0.74      0.91      0.82       398

    accuracy                           0.89      3804
   macro avg       0.86      0.89      0.87      3804
weighted avg       0.90      0.89      0.89      3804
```

Confusion Matrix:

```
[ [1539   45  114   49]
  [  33  803   22   42]
  [  29   24  672   34]
  [   8   13   13  364] ]
```

Question 4) 35pts – Report

Write a 3-4 lines summary of your work at the end of your notebook; this should be like an abstract of a paper (you aim for clarity and passing on information, not going to details about know facts such as what logistic regression are or what dataset is, assuming they are known to people in your research area).

> “We evaluated the performance of Logistic Regression and Bayes classifiers (Gaussian Naïve Bayes and Gaussian Bayes with general and shared covariance matrices) on the 4 topics of news dataset.
>
> We have obtained the best results with the ….. classifier , giving an accuracy of …% on test data….
>
> You can also comment on the second-best algorithm, or which algorithm was fast/slow in a summary fashion; or talk about errors or confusion matrix for your best approach.

Don’t forget to discuss, Naive Bayesian and Logistic Regression with the concept of conditional independence and decision boundary.

Note: You will get full points from here as long as you have a good (enough) summary of your work, regardless of your best performance or what you have decided to talk about in the last few lines.

We tested a real-life dataset taken from HuffPost on 4 different topics of news such as politics, wellness, entertainment, travel. We use 2 different classifiers and compare them based on their performances. Since we are using real life data and not an artificial data created by Gaussian distribution, Logistic Regression’s accuracy was way over the Gaussian Naïve Bayesian approach. This is what we expected because real life data are not conditionally independent, so Naïve Bayesian approach is not very helpful in our case.

We obtained the best results with the Linear Regression, which gives the accuracy of 89% on test data. Naïve Bayesian approach had 71% accuracy on test data since it has more general decision boundary. However, these results were based on accuracy of the classifiers. When we compare the time complexities of the classifiers, Logistic Regression took such a long time that we get an “iteration limit reached” error. We can solve this by letting it do more iterations if we want. We had a pretty big dataset yet; Naïve Bayesian approach was instant. In the end, Logistic Regression is undoubtedly giving the best results with a difference of nearly 20% since it has more complex algorithm and the decision boundary it generates is more specific. However, it took a really long time to do that.

Link to your Colab notebook (obtained via the share link in Colab):

Link to my Colab Notebook

https://colab.research.google.com/drive/1gzYIxQXWXD4Fm-o0XcpKLQOeq40v8fWS