Machine Learning & Data Mining
Caltech CS/CNS/EE 155
Homework 6

Changhao Xu
UID: 2103530
March 5th, 2020

1 Class-Conditional Densities for Binary Data

Question A:

$$p(x|y=c) = p(x_1|y=c)p(x_2|x_1, y=c)\cdots p(x_D|x_1,\cdots,x_{D-1}, y=c) = \prod_{j=1}^{D}\theta_{xjc}$$

Since all the D features are binary, $x_j \in \{0,1\}$,
for each class of C we need 2^{j-1} parameters for $p(x_j|x_1,\cdots,x_{j-1},y=c)$. In total $\sum_{j=1}^{D} 2^{j-1}$ = $O(2^D)$.
Therefore, we need $O(C \times 2^D)$ parameters for C classes.

Question B: Without factorization, since all the D features are binary, $x_j \in \{0,1\}$, and C classes for y, in total we need $O(C \times 2^D)$, which is the same as that with factorization.

Question C: For a small N, Naive Bayes is likely to give lower test set error. This is because full models are more likely to overfit with a small N.

Question D : For a large N, full models are likely to give lower test set error. This is because Naive Bayes is simple and is likely to underfit with a large N, while full models have more parameters and will do better.

Question E: For Naive Bayes,

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(y)}{p(x)}\prod_d p(x^d|y) = \frac{p(y)}{\sum_{i=1}^{C} p(x|y=c_i)}\prod_d p(x^d|y)$$

Since we assumed a uniform class prior p(y), O(p(y)) = O(1), and computation complexity of $p(y|x) = O(CD)$.
For full model,

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(y)}{p(x)}p(x|y)$$

, and computation complexity of D-dimensional vector is O(D). Different from Naive Bayes, full model doesn't take into account every y = c, but their overall probability. Therefore, computation complexity of $p(y|x) = O(D)$.

2 Sequence Prediction

Question A:

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


File #0:
Emission Sequence              Max Probability State Sequence
########################################################################
25421                          31033
01232367534                    22222100310
5452674261527433               1031003103222222
7226213164512267255            1310331000033100310
0247120602352051010255241      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      1310021333133133133133133
```

```
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
```

Question B: The results using Forward algorithm are as follows:

```
########################################################################
                    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
```

Changhao Xu
UID: 2103530
March 5th, 2020

The results using Backward algorithm are as follows:

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


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
```

Question C:

```
######################################################################
                      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


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
```

Question D:

Changhao Xu
UID: 2103530
March 5th, 2020

```
######################################################################
                     Running Code For Question 2D
######################################################################


test case: N_iters=0
Transition Matrix:
######################################################################
2.470e-01   6.958e-02   3.064e-01   3.770e-01
1.846e-01   3.635e-01   2.114e-01   2.404e-01
2.902e-01   1.522e-01   1.328e-01   4.248e-01
3.266e-01   2.547e-01   5.671e-02   3.620e-01


Observation Matrix:
######################################################################
9.093e-02   8.443e-02   1.399e-01   1.539e-01   1.214e-01   6.519e-02   8.377e-02   1.268e-01   9.067e-02   4.306e-02
4.669e-02   2.347e-01   1.511e-02   5.726e-03   1.281e-01   4.333e-02   5.775e-02   1.542e-01   1.871e-01   1.273e-01
1.797e-01   1.101e-01   6.323e-02   1.346e-01   3.886e-02   9.526e-02   1.189e-01   2.195e-02   8.021e-02   1.572e-01
1.381e-01   2.190e-01   6.830e-02   1.514e-01   1.058e-01   1.222e-02   1.477e-01   1.160e-01   7.535e-03   3.392e-02


test case: N_iters=1
Transition Matrix:
######################################################################
2.702e-01   7.208e-02   3.259e-01   3.318e-01
2.003e-01   3.712e-01   2.216e-01   2.069e-01
3.229e-01   1.585e-01   1.439e-01   3.747e-01
3.561e-01   2.646e-01   6.016e-02   3.192e-01


Observation Matrix:
######################################################################
9.941e-02   3.250e-02   1.594e-01   1.280e-01   1.222e-01   1.215e-01   5.508e-02   1.076e-01   1.409e-01   3.342e-02
5.500e-02   9.644e-02   1.872e-02   4.955e-03   1.382e-01   8.785e-02   4.052e-02   1.451e-01   3.061e-01   1.072e-01
1.987e-01   4.211e-02   7.324e-02   1.137e-01   4.002e-02   1.815e-01   8.075e-02   1.931e-02   1.253e-01   1.253e-01
1.880e-01   1.038e-01   9.799e-02   1.553e-01   1.338e-01   2.815e-02   1.223e-01   1.229e-01   1.453e-02   3.325e-02
```

```
Transition Matrix:
######################################################################
8.029e-04   5.274e-01   8.376e-05   4.717e-01
1.856e-03   6.830e-01   2.922e-01   2.303e-02
6.214e-01   8.278e-13   3.747e-01   3.940e-03
2.051e-02   7.025e-01   2.042e-02   2.566e-01


Observation Matrix:
######################################################################
1.577e-01   4.863e-02   1.835e-01   4.863e-02   2.072e-01   3.686e-21   6.874e-02   2.040e-02   2.653e-01   1.458e-35
1.120e-01   5.788e-02   1.132e-01   1.021e-01   1.309e-01   9.003e-02   8.507e-14   1.870e-01   1.437e-01   6.304e-02
9.063e-02   7.887e-02   1.014e-16   1.937e-01   6.975e-02   2.191e-01   1.481e-01   9.426e-17   5.298e-02   1.468e-01
3.079e-01   1.320e-01   9.116e-02   3.171e-02   5.488e-03   1.068e-06   2.835e-01   6.466e-02   8.364e-02   4.515e-07
```

Changhao Xu
UID: 2103530
March 5th, 2020

Question E: The transition and emission matrices from 2C and 2D are different, supervised matrices are more uniform, while unsupervised are sparse and nonuniform. 2C (supervised HMM) provides a more accurate representation. This is because for unsupervised 2D, we simply random the initial matrices, which cannot truly reflect Ron's moods. To improve the unsupervised learning data, we might need to initialize the transition and emission matrices A & O more precisely, such as using probability distributions from supervised matrices.

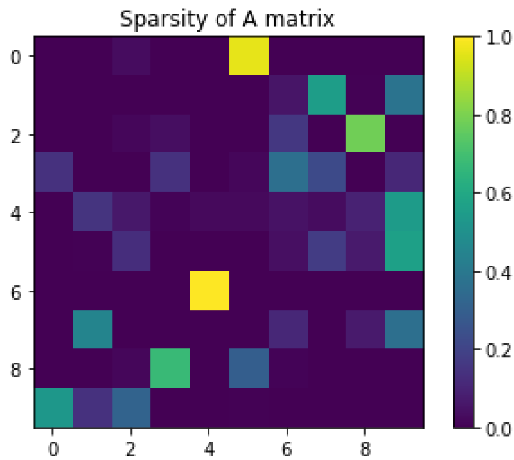
Question F:

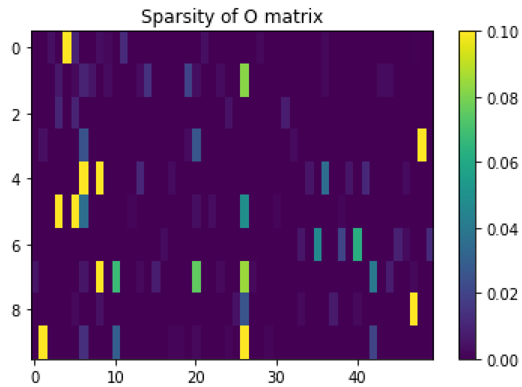
```
######################################################################
                    Running Code For Question 2F
######################################################################


File #0:
Generated Emission
######################################################################
45274511250275204412
05407242224726770147
72422713473031327660
42427312074107504324
65147177214245322422


File #1:
Generated Emission
######################################################################
40035770104447412207
45516552177027677212
04317422222000147762
35746073172105617607
05107555006525007447


File #2:
Generated Emission
######################################################################
17726563315712157156
90030125025149435211
62975620039788065245
45692563828145922702
52899330755675696052


File #3:
Generated Emission
######################################################################
03341033031362462112
15126633304022436420
62165152146652334362
12252065242314333013
21214502130111364421
```

```
File #4:
Generated Emission
######################################################################
34202331563531320223
16566131513143621520
43335062640460660662
62431603632366426636
16354030615436116252


File #5:
Generated Emission
######################################################################
32511566801635284208
81164464364484665615
04666240667481480436
60366565318454544314
45422451038680813715
```

Question G:

The trained A and O matrices are both sparse, with most elements near 0. The sparsity of transition matrix A means that for each state, there might be very few states to transit to, and the sparsity of observation matrix O means that for each state, there might be very few observations to belong to.





Question H: As the number of hidden states is increased, sample emission sentences from the HMM becomes more coherent and smooth. When there is only one hidden state, then there is no more transitions, meaning that observation words are just randomly picked from the dataset. Allowing more hidden states will increase the training data likelihood, but when we have too many hidden states for the fixed observation set, this will lead to overfitting and may increase the test error.

Question I: As shown in the figure, I think state 7 is semantically meaningful. This state is filled with

law-related words, such as 'law', 'court', 'representative', 'justice', 'rules'. I think this state represent all law-related words in the dataset, which is distinct from other states.