MACHINE LEARNING

-Sudhamayi Nanduri

(Discussed with Abhishek Hiregoudar)

Bernoulli Distribution:

Procedure:

- The data is partitioned into 10 sets using k-fold partitioning.
- Using, the k-fold cross validation, the 10 sets are divided into training sets and validation (testing) sets.
- The probability values of spam and non-spam classification of e-mails are calculated.
- These values are validated against the validation (testing) sets to estimate spam and non-spam.
- Applied **Laplace Smoothing** to avoid issues with zero probabilities.

The following observations have been recorded:

K-Partition	False Positive Error Rate	False Negative Error Rate	Overall Error Rate
1	0.0286738351254	0.142857142857	0.0737527114967
2	0.0934782608696	0.0359712230216	0.181318681319
3	0.0717391304348	0.0431654676259	0.115384615385
4	0.0891304347826	0.0430107526882	0.160220994475
5	0.104347826087	0.0501792114695	0.187845303867
6	0.136956521739	0.0824372759857	0.220994475138
7	0.117391304348	0.0716845878136	0.187845303867
8	0.100000000000	0.0573476702509	0.165745856354
9	0.0891304347826	0.0752688172043	0.110497237569
10	0.0891304347826	0.0573476702509	0.138121546961
Average Error Rate	0.09919781829516	0.075541853078	0.0965057059323

Gaussian Distribution:

- Used the below Gaussian distribution formula

$$\mathcal{N}(x|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- The procedure of k-cross validation is similar to Bernoulli's distribution.
- Used **Jelinek-Mercer** smoothing on the variances of spam and non-spam to avoid zero probability issues, by taking a lambda = 0.3

The following observations have been recorded.

K-Partition	False Positive Error Rate	False Negative Error Rate	Overall Error Rate
1	0.123644251627	0.168458781362	0.0549450549451
2	0.158695652174	0.197841726619	0.0989010989011
3	0.115217391304	0.165467625899	0.0384615384615
4	0.117391304348	0.146953405018	0.0718232044199
5	0.163043478261	0.197132616487	0.110497237569
6	0.169565217391	0.21146953405	0.104972375691
7	0.145652173913	0.182795698925	0.0883977900552
8	0.154347826087	0.21146953405	0.0662983425414
9	0.139130434783	0.186379928315	0.0662983425414
10	0.147826086957	0.218637992832	0.0386740331492
Average Error Rate	0.143451381684	0.1886606843557	0.143451381684

Distribution via Histrogram:

- The procedure of k-cross validation is similar to that of Bernoulli and Gaussian Distribution.
- The features are classified into the following 4 bins:
  - [min-value, low-mean-value]
  - (low-mean-value, overall-mean-value]
  - (overall-mean-value, high-mean-value]
  - (high-mean-value, max-value]
- The class conditional probabilities are calculated for the bins to estimate spam and non-spam e-mails.
- Applied **Laplace Smoothing** to avoid zero probability issues.

The following observations have been recorded:

K-Partition	**False Positive Error Rate**	**False Negative Error Rate**	**Overall Error Rate**
1	0.0824295010846	0.0430107526882	0.142857142857
2	0.10652173913	0.0431654676259	0.203296703297
3	0.0739130434783	0.0575539568345	0.0989010989011
4	0.102173913043	0.0501792114695	0.182320441989
5	0.0978260869565	0.0430107526882	0.182320441989
6	0.130434782609	0.0645161290323	0.232044198895
7	0.117391304348	0.0573476702509	0.209944751381
8	0.104347826087	0.0609318996416	0.171270718232
9	0.095652173913	0.0752688172043	0.127071823204
10	0.0934782608696	0.0501792114695	0.160220994475
Average Error Rate	0.1004168631519	0.10251638689049	0.100416863152

From the above tables, we can observe that the Error Rate for Gaussian is highest. The error rates for Bernoulli and Histogram are nearly the same.

Few Other Observations:

1. **Varying the number of partitions:**

   By varying the number of partitions to 5 and 15, we get the following average error rates:

   (i) For k = 5

Distribution	**Average Error Rate**
Bernoulli	0.0960640608035
Gaussian	0.146054147194
Histogram	0.100194023509

   (ii) For k = 15

Distribution	**Average Error Rate**
Bernoulli	0.0969293109933
Gaussian	0.143451633277
Histogram	0.0999744523216

From the above observations, we can see that Bernoulli and Gaussian distributions perform nearly the same while increasing or decreasing the partitions by a difference of 5. The difference in their error rates between k = 5 and k = 15 is nearly the same. We can observe a slight decrease in the error rate for Distribution via Histogram by doing the same. We can conclude that by increasing the number of partitions to an optimal number, we can reduce the error rate, as we are training over more sets of partitions.

2. **No. of records in each bin for training data set:**

   From Distribution via Histogram, we get the following observations:

Bin No.	Bin Property	Percentage of Records
1.	[min-value, low-mean-value]	82%
2.	(low-mean-value, overall-mean-value]	3%
3.	(overall-mean-value, high-mean-value]	3%
4.	(high-mean-value, max-value]	12%

We can deduce from the above table that, the e-mail records from training data set are not evenly distributed into the four bins. Bin one has highest number of records. Hence, the distribution is highly skewed towards low mean value.

But, Gaussian distribution is ideal in a case of normalized data set. Hence, it is not suitable for a skewed data set. We can also validate this observation from the average error rate of Gaussian Distribution, which being the highest among all the three distributions.

3. **Smoothing Observations:**

(i) For Bernoulli distribution, the average error rate without Laplace Smoothing is 0.0968883146279
But, the average error rate with Laplace Smoothing is 0.0965057059323
Hence, we can observe that it performs slightly better with Laplace smoothing.

(ii) For Gaussian distributions, the following values have been observed by varying the value of lambda, between 0.9 to 0.3

Lambda Value	**Average Error Rate**
0.9	0.262762897293
0.6	0.260808261813
0.3	0.143451381684

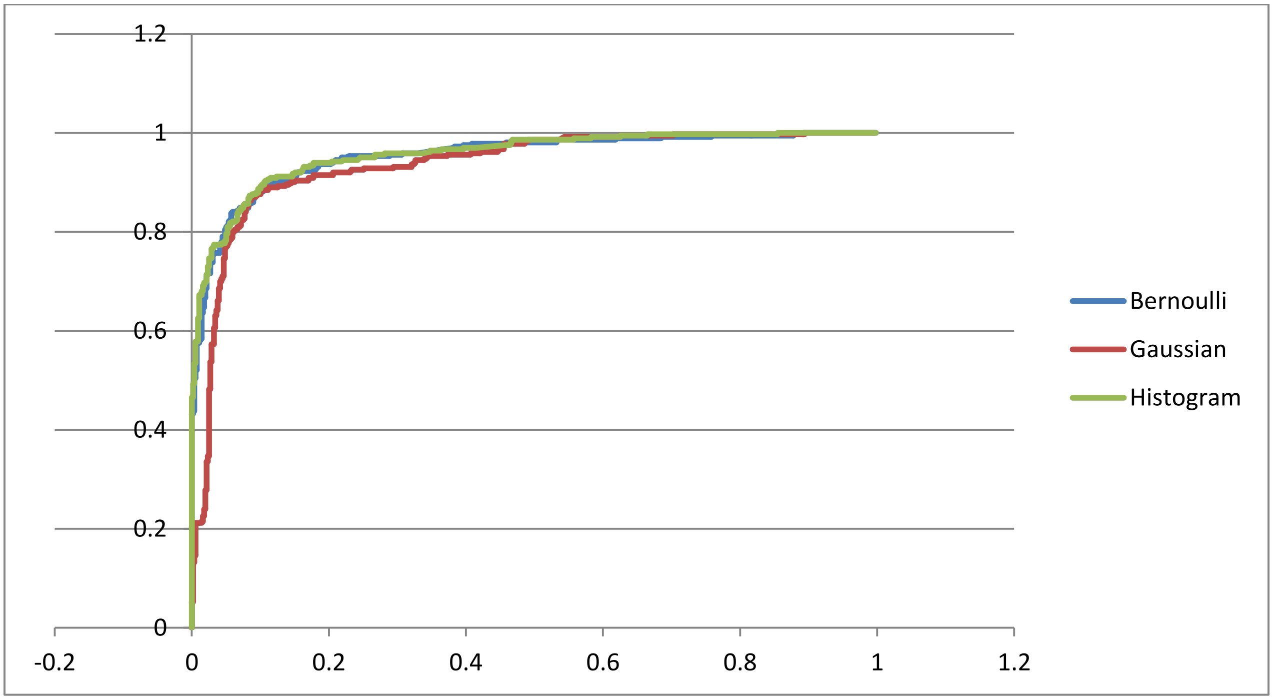
Hence, we can conclude that a higher value of lambda gives a higher error rate. An optimal error rate is obtained at lambda = 0.3

(iii) For Histogram, the average error rate without Laplace Smoothing is 0.103416863152
But, the average error rate with Laplace Smoothing is 0.100416863152

Thus, smoothing helps avoid zero probability issues and gives lower error rate.

Plot for ROC:

The following is the plot showing ROC curves for Bernoulli, Gaussian and Histogram Distributions.



Area Under Curve (AUC):

The following observations are recorded for the area under curve for each ROC plot.

Distribution	**AUC**
Bernoulli	0.961489227618
Gaussian	0.947181850408
Histogram	0.961971720036

We can see that the AUC value of Gaussian is less compared to Bernoulli and Histogram. An ideal value of AUC should be 1. Hence, this observation validates that Gaussian distribution is not suitable for this data set.

Conclusion:

From all the above observations made, we can conclude that Bernoulli and Histogram distributions perform similarly on the data. But, Gaussian distribution has a higher average error rate and a lower AUC value. Hence, it is not as suitable as the other distributions for this data set.