Student(s) Name: Onur Arda Bodur

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**

Share link: https://colab.research.google.com/drive/1mekBl5UozH05cG91C0L3mFZSpqhH5HTM

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.

Test Performance for both models are showing good results when the std = 1. When std = 1 is increased to std = 5 the test performance shows a decline. Precision checks how much of the returned results are correct and recall checks that from all relevant(correct results) how much of them is returned. For explaining the performance evaluations of the models , Precision and Recall is used in this part of the Question 1.

GNB STD=1;

When std = 1, it can be seen that the average of precision and recall are both over 0.90 which means that the model returns many results with all results labeled correctly.

```
Confusion Matrix for Naive Bayesian:
[[48  0  0]
 [ 0 44  5]
 [ 0  6 62]]

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

           0       1.00      1.00      1.00        48
           1       0.88      0.90      0.89        49
           2       0.93      0.91      0.92        68

    accuracy                           0.93       165
   macro avg       0.94      0.94      0.94       165
weighted avg       0.93      0.93      0.93       165
```

LR STD=1;

When std = 1, it can be seen that the average of precision and recall are both over 0.90 which means that the model returns many results with all results labeled correctly.

```
Confusion Matrix for Logistic Regression:
[[48  0  0]
 [ 0 44  5]
 [ 0  6 62]]

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

           0       1.00      1.00      1.00        48
           1       0.88      0.90      0.89        49
           2       0.93      0.91      0.92        68

    accuracy                           0.93       165
   macro avg       0.94      0.94      0.94       165
weighted avg       0.93      0.93      0.93       165
```

GNB STD=5;

When std = 5, it can be seen that the average of precision and recall are both 0.64 which means that the model returns many results with all results labeled correctly but worse than the previous std value. Increasing the std results the data to be more spread causing the model to choose the boundaries incorrectly. In make_blobs case when the std is increased the data points spread more causing some points to be inside wrong decision boundaries. This does not have to mean a bad condition since in real life such data can occur as an outlier.

```
Confusion Matrix for Naive Bayesian:
[[42  0  4]
 [ 5 25 27]
 [ 7 17 38]]

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

           0       0.78      0.91      0.84        46
           1       0.60      0.44      0.51        57
           2       0.55      0.61      0.58        62

    accuracy                           0.64       165
   macro avg       0.64      0.65      0.64       165
weighted avg       0.63      0.64      0.63       165
```

LR STD=5;

When std = 5, it can be seen that the average of precision and recall are both 0.65 which means that the model returns many results with all results labeled correctly. Logistic Regression performs slightly better but takes more time to compute.

```
Confusion Matrix for Logistic Regression:
[[42  0  4]
 [ 5 25 27]
 [ 7 17 38]]

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

           0       0.79      0.91      0.85        46
           1       0.58      0.54      0.56        57
           2       0.58      0.55      0.56        62

    accuracy                           0.65       165
   macro avg       0.65      0.67      0.66       165
weighted avg       0.64      0.65      0.64       165
```

Confusion matrix and classification report is also printed in Colab file.

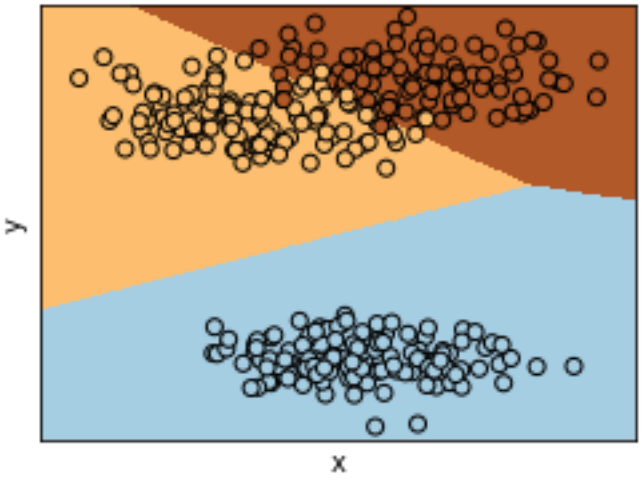
b) 10pts - Discuss the reason behind why Gaussian Naive Bayesian works better for artificial dataset with the concept of conditional independence.

Logistic Regression calculates the posterior probability by directly mapping from dataset where Naïve Bayes calculates the posterior probability by using Bayes rule. Gaussian Naïve Bayes classifier assumes that the effect of the value of Predictor Prior probability on a given class is **independent**. This process of assumption is called as the **conditional independence**. In the artificial dataset, this assumption holds and causes the Gaussian Naïve Bayes model to perform well. In normal life, datasets are not likely to display conditional independence and cause the performance of Gaussian Naïve Bayes to decline. Since Logistic Regression is a linear model, it might perform worse than normal when nonlinear data is used.

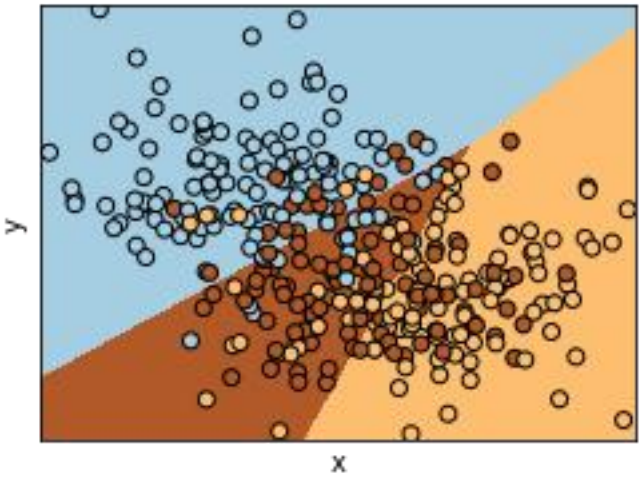
When the data is conditionally independent both Naïve Bayes and logistic regression learns same parameters but Naïve Bayes can be modified later.

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

Decision Boundary from STD=1 Logistic Regression



Decision Boundary from STD=1 Logistic Regression.



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?

Gaussian Naïve Bayes has 0.68 precision(averaged) and 0.69 recall(averaged).GNB model gives best test performance on class 0 where it can be seen in the classification report. The main problem in Gaussian Naïve Bayes is that it assumes all the classes are conditionally independent but that is less likely to happen. In text data lots of words are correlated to each other which causes the model to work less effectively.

O(logN)

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

```
Confusion Matrix for Naive Bayesian:
[[2106  170  286  108]
 [ 230 1005  101  104]
 [ 209   87  948  115]
 [ 142  110   70  525]]
```

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

           0       0.78      0.79      0.79      2670
           1       0.73      0.70      0.71      1440
           2       0.67      0.70      0.69      1359
           3       0.62      0.62      0.62       847

    accuracy                           0.73      6316
   macro avg       0.70      0.70      0.70      6316
weighted avg       0.73      0.73      0.73      6316
```

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?

Logistic Regression shows better performance than Gaussian Naïve Bayes since the model is built using mapping rather than trying to count as individually. Logistic Regression is more resistant to conditionally dependent cases.

It can be seen that model shows more over 0.80 for each label and each metric such precision and recall.Logistic Regression works with a very high performance regarding to results but it is time consuming.

O(N)

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

```
Confusion Matrix for Logistic Regression:
[[2550   60   47   13]
 [  93 1284   47   16]
 [ 161   55 1131   12]
 [  88   66   57  636]]


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

           0       0.88      0.96      0.92      2670
           1       0.88      0.89      0.88      1440
           2       0.88      0.83      0.86      1359
           3       0.94      0.75      0.83       847

    accuracy                           0.89      6316
   macro avg       0.89      0.86      0.87      6316
weighted avg       0.89      0.89      0.89      6316
```

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 share link in Colab): :
https://colab.research.google.com/drive/1mekBl5UozH05cG91C0L3mFZSpqhH5HTM