COM S 573 Lab1 Eliska Kloberdanz

In this lab, I implemented the Naive Bayes algorithm in Python from scratch and I used it to tackle the “20 Newsgroups “ classification problem. The model is multinomial and uses the “bag of words” approach. The summaries below show the class priors, overall accuracies, class accuracies, and confusion matrices for both training and testing data for two different posterior probability estimators: the maximum likelihood estimator (MLE) and the Bayesian estimator (BE).

The Naive Bayes assumes that features are mutually independent given a category, which means that the posterior probability is a product of terms. The classifier computes the following equation to decide the category for a given document:

$$\omega_{NB} = argmax_{\omega_j} P(\omega_j) \prod_{i\ in\ positions} P(x_i|\omega_j)$$

However, it is more convenient to compute the natural log of this, because the product of posteriors becomes a summation. Therefore, this solves the issue that a product of small probabilities may become virtually a zero. That is why I used the equation below for my Naive Bayes implementation:

$$\omega_{NB} = argmax_{\omega_j} [\ln P(\omega_j) + \sum_{i\ in\ positions} \ln P(x_i|\omega_j)].$$

Comparison of MLE and BE:

The MLE is calculated as follows:

$$P_{MLE}(w_k|\omega_j) = \frac{n_k}{n}$$

The MLE is a commonly used estimator that is asymptotically consistent. However, if n_k is zero (i.e.: a value is unobserved) the MLE will assign a zero probability to it. This is an issue. The BE avoids this issue by adding a 1 to n_k. Therefore, it will never assign a zero probability. The BE is calculated as follows:

$$P_{BE}(w_k|\omega_j) = \frac{n_k+1}{n+|Vocabulary|}$$

Because of this 1 that is always added to n_k in case of BE, I observed that P_{MLE} is zero while P_{BE} is not for a certain word in a certain category if that category doesn't contain that word.

Comparison of performance:

The overall accuracy on training data is sightly higher for BE (95%) than MLE (92%). However, the accuracy on testing data is quite different for BE (79%) and MLE (53%). Based on this we can conclude that the BE performed a lot better on testing data than MLE. One of the reasons for this is that BE performs better than MLE on smaller sample sizes of data (Pandey et al., 2011). And this is again because MLE assigns zero probabilities to unobserved values, which means that when estimating from small samples MLE will assign a zero probability to a value that may have occurred if the sample size was larger. And therefore, such probability would not be a zero.

```
Class priors:
P(Omega =  1 ) 0.04259472890229834
P(Omega =  2 ) 0.05155736977549028
P(Omega =  3 ) 0.05075871860857219
P(Omega =  4 ) 0.05208980388676901
P(Omega =  5 ) 0.051024935664211554
P(Omega =  6 ) 0.052533498979501284
P(Omega =  7 ) 0.051646108794036735
P(Omega =  8 ) 0.052533498979501284
P(Omega =  9 ) 0.052888455053687104
P(Omega =  10 ) 0.0527109770165942
P(Omega =  11 ) 0.05306593309078002
P(Omega =  12 ) 0.0527109770165942
P(Omega =  13 ) 0.05244475996095483
P(Omega =  14 ) 0.0527109770165942
P(Omega =  15 ) 0.052622237998047744
P(Omega =  16 ) 0.05315467210932647
P(Omega =  17 ) 0.04836276510781791
P(Omega =  18 ) 0.05004880646020055
P(Omega =  19 ) 0.04117490460555506
P(Omega =  20 ) 0.033365870973467035

Overall accuracy for MLE (Training)=  0.9245718342355134

Class Accuracy for MLE (Training):
Group  1 : 0.9520833333333333
Group  2 : 0.8537005163511188
Group  3 : 0.8548951048951049
Group  4 : 0.8347529812606473
Group  5 : 0.8556521739130435
Group  6 : 0.9358108108108109
Group  7 : 0.7474226804123711
Group  8 : 0.9391891891891891
Group  9 : 0.9664429530201343
Group  10 : 0.930976430976431
Group  11 : 0.959866220735786
Group  12 : 0.9747474747474747
Group  13 : 0.9170896785109983
Group  14 : 0.9797979797979798
Group  15 : 0.9797639123102867
Group  16 : 0.9532554257095158
Group  17 : 0.9798165137614679
Group  18 : 0.9769503546099291
Group  19 : 0.9655172413793104
Group  20 : 0.9547872340425532
```

MLE Confusion Matrix (Training):

```
      [1]  [2]  [3]  [4]  [5]  [6]  [7]  [8]  [9]  [10]  [11]  [12]  [13]  [14]  [15]  [16]  [17]  [18]  [19]  [20]
[1]   457    0    0    0    0    0    0    0    0     0     0     6     0     1     2     4     0     2     4     4
[2]     0  496    3    1    0   15    2    1    2     1     3    18     3     6    15     5     1     7     2     0
[3]     0   11  489    9    0    3    5    1    3     0     0    13     3     9    11     2     1     3     8     1
[4]     0    7    7  490    5   10    6    0    3     0     2    11     8    13    13     2     2     5     3     0
[5]     1    8    2    3  492   12    4    1    0     0     6    17     0     6    11     1     4     4     0     3
[6]     0    3    0    1    0  554    1    0    1     1     2     5     1     5     9     0     3     3     3     0
[7]     0   22    8    7    3   10  435    6    1     0     4    18    21     9    13     2     3    15     5     0
[8]     1    1    1    1    0    1    3  556    0     0     4     4     1     1     6     0     3     6     0     3
[9]     1    0    0    0    0    2    3    1  576     0     0     0     1     2     2     2     1     4     1     0
[10]    0    0    0    0    0    1    1    2    2   553     8     3     1     3     4     6     1     4     5     0
[11]    0    1    0    0    1    0    2    0    0     5   574     1     0     0     4     0     1     4     5     0
[12]    0    0    0    0    0    1    0    0    0     0     0   579     0     1     1     1     6     4     1     0
[13]    2    2    1    1    0    3    4    3    0     0     3     6   542     4     6     3     3     5     1     2
[14]    0    0    0    0    0    0    1    0    0     0     0     1     1   582     0     0     3     0     1     5
[15]    1    0    0    0    1    0    0    0    1     0     1     2     0     2   581     0     0     2     0     2
[16]    0    0    0    0    0    0    1    1    1     0     0     1     0     2     4   571     5     8     4     1
[17]    0    0    0    0    0    0    0    0    0     0     0     1     0     2     0     1   534     3     3     1
[18]    1    0    0    0    0    1    0    0    0     0     0     1     0     0     1     5     0   551     3     1
[19]    2    0    0    0    0    2    0    0    1     0     0     0     0     0     0     1     4     4   448     2
[20]    3    2    0    0    0    0    0    0    1     0     0     2     0     1     0     3     0     1     4   359
```

Overall accuracy for BE (Training)= 0.9481764131688704

Class Accuracy for BE (Training):
Group 1 : 0.98125
Group 2 : 0.9242685025817556
Group 3 : 0.8916083916083916
Group 4 : 0.9318568994889267
Group 5 : 0.9547826086956521
Group 6 : 0.9408783783783784
Group 7 : 0.8127147766323024
Group 8 : 0.9628378378378378
Group 9 : 0.9714765100671141
Group 10 : 0.9747474747474747
Group 11 : 0.9782608695652174
Group 12 : 0.9814814814814815
Group 13 : 0.9323181049069373
Group 14 : 0.9764309764309764
Group 15 : 0.9814502529510961
Group 16 : 0.9849749582637729
Group 17 : 0.9889908256880734
Group 18 : 0.9716312056737588
Group 19 : 0.9676724137931034
Group 20 : 0.8138297872340425

BE Confusion Matrix (Training):

```
      [1]  [2]  [3]  [4]  [5]  [6]  [7]  [8]  [9]  [10]  [11]  [12]  [13]  [14]  [15]  [16]  [17]  [18]  [19]  [20]
[1]   471    0    0    0    0    0    0    0    0     0     0     0     0     0     0     5     0     1     1     2
[2]     0  537    6   15    1   11    2    1    1     0     0     2     1     0     3     1     0     0     0     0
[3]     1   10  510   23    0   18    2    0    0     0     0     3     1     1     0     2     0     0     1     0
[4]     0   12    4  547    3    5    6    0    0     0     0     2     3     0     1     1     1     1     1     0
[5]     1    4    2    5  549    2    0    0    2     0     0     2     1     3     1     1     0     0     2     0
[6]     1   12    8    4    2  557    0    0    1     1     0     1     0     0     2     1     1     0     1     0
[7]     1    4    0   30    6    1  473   20    1     3     3    10    13     3     1     3     5     1     4     0
[8]     1    0    0    2    1    2    3  570    1     1     0     1     1     1     0     1     2     0     4     1
[9]     1    1    0    1    1    0    4    2  579     0     0     0     0     2     0     2     2     0     1     0
[10]    0    3    0    1    0    1    1    2    0   579     4     0     1     1     0     0     1     0     0     0
[11]    1    0    1    2    0    0    0    2    0     0   585     1     1     0     0     1     0     1     3     0
[12]    0    2    0    0    0    0    0    0    0     0     0   583     0     1     0     0     2     0     6     0
[13]    0    4    1   14    3    0    3    1    0     0     1     4   551     2     2     1     2     0     2     0
[14]    0    1    0    0    0    1    0    1    1     0     0     1     2   580     1     4     2     0     0     0
[15]    1    1    0    1    0    2    0    1    0     0     0     1     1     1   582     1     0     0     1     0
[16]    0    2    0    1    0    0    0    0    0     0     0     0     0     1     0   590     2     2     1     0
[17]    0    0    0    0    0    0    0    1    0     0     0     0     1     0     0     2   539     0     2     0
[18]    0    1    0    0    0    0    0    1    0     0     0     1     0     2     0     7     0   548     4     0
[19]    2    2    0    0    0    1    0    0    0     0     0     3     0     1     0     0     4     2   449     0
[20]   19    1    0    0    0    1    0    0    1     0     0     0     0     0     1    28    13     4     2   306
```

Overall accuracy for MLE (Testing Data) = 0.5272485009993337

Class Accuracy for MLE (Testing Data):
Group 1 : 0.5974842767295597
Group 2 : 0.3676092544987147
Group 3 : 0.18414322250639387
Group 4 : 0.2627551020408163
Group 5 : 0.22193211488250653
Group 6 : 0.5743589743589743
Group 7 : 0.21204188481675393
Group 8 : 0.5240506329113924
Group 9 : 0.7052896725440806
Group 10 : 0.5188916876574308
Group 11 : 0.7243107769423559
Group 12 : 0.739240506329114
Group 13 : 0.3994910941475827
Group 14 : 0.6921119592875318
Group 15 : 0.7244897959183674
Group 16 : 0.7512562814070352
Group 17 : 0.5164835164835165
Group 18 : 0.8351063829787234
Group 19 : 0.535483870967742
Group 20 : 0.41832669322709165

MLE Confusion Matrix (Testing Data):

```
       [1]  [2]  [3]  [4]  [5]  [6]  [7]  [8]  [9]  [10]  [11]  [12]  [13]  [14]  [15]  [16]  [17]  [18]  [19]  [20]
[1]    190    1    1    0    0    0    0    0    2     0     1    10     0     5     9    41     2    23    13    20
[2]      3  143    5    4    0   53    4    0    2     2    11    55     9    25    40    10     7    10     6     0
[3]      0   22   72   15    9   44    8    1    2     2     8    61     5    48    46     5    12    20    10     1
[4]      0   22   32  103    7   43   10    1    2     0     4    53    28    21    37     4     9     8     3     5
[5]      5   21   10   23   85   44    8    5    2     1     7    56    14    29    36     5    12    13     4     3
[6]      2   35    4    2    0  224    2    2    1     0     2    32     2    33    26     2     5    10     5     1
[7]      2   26   12   14    6   15   81   18    4     1    11    32    31    32    35     9    11    31     8     3
[8]      6    3    2    0    1    4    7  207   14     0     5    25    11    17    34    12    20    18     7     2
[9]      2    1    0    0    0    2    4   26  280     1     3     5     7    16    16     1    14    10     8     1
[10]     1    1    0    0    2    3    3    1    2   206    28    24     1    29    27     6    12    25    23     3
[11]     0    3    1    1    0    3    6    0    2     7   289     7     0     8    11     6     8    34    12     1
[12]     4    2    0    0    0    3    0    0    0     0     0   292     2     6    10     6    37    23     8     2
[13]     2   17    2    6    3   13    9    4    9     0     2    50   157    31    48     1    17    14     7     1
[14]     8    5    0    0    0    0    0    2    0     0     0    17     5   272    12    17     8    29    16     2
[15]     2   11    1    1    0    4    0    2    1     1     1    18     2    21   284     4    12    15    11     1
[16]     9    1    0    0    0    3    1    0    0     2     2     4     0     5    13   299     1    32    16    10
[17]     7    0    0    0    1    1    0    1    0     0     1    22     1    18    12     6   188    29    53    24
[18]     8    0    0    0    0    0    1    0    1     0     1     9     0     1     1    18     5   314    16     1
[19]    18    1    0    0    1    2    0    0    2     0     4    17     0    10    11     6    19    44   166     9
[20]    21    1    0    0    0    1    1    1    2     1     1    15     0    11    12    31    10    21    17   105
```

Overall accuracy for BE (Testing Data) = 0.7873417721518987

Class Accuracy for BE (Testing Data):
Group 1 : 0.7735849056603774
Group 2 : 0.7609254498714653
Group 3 : 0.5140664961636828
Group 4 : 0.7780612244897959
Group 5 : 0.7180156657963447
Group 6 : 0.7743589743589744
Group 7 : 0.6675392670157068
Group 8 : 0.8860759493670886
Group 9 : 0.9042821158690176
Group 10 : 0.9017632241813602
Group 11 : 0.9523809523809523
Group 12 : 0.9088607594936708
Group 13 : 0.6641221374045801
Group 14 : 0.8320610687022901
Group 15 : 0.875
Group 16 : 0.9346733668341709
Group 17 : 0.9093406593406593
Group 18 : 0.8351063829787234
Group 19 : 0.5870967741935483
Group 20 : 0.3705179282868526

BE Confusion Matrix (Testing Data):

```
       [1]  [2]  [3]  [4]  [5]  [6]  [7]  [8]  [9]  [10]  [11]  [12]  [13]  [14]  [15]  [16]  [17]  [18]  [19]  [20]
[1]    246    0    0    0    0    1    0    0    1     0     1     1     2     4     3    33     4     7     5    10
[2]      4  296    6   13   10   20    1    2    1     0     0    14     6     2     8     4     0     0     2     0
[3]      2   36  201   59   13   35    0    4    2     3     1    13     2     2     5     4     0     0     9     0
[4]      0    9   15  305   20    2    4    7    0     0     1     4    23     0     1     0     0     0     1     0
[5]      0   10    9   33  275    1    3    9    0     1     0     6    16     7     6     0     3     0     4     0
[6]      0   43   11    9    2  302    1    0    1     1     0    10     0     3     3     0     2     0     2     0
[7]      0    8    3   44   15    0  255   27    3     1     1     2    10     1     2     3     2     2     3     0
[8]      0    2    0    1    0    1    7  350   10     1     0     1     4     0     2     1     6     1     8     0
[9]      0    2    0    0    0    0    2   23  359     2     0     0     0     1     0     1     5     0     2     0
[10]     2    2    0    1    1    2    3    3    1   358    11     2     3     1     0     1     1     0     5     0
[11]     2    0    0    0    0    0    1    1    1     5   380     1     1     2     0     0     2     0     3     0
[12]     0    4    1    1    2    1    1    0    1     0     0   359     3     1     1     1    11     0     8     0
[13]     2   19    1   23   10    2    1   13    4     0     0    40   261     6     4     4     1     1     0     1
[14]     9    8    1    2    0    0    0    5    1     0     0     2     3   327     4    13     6     5     7     0
[15]     2   11    0    0    0    0    0    0    0     0     1     1     4     3   343     3     2     1    20     1
[16]     9    2    0    1    1    1    0    0    0     0     0     0     1     1     1   372     2     2     2     3
[17]     1    0    0    0    0    0    1    2    1     1     1     4     1     3     1     2   331     1    11     3
[18]    16    2    0    0    0    0    0    2    1     1     1     3     0     0     1     6     6   314    22     1
[19]     7    2    0    0    0    0    0    1    0     0     0     4     0     2     7     1    95     5   182     4
[20]    53    4    0    0    0    0    0    0    1     1     0     0     0     3     5    59    19     4     9    93
```