Student(s) Name: Boran İslamoğlu 24205

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.

When st.deviation is 1, test set performance is perfect. %100 Accuracy for both.

When st.deviation is test set performance is very good. %92 precision and recall for both GNB and LR.

Confusion matrixes and Classification reports are printed in my notebook.

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

Because, the artificial data is conditionally independent but the datasets from real life are usually conditionally dependent on some ways. Therefore, if we assume that the data are conditionally independent then Gaussian Naïve Bayesian will work better for artificial datasets because of the fact that Gaussian Naïve Bayesian assumes that all of the data are conditionally independent and artificial dataset is already conditionally independent too. In conclusion, Gaussian Naïve Bayesian does not make any wrong assumptions if we use an artificial dataset which is already conditionally independent.

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?

%70 Accuracy Rate and %71 Precision and Recall scores.

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

Confusion matrices and Classification reports are printed in my notebook.

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?

%89 Accuracy rate and %86 Precision Rate and %90 Recall rate.

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

Confusion matrixes and Classification reports are printed in my colab notebook.

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 observed that Logistic Regression worked much slower than Gaussian Naive Bayesian.

However, we have obtained much better accuracy, precision and recall scores with Logistic Regression than Gaussian Naive Bayesian.

We have obtained the best results with the Logistic Regression, giving an accuracy of 0.8924617817606747

We have obtained the second best results with the Gaussian Naive Bayesian, giving an accuracy of 0.7090142329994729

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

https://colab.research.google.com/drive/1hd43uBpwCyvIzTocUizmcSQCl1l9_6qi