Student(s) Name: Alper Bingol

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 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 a 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.

For Standard Deviation: 1

```
Classification Report for Naive Bayesian:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        28
           1       1.00      1.00      1.00        41
           2       1.00      1.00      1.00        31

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

Confusion matrix:

```
[[28  0  0]
 [ 0 41  0]
 [ 0  0 31]]
```

Test Score:100

```
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        28
           1       1.00      1.00      1.00        41
           2       1.00      1.00      1.00        31

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

Confusion matrix:

```
[[28  0  0]
 [ 0 41  0]
 [ 0  0 31]]
```

Test Score : 100

For Standard Deviation: 5

Classification Report for Naive Bayesian:

	precision	recall	f1-score	support
0	0.80	0.65	0.71	31
1	0.61	0.71	0.66	28
2	0.88	0.90	0.89	41
accuracy			0.77	100
macro avg	0.76	0.75	0.75	100
weighted avg	0.78	0.77	0.77	100

Confusion matrix:

```
[[20  5  0]
 [ 9 20  4]
 [ 2  3 37]]
```

Test Score: 0.77

Classification Report for Logistic Regression:

	precision	recall	f1-score	support
0	0.76	0.61	0.68	31
1	0.58	0.68	0.62	28
2	0.88	0.90	0.89	41
accuracy			0.75	100
macro avg	0.74	0.73	0.73	100
weighted avg	0.76	0.75	0.75	100

Confusion matrix:

```
[[19  6  0]
 [10 19  4]
 [ 2  3 37]]
```

Test Score:0.75

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

We know that Naive Bayes classifier is a generative model and logistic regression is a discriminative model.

Gaussian Distributed Data fits better to Gaussian Naive Bayesian rather than Logistic Regression.

The Naive Bayes is linear classifier and there is a strong independence condition among features. But the logistic regression is a generative model.

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
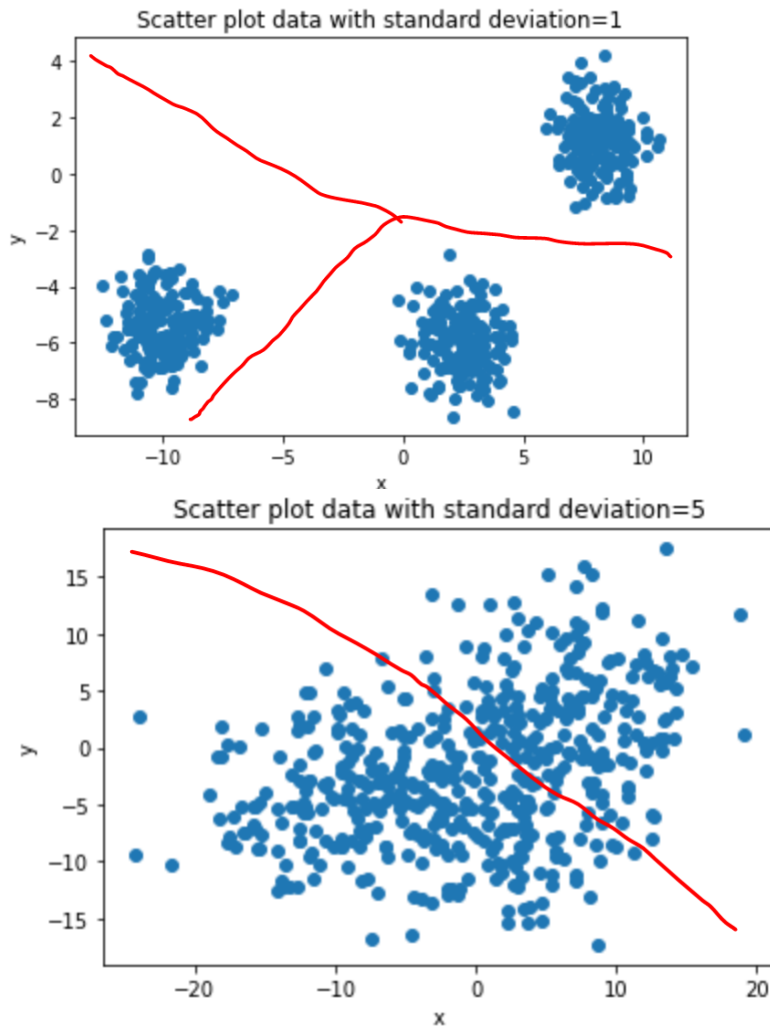
Question 2) 20pts – Use a Gaussian Naive Bayesian

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

Best test performance is 0.77 with class 0 and average accuracy is 0.70.

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

```
Classification Report for Naive Bayesian:
              precision    recall  f1-score   support

           0       0.76      0.78      0.77      1532
           1       0.65      0.71      0.68       853
           2       0.71      0.66      0.69       878
           3       0.63      0.57      0.60       530

    accuracy                           0.71      3793
   macro avg       0.69      0.68      0.68      3793
weighted avg       0.71      0.71      0.70      3793
```

```
Confusion Matrix:

[[1190  114  193   69]
 [ 155  604   66  104]
 [ 119   65  583   57]
 [  68   70   36  300]]

Test Score: 0.71
```

Question 3) 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 occurances 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?

Best Test Score is 0.92 for class 0 and average accuracy score is 0.89.

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

Classification Report for Logistic Regression:
              precision   recall f1-score   support

           0   0.96   0.89   0.92   1688
           1   0.89   0.89   0.89   925
           2   0.85   0.89   0.87   788
           3   0.76   0.92   0.83   392

    accuracy   0.89   3793
   macro avg   0.86   0.90   0.88   3793
weighted avg   0.90   0.89   0.89   3793

Confusion matrix:

[[1503   25   29   9]
 [66 825   26   12]
 [81   33 698   12]
 [38   42   35 359]]

Test Score:0.89

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.

Link to your Colab notebook (obtained via the <u>share link in Colab</u>):

https://colab.research.google.com/drive/1w30_a6PJYxHRzowmB0EUkm9975oLsPt1x