Student(s) Name: Efe Şencan

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 inde pendence 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.

Both Naïve Bayesian and Logistic Regression performs with 100% accuracy for the first artificial dataset with standard deviation 1. The reason Naïve Bayesian works perfectly well with the artificial dataset is because, the dataset's features are conditionally independent. For the second artificial dataset with standard deviation 5, both algorithms accuracy was decreased since the variance of the dataset increased, but they still have very high accuracy in the test set.

Naïve Bayes Result with first dataset

```
Classification Report for Naive Bayesian:
              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

Accuracy score for Naive Bayes Classifier for the first dataset:
1.0
Confusion matrix for Naive Bayes Classifier:
[[53  0]
 [ 0 47]]
```

Logistic Regression result with first dataset

```
Classification Report for Logistic Regression:
              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

Accuracy score for Logistic Regression for the first dataset:
1.0
Confusion matrix for Logistic Regression:
[[53  0]
 [ 0 47]]
```

Naive Bayes Result with second dataset

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

           0       0.91      0.91      0.91        47
           1       0.92      0.92      0.92        53

    accuracy                           0.92       100
   macro avg       0.92      0.92      0.92       100
weighted avg       0.92      0.92      0.92       100

Accuracy score for Naive Bayes for the second dataset:
0.92
Confusion matrix for Naive Bayesian:
[[43  4]
 [ 4 49]]
```

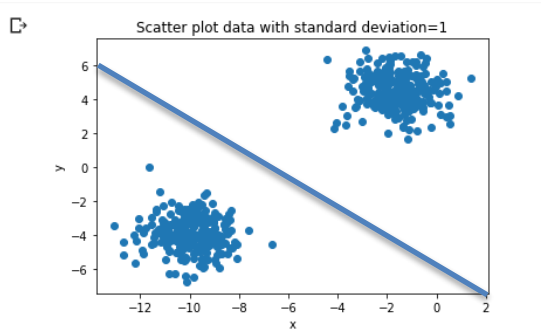
Logistic Regression result with second dataset

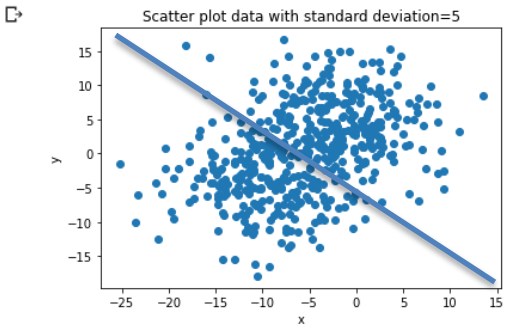
```
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

           0       0.91      0.91      0.91        47
           1       0.92      0.92      0.92        53

    accuracy                           0.92       100
   macro avg       0.92      0.92      0.92       100
weighted avg       0.92      0.92      0.92       100

Accuracy score for Logistic Regression for the second dataset:
0.92
Confusion matrix for Logistic Regression:
[[43  4]
 [ 4 49]]
```

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

GNB algorithm assumes that all the features in dataset are conditionally independent from each other and since the features are conditionally independent meaning that there is no correlation between the variables and they are Gaussian distributed in the artificial dataset, Gaussian Naïve Bayesian algorithm's prediction score is really high on that dataset.

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





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?

The best test performance was nearly 70.93% by Gaussian Naïve Bayesian algorithm. The accuracy performance of the algorithm was decreased compared to the artificial dataset since the features are not conditionally independent in the Kaggle dataset.

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

```
Accuracy score for Naive Bayes for the kaggle dataset:
0.7093389296956978
Classification Report for Naive Bayesian:
              precision    recall  f1-score   support

           0       0.79      0.78      0.78      1623
           1       0.70      0.71      0.70       917
           2       0.65      0.68      0.67       757
           3       0.57      0.54      0.56       515

    accuracy                           0.71      3812
   macro avg       0.68      0.68      0.68      3812
weighted avg       0.71      0.71      0.71      3812

Confusion matrix for Naive Bayes Classifier:
[[1259  101  167   68]
 [ 155  650   43   87]
 [ 123   72  515   80]
 [  86   94   32  280]]
```

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 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?

The best test performance of the Logistic Regression was 89.50%. The algorithm's accuracy score was really high since it assigns weight to the features and tries to determine relation between the data points. Hence, its prediction result is better than GNB.

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

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

           0       0.95      0.90      0.92      1685
           1       0.90      0.90      0.90       945
           2       0.86      0.87      0.87       780
           3       0.75      0.92      0.83       402

    accuracy                           0.90      3812
   macro avg       0.87      0.90      0.88      3812
weighted avg       0.90      0.90      0.90      3812

Accuracy score for Logistic Regression for the kaggle dataset:
0.8950682056663168
Confusion matrix for Logistic Regression Classifier:
[[1516   30   35   14]
 [  48  846   28   13]
 [  78   27  680    5]
 [  43   42   37  370]]
```

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 my Colab notebook (obtained via the share link in Colab):

https://colab.research.google.com/drive/1XuFInknE_7T3aVYoaVv7QQ3HgBz0hyk7

Report:

We evaluated the performance of Logistic Regression and Bayes classifiers on the 4 topics of news dataset

We have obtained the best results with the Logistic Regression classifier, giving an accuracy of 89.50% on test data. The second best algorithm was Gaussian Naive Bayes Classifier giving an accuracy of 70.93% on test data. However, Gaussian Naive Bayes algorithm works faster than the Logistic Regression since it assumes that all the features are conditionally independent from each other and calculates the posterior probability accordingly. On the other hand, Logistic Regression assign weights to every feature and it is more successful for determining the correlation between the features. Hence, it has more complex model and decision boundary compared with the GNB algorithm. However, it is slower than GNB method and it is more likely to overfit. You can access the results of the confusion matrix in the above cells.

As the dataset size increases, it is more likely that some features are dependent to each other. Therefore, it is more logical to use Logistic Regression Classifier to get high accuracy scores. However, if you have a dataset in which all the features are conditionally independent from each other, then Gaussian Naive Bayes algorithm would be a better choice since it would work fast and produce very accurate prediction scores.