CS155 Homework 6

Ty Limpasuvan

11 late hours used; 0 remaining

1 Class-Conditional Densities for Binary Data

1.1 A

1.1.1 i

$$P(x|y) = c$$

x has multiple dimensions so we have

$$= P(x_1, x_2, ...x_D|y = c)$$

Then we apply the chain rule with the following

$$P(x_D|x_{D-1}x_{D-2}, ..., y = c) * P(x_{D-1}, x_{D-2}, ...|y = c)$$

Which gives us

$$= \prod_{j=1}^{D} \Theta_{xjc}$$

We need $O(2^D C)$ parameters to represent the factorization. We need C parameters for each y value and for a given j value we need 2^{j-1} because we have that many permutations of $x_1, x_2, ...x_{j-1}$.

1.1.2 ii

Without the factorization, we cannot ignore the terms that were considered by previous iterations of j. However, this means that the number of parameters needed to compute the joint probability is still $O(2^D C)$, which is the same as the situation where we used the factorization.

1.2 B

With a small training set of size N, Naive Bayes will likely give lower test set error; it involves the use of fewer parameters than the full model. With fewer parameters, the amount of test error will be lower as there are fewer things to estimate on. In a sense, having fewer parameters will reduce overfitting.

1.3 C

With a large training set of size N, the full model will likely give lower test error. This is because we now have more parameters to estimate on a much larger data set, so the predictions are more robust.

1.4 D

For Naive Bayes, we start with

$$P(y = c|x)$$

which is equivalent to

$$\frac{P(x|y = c)P(y = c)}{P(x)}$$

We then substitute the equation from the question to get

$$\frac{P(x|y = c)P(y = c)}{\sum_{a\epsilon C} P(x|y = a)P(y = a)}$$

Since we assumed a uniform class, we can cancel some terms to get

$$\frac{P(x|y = c)}{\sum_{a\epsilon C} P(x|y = a)}$$

The numerator has a computational complexity of O(D) and the denominator has O(CD), so the overall computation has complexity O(CD).

For the full model, we start with the same first few steps.

$$P(y = c|x)$$

$$\frac{P(x|y = c)P(y = c)}{P(x)}$$

We end up with

$$\frac{P(x|y = c)}{\prod_{a\epsilon C} P(x|y = a)P(y = a)}$$

Both the numerator and denominator have complexities of O(D), so the computation has an overall complexity of O(D).

2 Sequence Prediction

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



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

```
01116644344441382533632626     1.440e-22


#######################################################################
                   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

######################################################################
                     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-0
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-0
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-0
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-0



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


Transition Matrix:
######################################################################
2.690e-117  2.889e-123  9.661e-121  8.339e-111
1.119e-91   3.470e-95   8.797e-93   5.790e-84
1.295e+00   1.908e+00   8.936e-01   3.165e-01
2.970e-07   1.723e-16   4.755e-05   1.906e-01
...

Observation Matrix:
######################################################################
3.371e-08   2.189e-07   7.753e-03   1.045e-04   4.022e-04   5.157e-03   9.157e-01   1.922e-05   6.792e-0
4.384e-08   5.936e-08   1.272e-02   6.332e-04   2.157e-04   2.143e-03   8.628e-01   1.146e-04   1.263e-0
1.361e-01   6.988e-02   9.451e-02   1.078e-01   1.137e-01   9.679e-02   7.753e-02   1.037e-01   1.338e-0
```

```
3.568e-03   2.325e-03   1.212e-01   3.190e-02   1.055e-02   1.345e-01   2.435e-01   1.227e-02   1.510e-0
...
```

2.1 E

Based on my results, the values of the matrices in 2D are much smaller than those in 2C. 2C should be better; we actually have labels and the full data set. 2D can be improved with more data points and a convergence criterion.