ECE 448 FALL 2020
Assignment 3:
Naive Bayes/Perceptron/Logistic Regression Classification

Nov. 16, 2020

Yucheng Jin

Yiqing Xie

Hangtao Jin

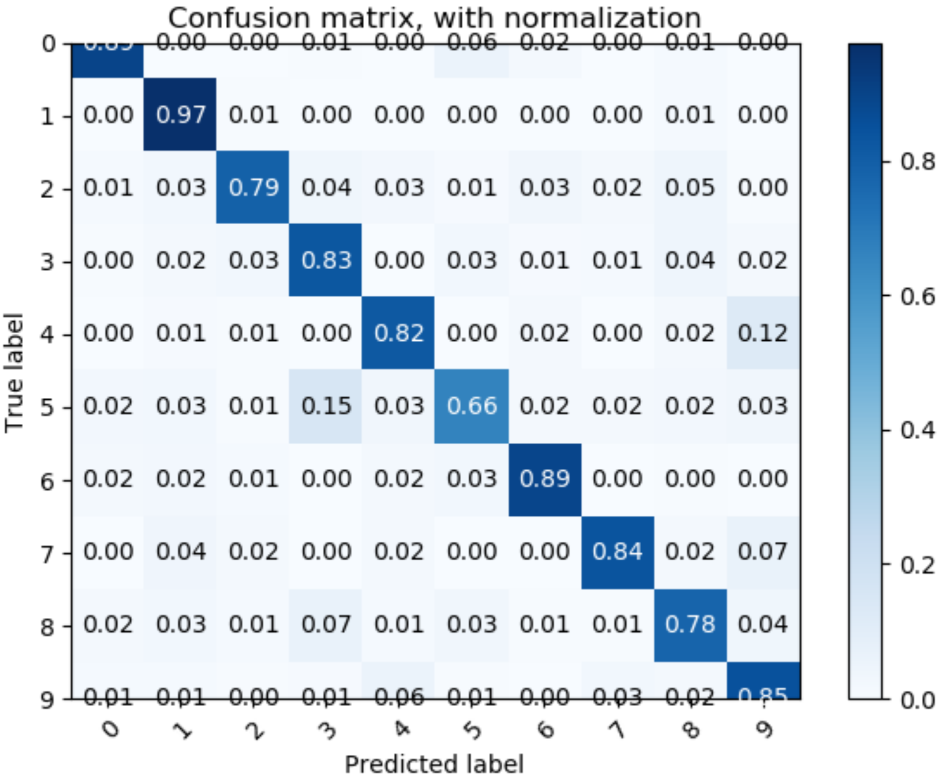
Section I

Average classification rate: for k = 1.0, the rate is 0.836

The classification rate for each class:

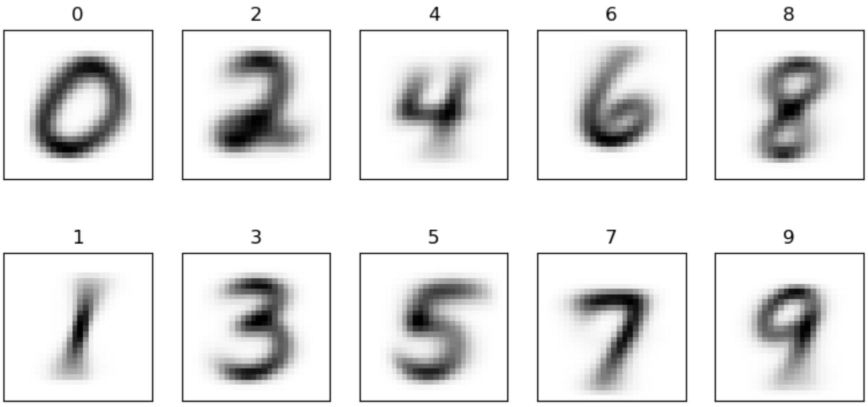
0	1	2	3	4	5	6	7	8	9
0.89	0.97	0.79	0.83	0.82	0.66	0.89	0.84	0.78	0.85

The confusion matrix:

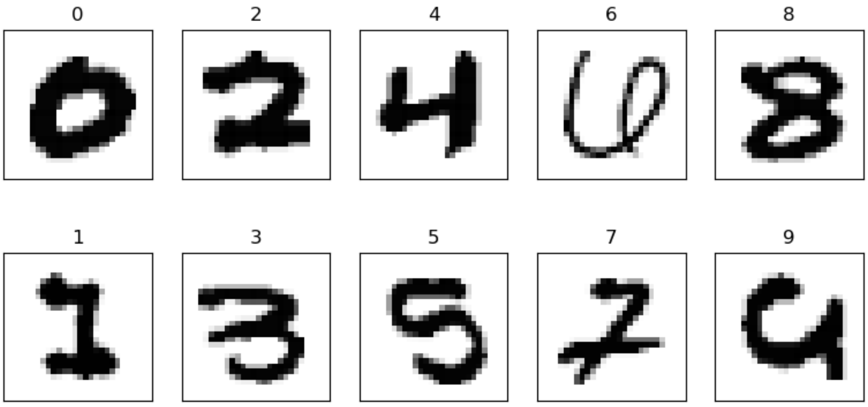


	0	1	2	3	4	5	6	7	8	9
0	0.89	0.00	0.00	0.01	0.00	0.06	0.02	0.00	0.01	0.00
1	0.00	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
2	0.01	0.03	0.79	0.04	0.03	0.01	0.03	0.02	0.05	0.00
3	0.00	0.02	0.03	0.83	0.00	0.03	0.01	0.01	0.04	0.02
4	0.00	0.01	0.01	0.00	0.82	0.00	0.02	0.00	0.02	0.12
5	0.02	0.03	0.01	0.15	0.03	0.66	0.02	0.02	0.02	0.03
6	0.02	0.02	0.01	0.00	0.02	0.03	0.89	0.00	0.00	0.00
7	0.00	0.04	0.02	0.00	0.02	0.00	0.00	0.84	0.02	0.07
8	0.02	0.03	0.01	0.07	0.01	0.03	0.01	0.01	0.78	0.04
9	0.01	0.01	0.00	0.01	0.06	0.01	0.00	0.03	0.02	0.85

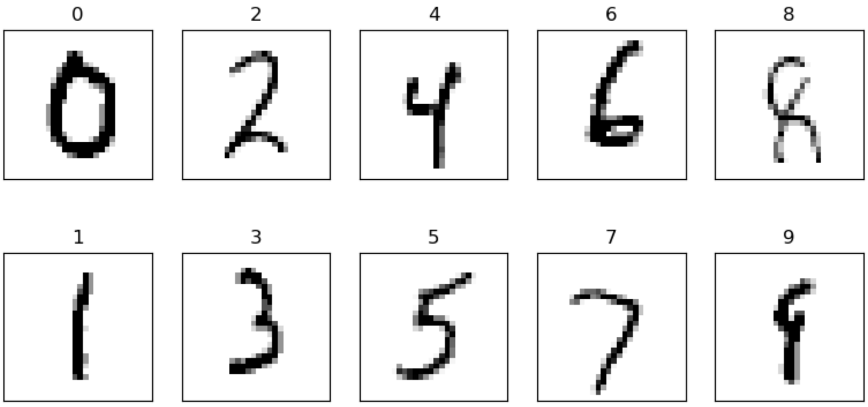
Plot(k=1.0):



Example with the lowest posterior probability:

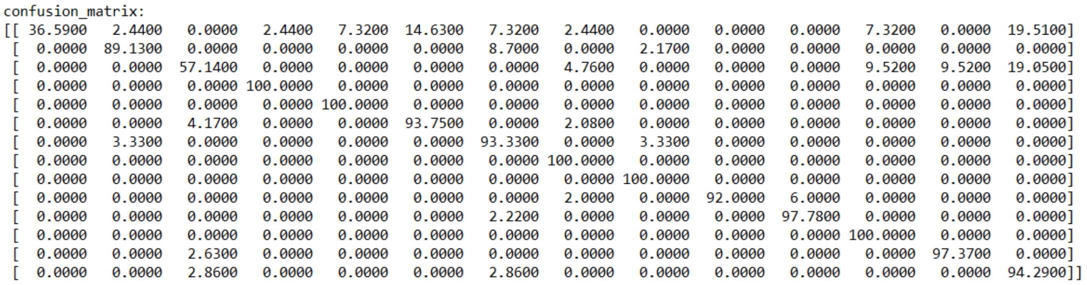


Example with the highest posterior probability:



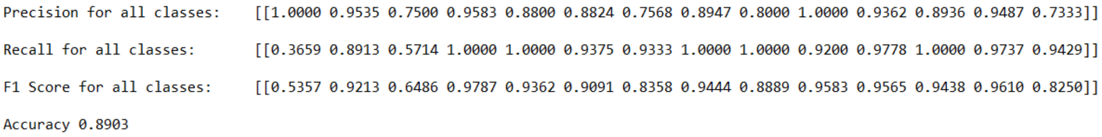
Section II

MAP confusion matrix:



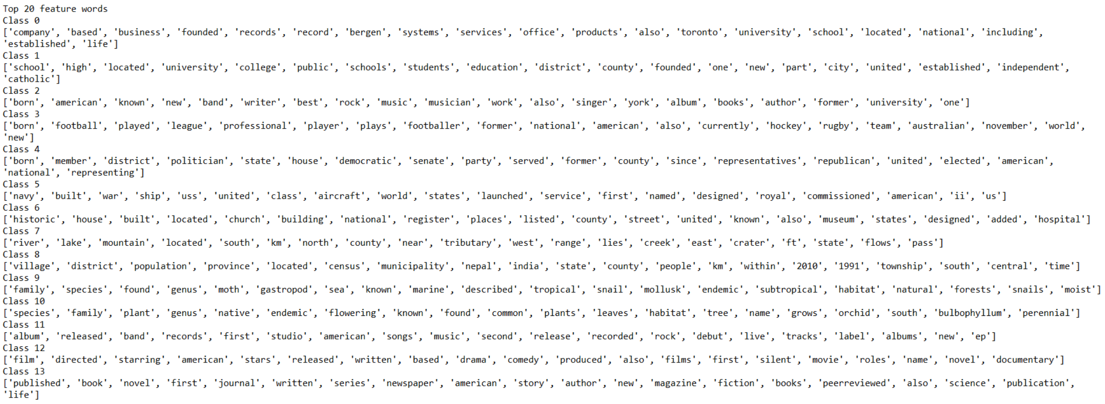
```
confusion_matrix:
[[ 36.5900   2.4400   0.0000   2.4400   7.3200  14.6300   7.3200   2.4400   0.0000   0.0000   0.0000   7.3200   0.0000  19.5100]
 [  0.0000  89.1300   0.0000   0.0000   0.0000   0.0000   8.7000   0.0000   2.1700   0.0000   0.0000   0.0000   0.0000   0.0000]
 [  0.0000   0.0000  57.1400   0.0000   0.0000   0.0000   0.0000   4.7600   0.0000   0.0000   0.0000   9.5200   9.5200  19.0500]
 [  0.0000   0.0000   0.0000 100.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000]
 [  0.0000   0.0000   0.0000   0.0000 100.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000]
 [  0.0000   0.0000   4.1700   0.0000   0.0000  93.7500   0.0000   2.0800   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000]
 [  0.0000   3.3300   0.0000   0.0000   0.0000   0.0000  93.3300   0.0000   3.3300   0.0000   0.0000   0.0000   0.0000   0.0000]
 [  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 100.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000]
 [  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 100.0000   0.0000   0.0000   0.0000   0.0000   0.0000]
 [  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   2.0000   0.0000  92.0000   6.0000   0.0000   0.0000   0.0000]
 [  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   2.2200   0.0000   0.0000   0.0000  97.7800   0.0000   0.0000   0.0000]
 [  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 100.0000   0.0000   0.0000]
 [  0.0000   0.0000   2.6300   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000  97.3700   0.0000]
 [  0.0000   0.0000   2.8600   0.0000   0.0000   0.0000   2.8600   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000  94.2900]]
```

MAP output:



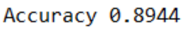
```
Precision for all classes:    [[1.0000 0.9535 0.7500 0.9583 0.8800 0.8824 0.7568 0.8947 0.8000 1.0000 0.9362 0.8936 0.9487 0.7333]]

Recall for all classes:       [[0.3659 0.8913 0.5714 1.0000 1.0000 0.9375 0.9333 1.0000 1.0000 0.9200 0.9778 1.0000 0.9737 0.9429]]

F1 Score for all classes:     [[0.5357 0.9213 0.6486 0.9787 0.9362 0.9091 0.8358 0.9444 0.8889 0.9583 0.9565 0.9438 0.9610 0.8250]]

Accuracy 0.8903
```

Top 20 feature words:



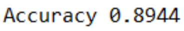
```
Top 20 feature words
Class 0
['company', 'based', 'business', 'founded', 'records', 'record', 'bergen', 'systems', 'services', 'office', 'products', 'also', 'toronto', 'university', 'school', 'located', 'national', 'including',
'established', 'life']
Class 1
['school', 'high', 'located', 'university', 'college', 'public', 'schools', 'students', 'education', 'district', 'county', 'founded', 'one', 'new', 'part', 'city', 'united', 'established', 'independent',
'catholic']
Class 2
['born', 'american', 'known', 'new', 'band', 'writer', 'best', 'rock', 'music', 'musician', 'work', 'also', 'singer', 'york', 'album', 'books', 'author', 'former', 'university', 'one']
Class 3
['born', 'football', 'played', 'league', 'professional', 'player', 'plays', 'footballer', 'former', 'national', 'american', 'also', 'currently', 'hockey', 'rugby', 'team', 'australian', 'november', 'world',
'new']
Class 4
['born', 'member', 'district', 'politician', 'state', 'house', 'democratic', 'senate', 'party', 'served', 'former', 'county', 'since', 'representatives', 'republican', 'united', 'elected', 'american',
'national', 'representing']
Class 5
['navy', 'built', 'war', 'ship', 'uss', 'united', 'class', 'aircraft', 'world', 'states', 'launched', 'service', 'first', 'named', 'designed', 'royal', 'commissioned', 'american', 'ii', 'us']
Class 6
['historic', 'house', 'built', 'located', 'church', 'building', 'national', 'register', 'places', 'listed', 'county', 'street', 'united', 'known', 'also', 'museum', 'states', 'designed', 'added', 'hospital']
Class 7
['river', 'lake', 'mountain', 'located', 'south', 'km', 'north', 'county', 'near', 'tributary', 'west', 'range', 'lies', 'creek', 'east', 'crater', 'ft', 'state', 'flows', 'pass']
Class 8
['village', 'district', 'population', 'province', 'located', 'census', 'municipality', 'nepal', 'india', 'state', 'county', 'people', 'km', 'within', '2010', '1991', 'township', 'south', 'central', 'time']
Class 9
['family', 'species', 'found', 'genus', 'moth', 'gastropod', 'sea', 'known', 'marine', 'described', 'tropical', 'snail', 'mollusk', 'endemic', 'subtropical', 'habitat', 'natural', 'forests', 'snails', 'moist']
Class 10
['species', 'family', 'plant', 'genus', 'native', 'endemic', 'flowering', 'known', 'found', 'common', 'plants', 'leaves', 'habitat', 'tree', 'name', 'grows', 'orchid', 'south', 'bulbophyllum', 'perennial']
Class 11
['album', 'released', 'band', 'records', 'first', 'studio', 'american', 'songs', 'music', 'second', 'release', 'recorded', 'rock', 'debut', 'live', 'tracks', 'label', 'albums', 'new', 'ep']
Class 12
['film', 'directed', 'starring', 'american', 'stars', 'released', 'written', 'based', 'drama', 'comedy', 'produced', 'also', 'films', 'first', 'silent', 'movie', 'roles', 'name', 'novel', 'documentary']
Class 13
['published', 'book', 'novel', 'first', 'journal', 'written', 'series', 'newspaper', 'american', 'story', 'author', 'new', 'magazine', 'fiction', 'books', 'peerreviewed', 'also', 'science', 'publication',
'life']
```

ML accuracy:



```
Accuracy 0.8944
```

Uniform distribution accuracy:



```
Accuracy 0.8944
```

Explanation for different accuracy:

If we ignore class prior or apply a uniform distribution, the accuracy of the result increase. Including the class prior doesn’t always beneficial. That’s because the class distribution between training and test sets may be different. What’s more, the result ignoring the class prior and that using uniform distribution share the same result.

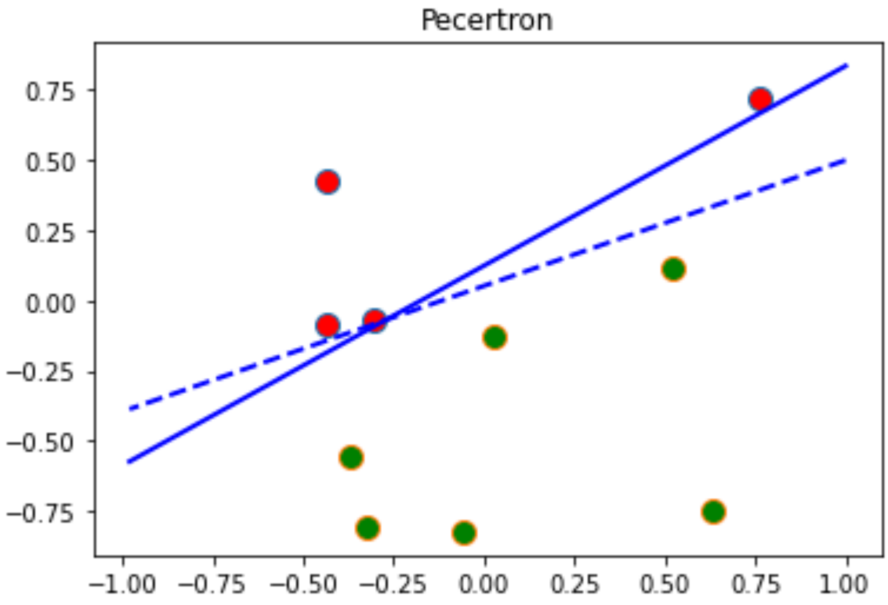
Section III

Part 3.1: Perceptron model

Perceptron:

E_train is 0.0, E_test is 0.10804179999999984

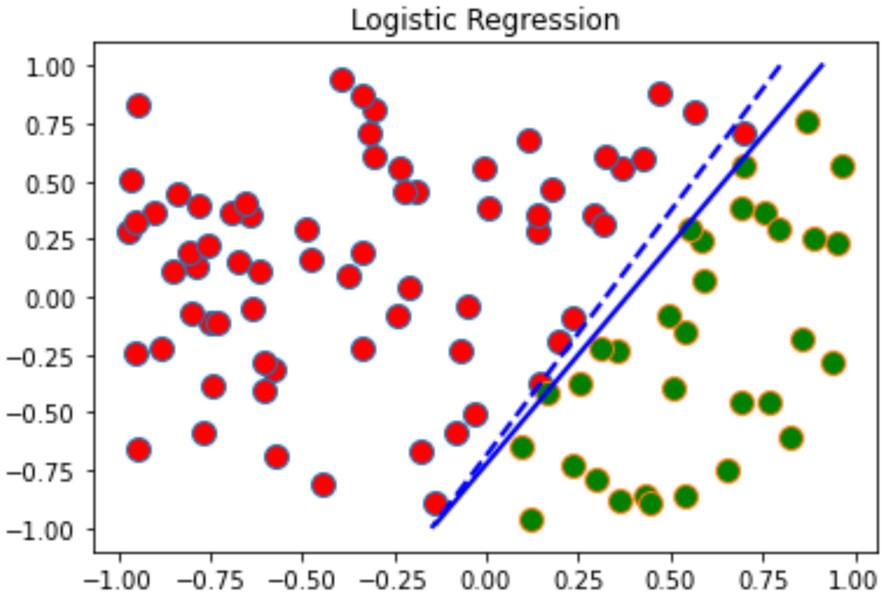
Average number of iterations is 3.7419999999999636



Part 3.2: Logistic regression model

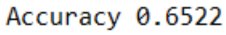
Logistic:

E_train is 0.05280000000000003, E_test is 0.05781600000000003



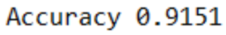
Extra Credit for Section II

Bigram model:



```
Accuracy 0.6522
```

Optimal mixture model ($\lambda = 0.33$):



```
Accuracy 0.9151
```

Question 1:

Relaxing the naïve assumption isn’t always a good thing. As we can see, when using $\lambda = 1$, which means only consider bigram model and relax the assumption the most, the accuracy decreases. This may because such strict test cases cause a lot of documents share similarly low probability. They have small probability in all classes, which let them be randomly classified.

Question 2:

When N is a really large number, we may find that almost all test documents have minimal probability and because they all don’t fit any test case. Thus such model is useless.

Contribution

Part I: Yucheng Jin
Part II: Yiqing Xie
Part III: Hangtao Jin