CS155 Set 6

Timothy Liu

March 2, 2018

1 Problem 1

1.1 Problem A

1.1.1 Part i

$$p(x|y=c) = p\Bigg((x_1|y=c),(x_2|y=c),(x_3|y=c).....(x_j|y=c)\Bigg)$$

$$p(x|y=c) = \prod_{j=1}^{D} \theta_{xjc}$$

This requires $O(DC)$ parameters to be stored.

1.1.2 Part ii

We would need to store $O(D2^DC)$ parameters because we have to store one parameter for every combination of other parameters and then multiply this by the number of values of C. This is drastically more parameters to store.

1.2 Problem B

The Naive Bayes model has fewer parameters, and for a small training set it will be less prone to overfitting. The Naive Bayes will have lower test set error.

1.3 Problem C

If the sample size N is very large, then the full model will have lower test set error. The larger number of parameters will be able to better capture the large sample size, and it will not have the same overfitting problem if the training set is large enough.

1.4 Problem D

Making a prediction with Naive Bayes requires multiplying D parameters together and there are C classes, making this a $O(DC)$ operation. In the complete case, we must multiply D parameters together, making it a $O(D)$ operation.

2 Problem 2

2.1 Problem A

```
######################################################################
                      Running Code For Question 2A
######################################################################


File #0:
Emission Sequence              Max Probability State Sequence
######################################################################
25421                          31033
0123236753 4                   2222100310
545267426152 7433              1031003103222222
722621316451226 7255           1310331000033100310
02471206023520510102 55241     2222222222222222222222103


File #1:
Emission Sequence              Max Probability State Sequence
######################################################################
77550                          22222
7224523677                     2222221000
505767442426747                222100003310031
72134131645536112267           10310310000310333100
4733667771450051060253041      2221000003222223103222223


File #2:
Emission Sequence              Max Probability State Sequence
######################################################################
60622                          11111
4687981156                     2100202111
815833657775062                021011111111111
21310222515963505015           02020111111111111021
6503199452571274006320025      1110202111111102021110211


File #3:
Emission Sequence              Max Probability State Sequence
######################################################################
13661                          00021
2102213421                     3131310213
166066262165133                133333133133100
53164662112162634156           20000021313131002133
1523541005123230226306256      1310021333133133313133133


File #4:
Emission Sequence              Max Probability State Sequence
######################################################################
23664                          01124
3630535602                     0111201112
350201162150142                011244012441112
00214005402015146362           11201112412444011112
2111266524665143562534450      2012012424124011112411124


File #5:
Emission Sequence              Max Probability State Sequence
######################################################################
68535                          10111
4546566636                     1111111111
638436858181213                110111010000011
13240338308444514688           00010000000111111100
0111664434441382533632626      2111111111111100111110101
```

Figure 1: Output for viterbi algorithm.

2.2 Problem B

2.2.1 Part i

```
######################################################################
                    Running Code For Question 2Bi
######################################################################

File #0:
Emission Sequence              Probability of Emitting Sequence
######################################################################
25421                          4.537e-05
01232367534                    1.620e-11
5452674261527433               4.348e-15
7226213164512267255            4.739e-18
0247120602352051010255241      9.365e-24


File #1:
Emission Sequence              Probability of Emitting Sequence
######################################################################
77550                          1.181e-04
7224523677                     2.033e-09
505767442426747                2.477e-13
72134131645536112267           8.871e-20
4733667771450051060253041      3.740e-24


File #2:
Emission Sequence              Probability of Emitting Sequence
######################################################################
60622                          2.088e-05
4687981156                     5.181e-11
815833657775062                3.315e-15
21310222515963505015           5.126e-20
6503199452571274006320025      1.297e-25


File #3:
Emission Sequence              Probability of Emitting Sequence
######################################################################
13661                          1.732e-04
2102213421                     8.285e-09
166066262165133                1.642e-12
53164662112162634156           1.063e-16
1523541005123230226306256      4.535e-22


File #4:
Emission Sequence              Probability of Emitting Sequence
######################################################################
23664                          1.141e-04
3630535602                     4.326e-09
350201162150142                9.793e-14
00214005402015146362           4.740e-18
2111266524665143562534450      5.618e-22


File #5:
Emission Sequence              Probability of Emitting Sequence
######################################################################
68535                          1.322e-05
4546566636                     2.867e-09
638436858181213                4.323e-14
13240338308444514688           4.629e-18
0111664434441382533632626      1.440e-22
```

Figure 2: Probabilities calculated from forwards algorithm.

2.2.2 Part ii

```
######################################################################
                    Running Code For Question 2Bii
######################################################################


File #0:
Emission Sequence              Probability of Emitting Sequence
######################################################################
25421                          4.537e-05
01232367534                    1.620e-11
545267426152743 3              4.348e-15
72262131645122672 55           4.739e-18
02471206023520510102 55241     9.365e-24


File #1:
Emission Sequence              Probability of Emitting Sequence
######################################################################
77550                          1.181e-04
7224523677                     2.033e-09
505767442426747                2.477e-13
72134131645536112267           8.871e-20
4733667771450051060253041      3.740e-24


File #2:
Emission Sequence              Probability of Emitting Sequence
######################################################################
60622                          2.088e-05
4687981156                     5.181e-11
815833657775062                3.315e-15
21310222515963505015           5.126e-20
6503199452571274006320025      1.297e-25


File #3:
Emission Sequence              Probability of Emitting Sequence
######################################################################
13661                          1.732e-04
2102213421                     8.285e-09
166066262165133                1.642e-12
53164662112162634156           1.063e-16
1523541005123230226306256      4.535e-22


File #4:
Emission Sequence              Probability of Emitting Sequence
######################################################################
23664                          1.141e-04
3630535602                     4.326e-09
350201162150142                9.793e-14
00214005402015146362           4.740e-18
2111266524665143562534450      5.618e-22


File #5:
Emission Sequence              Probability of Emitting Sequence
######################################################################
68535                          1.322e-05
4546566636                     2.867e-09
638436858181213                4.323e-14
13240338308444514688           4.629e-18
0111664434441382533632626      1.440e-22
```

Figure 3: Probabilities calculated from backwards algorithm.

2.3 Problem C

Transition matrix:

```
#####################################################################
                     Running Code For Question 2C
#####################################################################


Transition Matrix:
#####################################################################
2.833e-01   4.714e-01   1.310e-01   1.143e-01
2.321e-01   3.810e-01   2.940e-01   9.284e-02
1.040e-01   9.760e-02   3.696e-01   4.288e-01
1.883e-01   9.903e-02   3.052e-01   4.075e-01
```

Figure 4: Transition matrix.

Observation matrix:

```
Observation Matrix:
#####################################################################
1.486e-01   2.288e-01   1.533e-01   1.179e-01   4.717e-02   5.189e-02   2.830e-02   1.297e-01   9.198e-02   2.358e-03
1.062e-01   9.653e-03   1.931e-02   3.089e-02   1.699e-01   4.633e-02   1.409e-01   2.394e-01   1.371e-01   1.004e-01
1.194e-01   4.299e-02   6.529e-02   9.076e-02   1.768e-01   2.022e-01   4.618e-02   5.096e-02   7.803e-02   1.274e-01
1.694e-01   3.871e-02   1.468e-01   1.823e-01   4.839e-02   6.290e-02   9.032e-02   2.581e-02   2.161e-01   1.935e-02
```

Figure 5: Observation matrix.

2.4 Problem D

Transition matrix:

```
Transition Matrix:
######################################################################
4.511e-01   2.475e-01   3.013e-01   1.103e-04
3.584e-01   2.123e-01   3.123e-01   1.170e-01
4.402e-09   9.756e-01   2.435e-02   1.640e-12
2.673e-07   2.993e-04   6.030e-01   3.967e-01
```

Figure 6: Learned transition matrix from unsupervised learning.

Observation matrix:

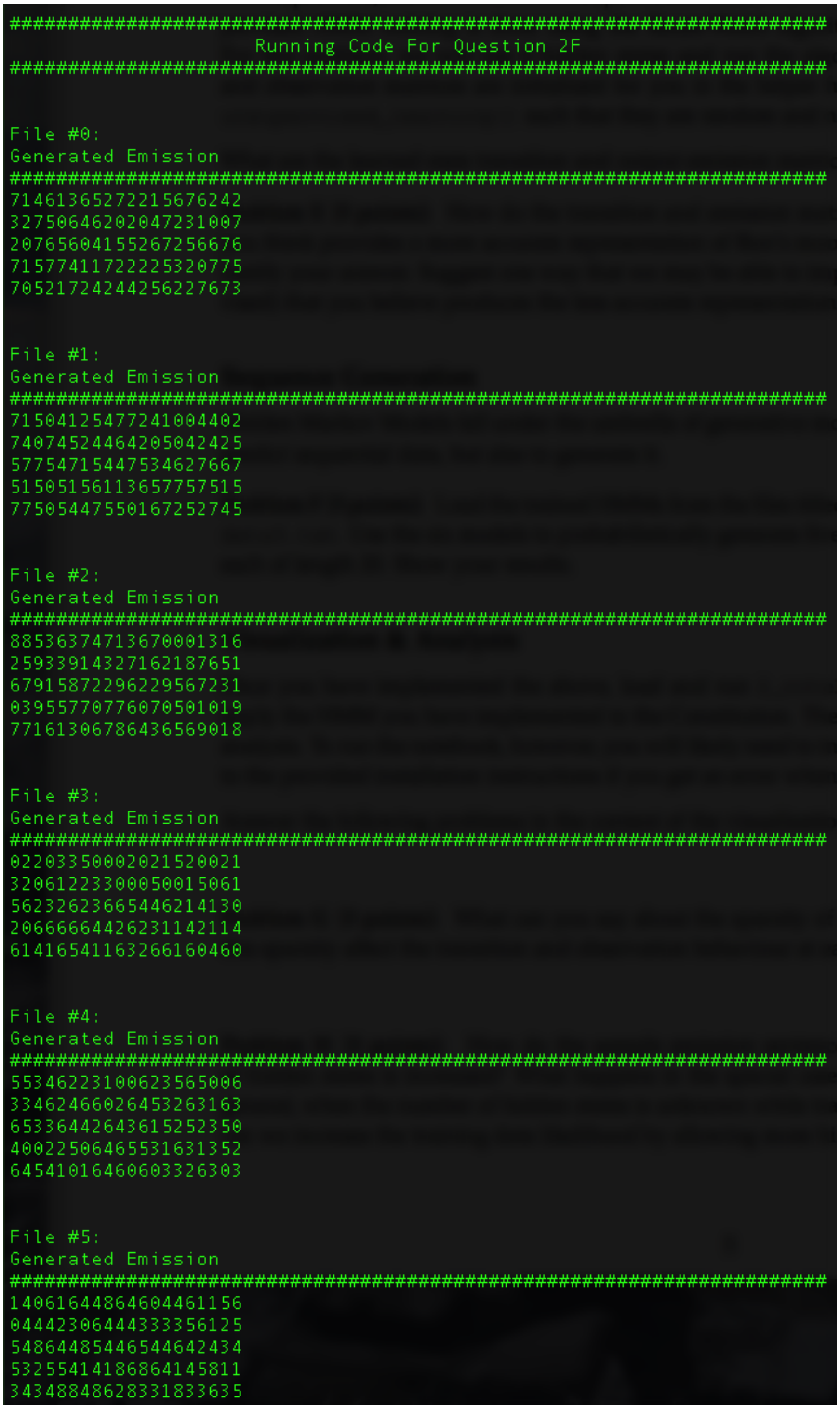
```
Observation Matrix:
##########################################################################
4.836e-03   2.849e-20   1.700e-02   2.280e-01   1.056e-01   1.326e-01   1.548e-01   2.587e-01   3.493e-02   6.344e-02
1.575e-01   8.094e-02   1.780e-01   2.047e-18   7.979e-02   1.139e-01   6.531e-02   9.053e-02   1.682e-01   6.590e-02
2.133e-01   1.466e-01   5.351e-02   1.219e-05   1.763e-01   5.630e-02   4.274e-02   1.614e-07   2.197e-01   9.155e-02
2.166e-01   2.359e-46   4.571e-02   6.014e-01   1.156e-01   2.066e-02   6.573e-47   1.024e-10   5.827e-27   2.985e-47
```

Figure 7: Learned observation matrix from unsupervised learning.

2.5 Problem E

The supervised model is likely more accurate because it's trained on both the values and the labels. The supervised model is guaranteed to have the correct states while the unsupervised model is guessing at how many states there are. The unsupervised transition matrix sometimes pushes the transition probability between two states to near zero, which the supervised matrix does not. The two transition matrices have little resemblance. Similarly, the observation matrix for unsupervised has some values very close to zero, which the transition matrix does not. The supervised matrix is likely a better representation because it's trained on labeled data. One way to improve the unsupervised model is to either feed it even more data or to somehow include regularization.

2.6 Problem F



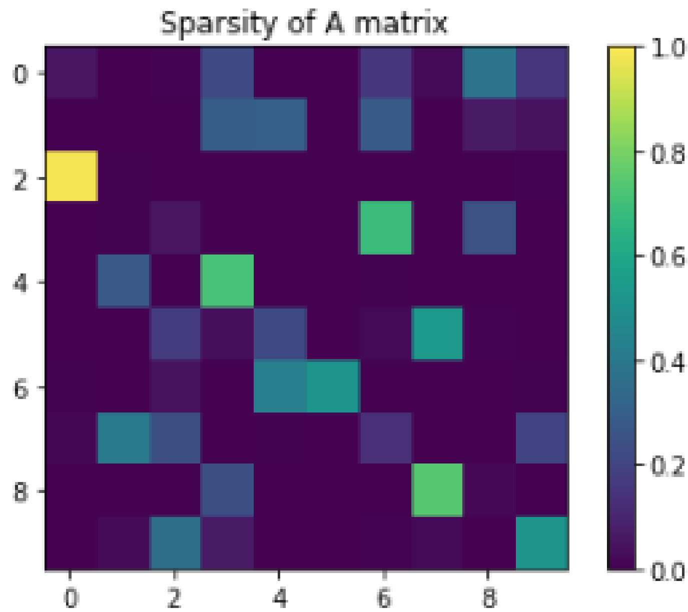
```
#######################################################################
                     Running Code For Question 2F
#######################################################################


File #0:
Generated Emission
#######################################################################
71461365272215676242
32750646202047231007
20765604155267256676
71577411722225320775
70521724244256227673


File #1:
Generated Emission
#######################################################################
71504125477241004402
74074524464205042425
57754715447534627667
51505156113657757515
77505447550167252745


File #2:
Generated Emission
#######################################################################
88536374713670001316
25933914327162187651
67915872296229567231
03955770776070501019
77161306786436569018


File #3:
Generated Emission
#######################################################################
02203350002021520021
32061223300050015061
56232623665446214130
20666664426231142114
61416541163266160460


File #4:
Generated Emission
#######################################################################
55346223100623565006
33462466026453263163
65336442643615252350
40022506465531631352
64541016460603326303


File #5:
Generated Emission
#######################################################################
14061644864604461156
04442306444333356125
54864485446544642434
53255414186864145811
34348848628331833635
```

Figure 8: Emitted sequences.

2.7 Problem G

Word cloud visualization of constitution:

Figure 9: Word cloud representing US constitution.



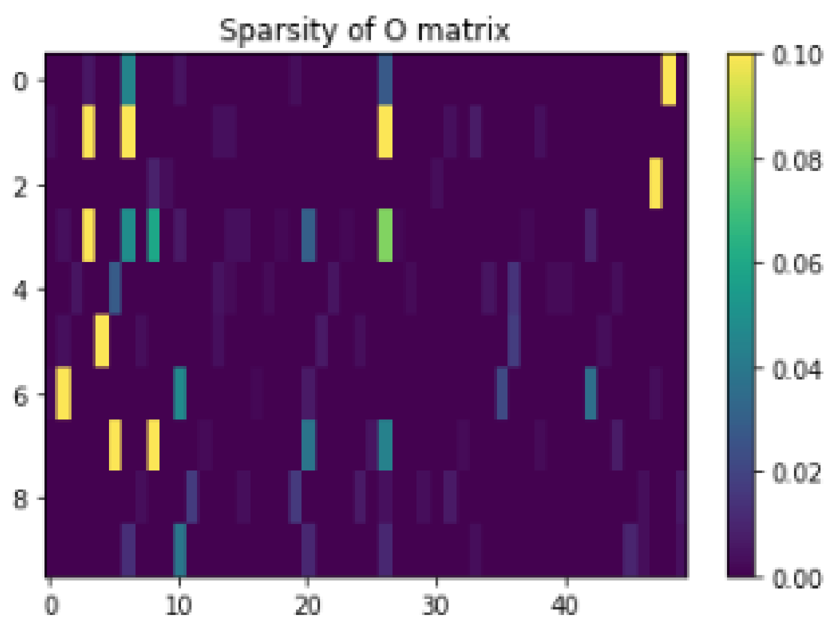


Figure 10: Sparsity of matrices.

Both the trained A and O matrices are fairly sparse. A large portion of both matrices are 0. This means that there are many possible state transitions that never happen, or there are some states that cannot emit certain observations. This makes sense, as there are some states - such as verb to verb - that may be unlikely.

2.8 Problem H

```
Sample Sentence:
====================
And of their shall disorderly having to trial in except shall being which person any page be impeachment be writ the
house shall interlined records...
```

Figure 11: Emission with 1 state.

```
Sample Sentence:
====================
Of power may state such shall as as shall which houses respective inferior be manner for from the of such be not prop
ose be according...
```

Figure 12: Emission with 2 states.

```
Sample Sentence:
====================
Of power may state such shall as as shall which houses respective inferior be manner for from the of such be not prop
ose be according...
```

Figure 13: Emission with 4 states.

```
print('Sample Sentence:\n====================')
print(sample_sentence(hmm8, obs_map, n_words=25))
```

```
Sample Sentence:
====================
Years and direct shall day expedient facto be for by publish a to be as ministers to approved majority which from the
united of the...
```

Figure 14: Emission with 10 states.

```
Sample Sentence:
====================
Be secure the trial but present necessary and insure deem appoint power shall than not this entered of the thing impa
iring before not of the...
```

Figure 15: Emission with 16 states.

The sample emissions become somewhat more meaningful as the number of states changes. In the special case of only one state, the sentence makes no grammatical sense. Adding more hidden states will likely improve performance but only to a certain point. If there are too many hidden states, then the HMM may try to extract meaning that does not exist.

2.9 Problem I

State 4 seems to represent descriptors of the legislative branch. The words "congress," "house," and "senate" all appear in the word cloud. This states differs from some others like state 8, which has no dominant words. The words in state 4 all appear to be describing something similar, rather than being a part of speech.