Umair Afzal
HW2
CS589

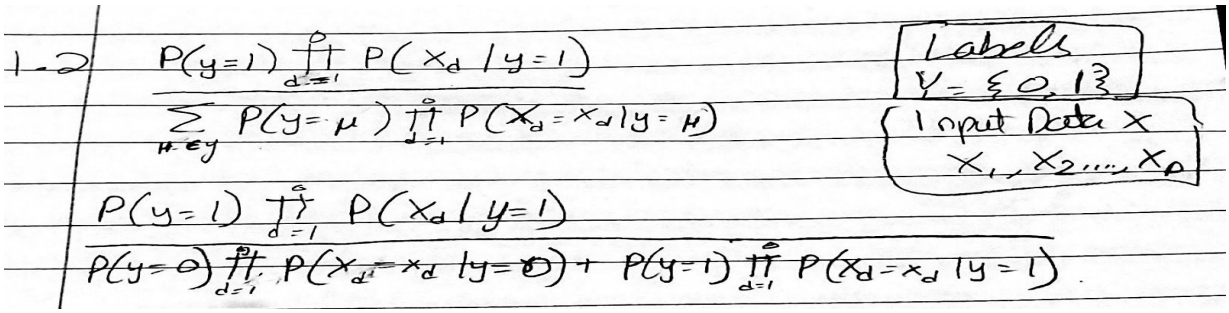
Collaborators:
Debankita,
Bill

1. Nave Bayes and Bag of Words [30 points]

1.

'the', 'and', 'it', 'is', 'this', 'to', 'in', 'but', 'with', 'of'
357, 203, 181, 129, 116, 106, 99, 74, 65, 49

2.



3.

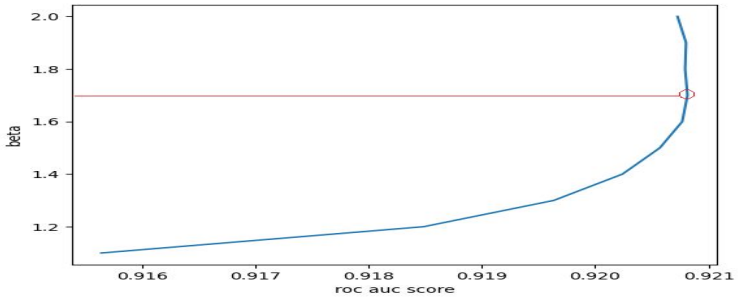
--- Beta = 1.7 ---
Confusion Matrix

Predicted	0	1
True		
0	336	44
1	312	2607

ROC AUC score = 0.920815077261499
f1 score = 0.9360861759425493
accuracy = 89.20885116702031 percent
Recall Score = 0.8931140801644398
Precision Score = 0.983402489626556

Full list of scores available in the “output_from_nb.txt” in the submission folder

4. Getting max ROC AUC when beta is 1.7 after which it starts to decrease



2. Probabilistic Classication [20 points + 5 Extra Credit]

1.

P(A)P(B) = (P(B|A) * P(A)) / P(B)
P(B|A) * P(A)

2.

The parameters needed to estimate to construct the Bayes optimal Classifier are the ones that do not take into consideration the prediction strength of the variables.

3.

The MLE estimates for these parameters as a function of the data are the ones with estimated normal distribution and variance and likelihood functions. Possibly maximize the likelihood function when calculating prob distribution. Also, when the sample size is more we get unbiased result and smallest variance.

4.

Risk can be loss of information as it is provided in shorthand form. In order to mitigate the risks, we could assume a probability and independence of variables of our data.

5.

They could be Discrete Random Variable, max likelihood, posterior distribution

3. Logistic Regression Revisited [40 points]

1.

regularization_weights = 0, 0.05, 0.1, 0.2
learning_rates = 10^-4, 10^-1, 1, 10
num_iterations = 10, 750, 1000, 1500

2.

Green Line = validation-set AUC
Magenta Line = train-set AUC
validation data with the best hyper parameters
learning_rate=1
num_epochs=1000
reg=0

3.

Confusion Matrix on the validation data with the best hyper parameters
learning_rate=1
num_epochs=1000
reg=0

Predicted	0	1
True		
0	252	128
1	87	2832

4. Train Your Best Model [10 points]

Metric	Precision	Recall	F1 Score	AUC
Validation Set	0.8848135798726887	1.0	0.9388871019620456	0.5

I chose Logistic Regression as my best model because it incorporates different method for estimating the parameters. This provide us with best results like results that are not biased and have low variances.

Sources:

https://compsci682-fa20.github.io/notes/optimization-1/#analytic
https://philippmuens.com/logistic-regression-from-scratch/
https://docs.google.com/presentation/d/1gQMaX0kU3m8rgK7ltbT2MsuOo3El8dDwSjie2RALWEg/edit#slide=id.p