COMS 573: LAB REPORT 1

Saunak Saha | saha@iastate.edu

2.1 Learning the Naïve Bayes Model

1. Class Priors:

P(Omega=1) = [0.04259473]

P(Omega=2) = [0.05155737]

P(Omega=3) = [0.05075872]

P(Omega=4) = [0.0520898]

P(Omega=5) = [0.05102494]

P(Omega=6) = [0.0525335]

P(Omega=7) = [0.05164611]

P(Omega=8) = [0.0525335]

P(Omega=9) = [0.05288846]

P(Omega=10) = [0.05271098]

P(Omega=11) = [0.05306593]

P(Omega=12) = [0.05271098]

P(Omega=13) = [0.05244476]

P(Omega=14) = [0.05271098]

P(Omega=15) = [0.05262224]

P(Omega=16) = [0.05315467]

P(Omega=17) = [0.04836277]

P(Omega=18) = [0.05004881]

P(Omega=19) = [0.0411749]

P(Omega=20) = [0.03336587]

2. Bayesian Estimators vs Maximum Likelihood Estimators:

Maximum Likelihood estimators for words w1 to w9 for classes ω1 to ω9

```
[[8.73585464e-05 5.43685098e-04 1.21189419e-04 8.06890848e-05
  6.96136443e-05 3.07499051e-04 0.00000000e+00 7.88767944e-05
  1.36411026e-04]
 [4.23352955e-04 5.34623679e-04 7.60188174e-04 3.12670204e-04
  3.82875044e-04 1.45244233e-03 4.58310145e-04 4.73260767e-04
  6.52824195e-04]
 [1.84796925e-03 0.00000000e+00 0.00000000e+00 0.00000000e+00
  0.00000000e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00
  0.00000000e+00]
 [6.04789936e-05 1.54044111e-04 1.87292739e-04 0.00000000e+00
  1.16022740e-05 5.16860108e-04 3.27364389e-05 0.00000000e+00
  3.89745788e-05]
 [5.51030831e-04 1.26859856e-04 2.31361618e-04 1.00861356e-04
  1.16022740e-05 9.81379951e-05 3.27364389e-05 1.13933148e-04
  3.89745788e-05]
 [2.75515415e-04 5.25562261e-04 3.74585477e-04 4.84134509e-04
  5.45306880e-04 3.40211716e-04 5.23783023e-04 7.53711591e-04
  8.08722511e-04]
 [4.03193291e-05 9.06141829e-05 2.20344398e-05 2.01722712e-05
  1.16022740e-05 1.37393193e-04 0.00000000e+00 6.13486179e-05
  2.92309341e-05]
 [6.71988818e-06 5.43685098e-05 2.20344398e-05 1.00861356e-05
  0.00000000e+00 1.96275990e-05 0.00000000e+00 2.62922648e-05
  9.74364471e-06]
 [2.28476198e-04 1.55856395e-03 1.58647967e-03 4.84134509e-04
  5.45306880e-04 1.55058032e-03 4.58310145e-04 1.57753589e-04
  3.89745788e-05]]
```

Bayesian estimators for words w1 to w9 for classes ω1 to ω9

```
[[6.66666667e-05 3.55589754e-04 7.89707479e-05 5.61328227e-05
  4.74969127e-05 2.24263435e-04 8.17781849e-06 5.70483199e-05
  9.15644705e-05]
 [3.04761905e-04 3.49760414e-04 4.60662696e-04 1.99583370e-04
  2.30699290e-04 1.04189054e-03 2.37156736e-04 3.13765760e-04
  4.15092266e-04]
 [1.31428571e-03 5.82934024e-06 6.58089566e-06 6.23698030e-06
  6.78527324e-06 4.67215489e-06 8.17781849e-06 5.70483199e-06
  6.10429804e-06]
 [4.76190476e-05 1.04928124e-04 1.18456122e-04 6.23698030e-06
  1.35705465e-05 3.73772391e-04 2.45334555e-05 5.70483199e-06
  3.05214902e-05]
 [3.95238095e-04 8.74401035e-05 1.44779705e-04 6.86067833e-05
  1.35705465e-05 7.47544783e-05 2.45334555e-05 7.98676479e-05
  3.05214902e-05]
 [2.00000000e-04 3.43931074e-04 2.30331348e-04 3.05612035e-04
  3.25693116e-04 2.47624209e-04 2.69868010e-04 4.96320383e-04
  5.12761035e-04]
 [3.33333333e-05 6.41227426e-05 1.97426870e-05 1.87109409e-05
  1.35705465e-05 1.02787408e-04 8.17781849e-06 4.56386559e-05
  2.44171921e-05]
 [9.52380952e-06 4.08053816e-05 1.97426870e-05 1.24739606e-05
  6.78527324e-06 1.86886196e-05 8.17781849e-06 2.28193280e-05
  1.22085961e-05]
 [1.66666667e-04 1.00847586e-03 9.54229871e-04 3.05612035e-04
  3.25693116e-04 1.11197286e-03 2.37156736e-04 1.08391808e-04
  3.05214902e-05]]
```

We notice that Maximum Likelihood estimators attain value 0 whenever there are no appearances of a particular word in any document of a given class. (i.e. $n_k=0$). This causes the entire product to be zero or in case of logarithms, a mathematical domain error (i.e. negative infinite). This is rectified as seen above, by the use of Bayesian estimators. (corresponding elements of estimator matrix are not zero but $1/n+|V|$)

2.2 Performance of the Classifier

2.2.1 Evaluation on Training data

The classifier uses only the Bayesian Estimators when evaluating performance on the training data and outputs the Overall Accuracy, Class Accuracy values and the confusion matrix.

1. **Overall Accuracy** = 0.9481764131688704 or **94.81764131688703 %**

2. **Class Accuracy:**

```
Class1: 0.98125
Class2: 0.9242685025817556
Class3: 0.8916083916083916
Class4: 0.9318568994889267
Class5: 0.9547826086956521
Class6: 0.9408783783783784
Class7: 0.8127147766323024
Class8: 0.9628378378378378
Class9: 0.9714765100671141
Class10: 0.9747474747474747
Class11: 0.9782608695652174
Class12: 0.9814814814814815
Class13: 0.9323181049069373
Class14: 0.9764309764309764
Class15: 0.9814502529510961
Class16: 0.9849749582637729
Class17: 0.9889908256880734
Class18: 0.9716312056737588
Class19: 0.9676724137931034
Class20: 0.8138297872340425
```

3. Confusion Matrix:

```
[[471.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0   0.   5.   0.   1.   1.   2.]
 [  0. 537.   6.  15.   1.  11.   2.   1.   1.   0.   0.   2.   1.   0.
    3.   1.   0.   0.   0.   0.]
 [  1.  10. 510.  23.   0.  18.   2.   0.   0.   0.   0.   3.   1.   1.
    0.   2.   0.   0.   1.   0.]
```

```
[  0.  12.   4. 547.   3.   5.   6.   0.   0.   0.   0.   2.   3.   0.
   1.   1.   1.   1.   1.   0.]
[  1.   4.   2.   5. 549.   2.   0.   0.   2.   0.   0.   2.   1.   3.
   1.   1.   0.   0.   2.   0.]
[  1.  12.   8.   4.   2. 557.   0.   0.   1.   1.   0.   1.   0.   0.
   2.   1.   1.   0.   1.   0.]
[  1.   4.   0.  30.   6.   1. 473.  20.   1.   3.   3.  10.  13.   3.
   1.   3.   5.   1.   4.   0.]
[  1.   0.   0.   2.   1.   2.   3. 570.   1.   1.   0.   1.   1.   1.
   0.   1.   2.   0.   4.   1.]
[  1.   1.   0.   1.   1.   0.   4.   2. 579.   0.   0.   0.   0.   2.
   0.   2.   2.   0.   1.   0.]
[  0.   3.   0.   1.   0.   1.   1.   2.   0. 579.   4.   0.   1.   1.
   0.   0.   1.   0.   0.   0.]
[  1.   0.   1.   2.   0.   0.   0.   2.   0.   0. 585.   1.   1.   0.
   0.   1.   0.   1.   3.   0.]
[  0.   2.   0.   0.   0.   0.   0.   0.   0.   0.   0. 583.   0.   1.
   0.   0.   2.   0.   6.   0.]
[  0.   4.   1.  14.   3.   0.   3.   1.   0.   0.   1.   4. 551.   2.
   2.   1.   2.   0.   2.   0.]
[  0.   1.   0.   0.   0.   1.   0.   1.   1.   0.   0.   1.   2. 580.
   1.   4.   2.   0.   0.   0.]
[  1.   1.   0.   1.   0.   2.   0.   1.   0.   0.   0.   1.   1.   1.
 582.   1.   0.   0.   1.   0.]
[  0.   2.   0.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.   1.
   0. 590.   2.   2.   1.   0.]
[  0.   0.   0.   0.   0.   0.   0.   1.   0.   0.   0.   0.   1.   0.
   0.   2. 539.   0.   2.   0.]
[  0.   1.   0.   0.   0.   0.   0.   1.   0.   0.   0.   1.   0.   2.
   0.   7.   0. 548.   4.   0.]
[  2.   2.   0.   0.   0.   1.   0.   0.   0.   0.   0.   3.   0.   1.
   0.   0.   4.   2. 449.   0.]
[ 19.   1.   0.   0.   0.   1.   0.   0.   1.   0.   0.   0.   0.   0.
   1.   28.  13.   4.   2. 306.]]
```

2.2.2 Evaluation on Testing data

The classifier uses both the Bayesian and ML Estimators when evaluating performance on the training data and outputs the Overall Accuracy, Class Accuracy values and the confusion matrix.

- Using Bayesian Estimators

1. Overall Accuracy = 0.7873417721518987 or **78.73417721518987 %**

2. Class Accuracy:

```
Class1: 0.7735849056603774
Class2: 0.7609254498714653
Class3: 0.5140664961636828
Class4: 0.7780612244897959
Class5: 0.7180156657963447
Class6: 0.7743589743589744
Class7: 0.6675392670157068
Class8: 0.8860759493670886
Class9: 0.9042821158690176
Class10: 0.9017632241813602
Class11: 0.9523809523809523
Class12: 0.9088607594936708
Class13: 0.6641221374045801
Class14: 0.8320610687022901
Class15: 0.875
Class16: 0.9346733668341709
Class17: 0.9093406593406593
Class18: 0.8351063829787234
Class19: 0.5870967741935483
Class20: 0.3705179282868526
```

3. Confusion Matrix:

```
[[246.   0.   0.   0.   0.   1.   0.   0.   1.   0.   1.   1.   2.   4.
    3.  33.   4.   7.   5.  10.]
 [  4. 296.   6.  13.  10.  20.   1.   2.   1.   0.   0.  14.   6.   2.
    8.   4.   0.   0.   2.   0.]
 [  2.  36. 201.  59.  13.  35.   0.   4.   2.   3.   1.  13.   2.   2.
    5.   4.   0.   0.   9.   0.]
 [  0.   9.  15. 305.  20.   2.   4.   7.   0.   0.   1.   4.  23.   0.
    1.   0.   0.   0.   1.   0.]
 [  0.  10.   9.  33. 275.   1.   3.   9.   0.   1.   0.   6.  16.   7.
    6.   0.   3.   0.   4.   0.]
 [  0.  43.  11.   9.   2. 302.   1.   0.   1.   1.   0.  10.   0.   3.
    3.   0.   2.   0.   2.   0.]
 [  0.   8.   3.  44.  15.   0. 255.  27.   3.   1.   1.   2.  10.   1.
    2.   3.   2.   2.   3.   0.]
 [  0.   2.   0.   1.   0.   1.   7. 350.  10.   1.   0.   1.   4.   0.
    2.   1.   6.   1.   8.   0.]
 [  0.   2.   0.   0.   0.   0.   2.  23. 359.   2.   0.   0.   0.   1.
    0.   1.   5.   0.   2.   0.]
 [  2.   2.   0.   1.   1.   2.   3.   3.   1. 358.  11.   2.   3.   1.
    0.   1.   1.   0.   5.   0.]
 [  2.   0.   0.   0.   0.   0.   1.   1.   1.   5. 380.   1.   1.   2.
    0.   0.   2.   0.   3.   0.]
 [  0.   4.   1.   1.   2.   1.   1.   0.   1.   0.   0. 359.   3.   1.
    1.   1.  11.   0.   8.   0.]
 [  2.  19.   1.  23.  10.   2.   1.  13.   4.   0.   0.  40. 261.   6.
    4.   4.   1.   1.   0.   1.]
 [  9.   8.   1.   2.   0.   0.   0.   5.   1.   0.   0.   2.   3. 327.
    4.  13.   6.   5.   7.   0.]
 [  2.  11.   0.   0.   0.   0.   0.   0.   0.   0.   1.   1.   4.   3.
  343.   3.   2.   1.  20.   1.]
 [  9.   2.   0.   1.   1.   1.   0.   0.   0.   0.   0.   0.   1.   1.
    1. 372.   2.   2.   2.   3.]
 [  1.   0.   0.   0.   0.   0.   1.   2.   1.   1.   1.   4.   1.   3.
    1.   2. 331.   1.  11.   3.]
 [ 16.   2.   0.   0.   0.   0.   0.   2.   1.   1.   1.   3.   0.   0.
    1.   6.   6. 314.  22.   1.]
 [  7.   2.   0.   0.   0.   0.   0.   1.   0.   0.   0.   4.   0.   2.
    7.   1.  95.   5. 182.   4.]
 [ 53.   4.   0.   0.   0.   0.   0.   0.   1.   1.   0.   0.   0.   3.
    5.  59.  19.   4.   9.  93.]]
```

- **Using Maximum Likelihood Estimators**

(Using -1.00e+100 in place of negative infinite. Using further negative values will further increase misclassification)

1. Overall Accuracy = 0.7221852098600933 or 72.21852098600932 %
2. Class Accuracy:

```
Class1: 0.7767295597484277
Class2: 0.6683804627249358
Class3: 0.42710997442455245
Class4: 0.6071428571428571
Class5: 0.5274151436031331
Class6: 0.7128205128205128
Class7: 0.4973821989528796
Class8: 0.8126582278481013
Class9: 0.8664987405541562
Class10: 0.8211586901763224
Class11: 0.924812030075188
Class12: 0.9113924050632911
Class13: 0.5521628498727735
Class14: 0.811704834605598
Class15: 0.8367346938775511
Class16: 0.8743718592964824
Class17: 0.75
Class18: 0.8803191489361702
Class19: 0.5870967741935483
Class20: 0.47808764940239046
```

3. Confusion Matrix:

```
[[247.   0.   0.   0.   0.   1.   0.   0.   1.   0.   0.   4.   0.   4.
    7.  25.   1.   7.   6.  15.]
 [  2. 260.  10.  10.   5.  31.   3.   1.   0.   0.   2.  25.   6.   9.
   18.   3.   2.   1.   1.   0.]
 [  1.  50. 167.  35.  14.  36.   2.   1.   2.   0.   3.  33.   4.  20.
   11.   5.   2.   1.   4.   0.]
 [  1.  24.  48. 238.  15.   7.  11.   2.   0.   0.   0.  10.  28.   3.
    5.   0.   0.   0.   0.   0.]
 [  5.  40.  15.  42. 202.   8.   5.   3.   1.   0.   2.  10.  20.  21.
```

```
   6.   0.   2.   0.   1.   0.]
[  1.  60.  10.   4.   2. 278.   0.   1.   2.   0.   0.  11.   3.   9.
   6.   0.   2.   0.   1.   0.]
[  1.  35.  13.  36.  17.   4. 190.  17.   8.   0.   3.   6.  14.  15.
   9.   1.   3.   6.   2.   2.]
[  1.   6.   2.   1.   2.   0.  13. 321.  17.   0.   0.   3.   7.   3.
   9.   1.   2.   1.   5.   1.]
[  0.   2.   0.   0.   0.   0.   2.  32. 344.   0.   0.   2.   3.   3.
   2.   0.   3.   2.   2.   0.]
[  1.   1.   0.   1.   0.   1.   3.   3.   3. 326.  16.   4.   0.  10.
   8.   3.   5.   3.   9.   0.]
[  1.   3.   2.   1.   0.   0.   3.   0.   3.   5. 369.   1.   0.   1.
   1.   1.   4.   0.   3.   1.]
[  2.   5.   0.   1.   3.   2.   0.   0.   0.   0.   0. 360.   1.   1.
   4.   1.   7.   3.   4.   1.]
[  3.  31.   2.  17.   5.   6.  10.  12.   9.   0.   0.  41. 217.  17.
  19.   0.   3.   0.   1.   0.]
[ 10.   7.   0.   2.   0.   0.   2.   4.   3.   0.   0.   8.   7. 319.
   8.   6.   7.   6.   3.   1.]
[  3.  12.   0.   1.   0.   4.   1.   2.   1.   1.   1.   8.   3.  13.
 328.   2.   4.   3.   4.   1.]
[ 19.   2.   0.   0.   0.   2.   0.   0.   0.   0.   0.   0.   0.   3.
   6. 348.   0.   6.   2.  10.]
[  6.   3.   0.   0.   0.   0.   0.   3.   0.   0.   0.  12.   0.  10.
   4.   4. 273.   6.  26.  17.]
[ 11.   1.   0.   1.   0.   0.   1.   2.   0.   1.   0.   2.   0.   1.
   1.  10.   5. 331.   9.   0.]
[ 15.   1.   0.   0.   1.   1.   0.   0.   1.   0.   0.  11.   0.   8.
   9.   5.  49.  20. 182.   7.]
[ 41.   0.   0.   0.   0.   0.   1.   1.   1.   1.   0.   6.   0.   6.
   7.  42.  14.   5.   6. 120.]]
```

Performance Evaluation: Training data vs Testing data

The training data gives a much higher percentage of accuracy in classification than the testing data. This is because the model has been trained using the same data and is being evaluated on the same. Whereas, when we use the testing data it is a completely different dataset. The Estimators have not been freshly restructured using that test data. Therefore, the accuracy is lower. However, this is a truer figure of how accurate the model actually is.

Performance Evaluation: Bayesian vs Maximum Likelihood Estimators

There is a large difference between the accuracy provided by the Bayesian against that of the ML estimators. This is because, Bayesian estimators are never zero. Even when a certain word W_k does not appear in any document of a class Omega=j, the $P(W_k|\text{Omega}=j)$ does not fall to zero, thereby avoiding the logarithm to go to negative infinite and removing all chances of classifying in that class even if there are other words having high probabilities in that class. This is precisely what happens with ML estimators and therefore, the misclassification rate is large.