Homework 5

1 Naive Bayes

Question A:
For the Red Domestic SUV:

$$
\begin{aligned}
& P(\text{stolen}) \prod P(a_i|v_j) \\
= \; & P(\text{stolen}) P(\text{Red}|\text{stolen}) P(\text{SUV}|\text{stolen}) P(\text{Domestic}|\text{stolen}) \\
= \; & \frac{3}{10} \cdot \frac{4.5}{6} \cdot \frac{3.5}{6} \cdot \frac{1.5}{6} \\
= \; & 21/640 = 0.0328125
\end{aligned}
$$

$$
\begin{aligned}
& P(\text{not stolen}) \prod P(a_i|v_j) \\
= \; & P(\text{not stolen}) P(\text{Red}|\text{not stolen}) P(\text{SUV}|\text{not stolen}) P(\text{Domestic}|\text{not stolen}) \\
= \; & \frac{7}{10} \cdot \frac{3.5}{10} \cdot \frac{5.5}{10} \cdot \frac{4.5}{10} \\
= \; & 4851/80000 = 0.0606375
\end{aligned}
$$

The resulting $V_{nb} = 0.0606375$, and my Red Domestic SUV car is more likely not to be stolen than stolen.

Question B:
No, Naive Bayes should not be used. Naive Bayes makes the assumption that the features are strongly independent of one another. In this case, we can clearly see that how much one sleeps one day most definitely affects how much one sleeps the next day and so forth. Thus, the assumption is invalid in this case.

2 Sequence Prediction

Question A:
The naive algorithm has time complexity $O(L^M)$ where L is the number of possible states, and M is the length of the sequence.

Question B:
The Viterbi algorithm has time complexity $O(M \cdot |L|^2)$.

Question C: True. When we increase the number of hidden states, it is essentially increasing the number of possibilities we have for classifying the observation. We take the speech example presented in lecture. If we have a limited amount of hidden states (parts of speech) such that the input sequence consists of words that if correctly labeled, would fall into hidden states/ parts of speech not in our current HMM, then the likelihood would be very low. However, if we add more hidden states/ parts of speech into our HMM such that the words in our limited HMM can now be correctly classified, the likelihood increases.

Question D:
From lecture slides, recall that $\alpha_z(1) = P(y^1 = z \mid y^0)P(x^1 \mid y^1 = z)$. If some coefficient of initial state is initially 0, then we get that for some z, $P(y^1 = z) = 0$ which then means that $P(y^1 = z \mid y^0) = 0$, and thus $\alpha_z(1) = 0$. Now, recall that our initial state distribution and transition matrix are updated as follows:

$$\pi_z^* = \frac{\alpha_z(1)\beta_z(1)}{\sum_{z'} \alpha_{z'}(1)\beta_{z'}(1)}$$
$$A_{b,a}^* = \frac{\alpha_a(i-1)P(y^i = b \mid y^{i-1} = a)P(x^i \mid y^i = b)\beta_b(i)}{\sum_{a',b'} \alpha_{a'}(i-1)P(y^i = b' \mid y^{i-1} = a')P(x^i \mid y^i = b')\beta_{b'}(i)}$$

Because π_z^* relies on $\alpha_z(1)$, $\pi_z^* = 0$ when $\alpha_z(1) = 0$ and if any coefficient of the initial state is initially 0, it will remain 0.

Similarly, if the coefficient of the state transition probability of a HMM is initially 0, then $P(y^i = b \mid y^{i-1} = a) = 0$ for that specific transition. Note that this term is included in the numerator of the transition matrix so if $P(y^i = b \mid y^{i-1} = a) = 0$, then $A_{b,a}^* = 0$ and thus the transition probabilities will not be updated. Hence, if a coefficient of the state transition probability matrix is initially 0, it will remain 0.

Question E:

```
File #0:
Emission Sequence              Max Probability State Sequence
######################################################################
25421                          31033
01232367534                    22222100310
5452674261527433               1031003103222222
72262131645122672 55           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
```

Question F:

i.

```
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
```

ii.

```
File #0:
Emission Sequence             Probability of Emitting Sequence
#######################################################################
25421                         4.537e-05
01232367534                   1.620e-11
5452674261527433              4.348e-15
722621316451226725 5          4.739e-18
024712060235205101025524 1    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 G:

```
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 H:

```
Transition Matrix:
##########################################################################
4.004e-01   1.453e-01   3.973e-01   5.703e-02
1.767e-01   2.530e-01   1.756e-01   3.946e-01
2.534e-01   1.292e-01   1.553e-01   4.621e-01
1.695e-01   4.725e-01   2.718e-01   8.623e-02


Observation Matrix:
##########################################################################
1.038e-01   1.257e-01   8.743e-02   1.585e-01   1.585e-01   7.104e-02   8.743e-02   1.585e-01   1.038e-01   1.475e-01
1.038e-01   1.257e-01   8.743e-02   1.585e-01   1.585e-01   7.104e-02   8.743e-02   1.585e-01   1.038e-01   1.475e-01
1.038e-01   1.257e-01   8.743e-02   1.585e-01   1.585e-01   7.104e-02   8.743e-02   1.585e-01   1.038e-01   1.475e-01
1.038e-01   1.257e-01   8.743e-02   1.585e-01   1.585e-01   7.104e-02   8.743e-02   1.585e-01   1.038e-01   1.475e-01
```

Question I: It seems that the matrices from 2G provide a more accurate representation of Ron's mood and how they affect his music choices. First of all, 2G is trained with fully supervised data while 2H is trained with unsupervised data(half of data in 2G). It makes intuitive sense that when given the moods of training data, we are better able to make predictions on moods than if we did not have these labels. Moreover, patterns in unsupervised learning is generally harder to decipher because there is a tendency to go to a local minimum which may not be meaningful.

Also, in comparing the transition matrices, we see that a majority of the values lying on the principal diagonal for 2G are greater than the corresponding diagonal values of 2H. These values correspond to the probabilities of staying in a current mood. Not an expert on human behavior, but it does seem that moods tend to linger. In addition, looking at the observation matrix for 2H and the same values for each column, it's unlikely that an observation (song choice)is equally likely for all states (moods).

One way to improve the unsupervised method is to increase the training data set by getting more observations.

Question J: My favorite sequence is the second of file #5. It reads "84883333312318283228." This is my favorite because it's essentially dominated by the two numbers, 8 and 3, a level of domination by two digits that is hard to see in the other generated emissions.

```
File #0:
Generated Emission
######################################################################
04424077452222467476
22744052262765741507
71735017552646604544
16270157573667272433
71750264155444346777


File #1:
Generated Emission
######################################################################
75770357104663074575
22427267005756072766
44704756075272647067
00450072572477710724
71705754512611145037


File #2:
Generated Emission
######################################################################
55663130355536654621
27975393002537357711
45628617709292863535
92751377756235288135
37438776621161925737


File #3:
Generated Emission
######################################################################
66111462134412263051
20110265446302400636
11412546303506145061
31341630252136121510
33221043341616663506


File #4:
Generated Emission
######################################################################
62104111601243221053
31454622650516230364
60124364611155666062
01620314063660222326
44546661554133166324


File #5:
Generated Emission
######################################################################
34838745114836888650
84883333312318283228
63261381180086136121
11851524236562286660
50144353630488133450
```