Name: ROHIT GARLA NET ID:rg3365

1. What was the estimated value of P(C) for C = 1?

Answer: 0.4018006002000667

2. What was the estimated value of P(C) for C = 0?

Answer: 0.5981993997999333

3. What were the estimated values for (μˆ, σˆ2) for the Gaussian corresponding to attribute capital run length longest and class 1 (Spam).

Answer: Variance: 36369.99111261216
Mean: 97.2091286307054

4. What were the estimated values for (μˆ, σˆ2) for the Gaussian corresponding to attribute char freq ; and Class 0.

Answer: Variance: 0.08830560325706198
Mean: 0.0484258639910814 25

5. Which classes were predicted for the first 5 examples in the test set?

Answers: 0,0,0,0,0

6. Which classes were predicted for the last 5 examples in the test set?

Answers: 0,0,0,0,0

7. What was the percentage error on the examples in the test file?

Answers: 20%

8. Sometimes a not-very-intelligent learning algorithm can achieve high ac- curacy on a particular learning task simply because the task is easy. To check for this, you can compare the performance of your algorithm to the performance of some very simple algorithms. One such algorithm just pre- dicts the majority class (the class that is most frequent in the training set). This algorithm is sometimes called Zero-R. It can achieve high accuracy in a 2-class problem if the dataset is very imbalanced (i.e., if the fraction

of examples in one class is much larger than the fraction of examples in the other). What accuracy is attained is you use Zero-R instead of Gaussian Naive Bayes?

Answer:

Accuracy : 0.59

9. Gaussian Naive Bayes is based on two assumptions: (1) the conditional in- dependence assumption, and (2) the assumption that the pdfs for p(xj|C) are Gaussian. These assumptions are more reasonable for some datasets than for others. Do you think these assumptions are reasonable for the spam dataset you just used? Why or why not? In answering this ques- tion, you can give a common-sense argument and/or show relevant plots, graphs, or statistical information.

Answer: Gaussian Naïve Bayes works really well for small datasets. In the case of the spam folder, the number of spam emails would be considerably less than the number of non-spam emails. So the dataset would be quite small in this case and Gaussian Naïve Bayes will work really well. I got an accuracy of 80% which is quite high.

The reason for this is because of the independence of the attributes.

It can perform particularly well in avoiding false positives, where legitimate email is incorrectly classified as spam. For example, if the email contains the word "Nigeria", which is frequently used in Advance fee fraud spam, a pre-defined rules filter might reject it outright. A Bayesian filter would mark the word "Nigeria" as a probable spam word but would take into account other important words that usually indicate legitimate e-mail. For example, the name of a spouse may strongly indicate the e-mail is not spam, which could overcome the use of the word "Nigeria."