Machine Learning & Data Mining — Caltech CS/CNS/EE 155
Homework 6 — February 2019

---

7 LATE HOURS

1 Class-Conditional Densities for Binary Data [25 Points]

Problem A [5 points]: Parameters of Full Model with Factorizing

Solution A.:

Problem 1 A. We will start by factorizing $p(x|y)$
and we let $\Theta_{x_jc} = P(x_j|x_1,\ldots,x_{j-1}, y=c)$
Thus, we can factorize as $\prod_{i=1}^{D}(\Theta_{x_jc})$
We want to store every Θ_{x_jc}. X is binary and
D dimensional as stated. There are C classes to
mutiply these by so we get a total complexity
of $O(2^D C)$

Problem B [5 points]: Parameters of Full Model without Factorizing

Solution B.:

B. We still need to store the estimates of all the parameters in this case so the complexity is the same as the case where we factorize, $O(2^D C)$

Problem C [2 points]: Naive Bayes vs. Full Model for Small N

Solution C.: *If the training set is small, then Naive Bayes is better to use than Full Bayes. This is because Full Bayes assumes that the x of the input are dependent on each other. This will cause Full Bayes to overfit the data as there are so many possible dependencies it has to learn. This will cause convergence to an unideal local minimum or no convergence which will lead to high error. Naive Bayes assumes the x are independent so it won't run into these problems. Thus, Naive Bayes will have lower test error.*

Problem D [2 points]: Naive Bayes vs. Full Model for Large N

Solution D.: *For large N Full Bayes is better because there is enough data such that the model can be trained fully and it won't overfit. Naive Bayes won't perform as well because its independence assuption won't allow it to pick up on the real patterns in the data. Because Full Bayes will be able to capture the dependencies in the input and there is enough data for those to be captured, Full Bayes will have lower test error.*

Problem E [11 points]: Computational Complexity of Making a Prediction Using Naive Bayes vs Full Model

Solution E.:

E. The naive bayes for a single test case in the case where we have all the parameters already computed we get $O(CD)$ because all we need to worry about is $P(x) = \sum_y P(y,x)$ and we need to go through all classes and all features which is $C*D$.

For the time complexity of the full model we have a similar situation except we convert a D-bit vector which takes $O(D)$ but we end up getting $O(CD)$ because $O(2DC) \approx O(DC)$

2 Sequence Prediction [75 Points]

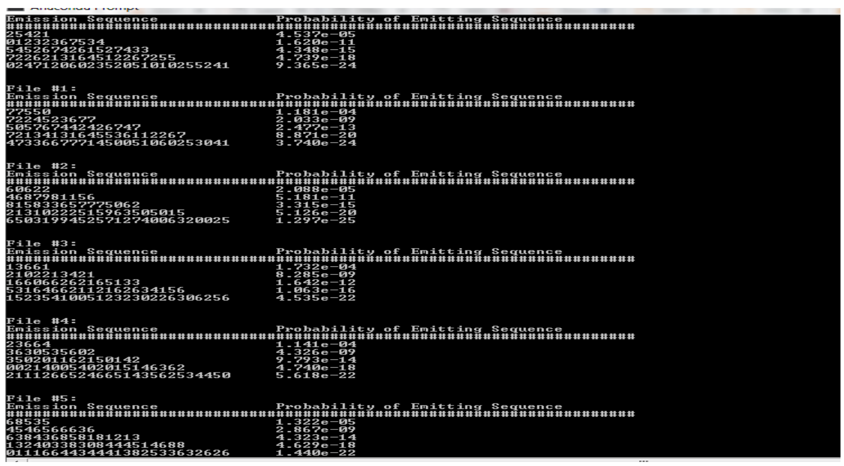
Problem A [10 points]: Max-Probability State Sequences for 6 Trained HMMs

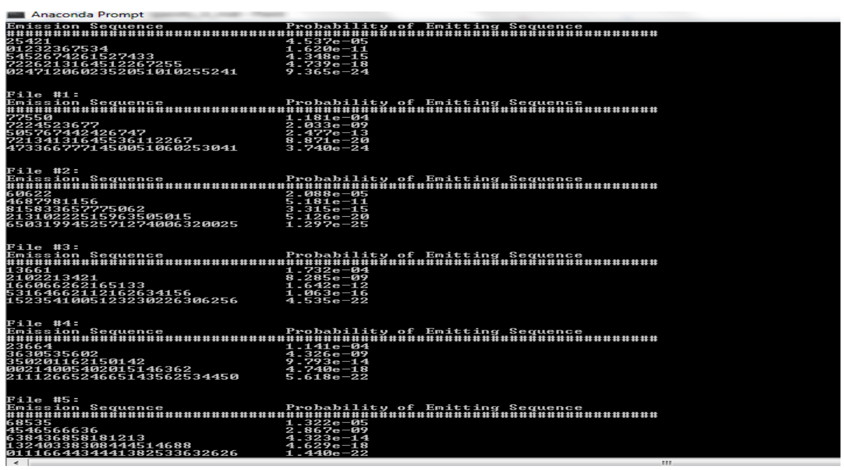
Solution A.:

```
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
5057674424 26747              222100003310031
7213413164553611 2267         1031031000031033 3100
4733667771450051060253041     2221000003222223103222223

File #2:
Emission Sequence             Max Probability State Sequence
######################################################################
60622                         11111
4687981156                    2100202111
8158336577 75062              021011111111111
2131022251596350 5015         0202011111111111 1021
5503199452571274006320025     1110202111111102021110211

File #3:
Emission Sequence             Max Probability State Sequence
######################################################################
13661                         00021
2102213421                    3131310213
166066262165133               133333133133100
5316466211216263 4156         2000002131313100 2133
1523541005123230226306256     1310021333133133313133133

File #4:
Emission Sequence             Max Probability State Sequence
######################################################################
23664                         01124
3630535602                    0111201112
350201162150142               011244012441112
0021400540201514 6362         1120111241244401 1112
2111266524665143562534450     2012012424124011112411124

File #5:
Emission Sequence             Max Probability State Sequence
######################################################################
68535                         10111
4546566636                    1111111111
638436858181213               110111010000011
1324033830844451 4688         0001000000001111 1100
0111664434441382533632626     2111111111111100111110101
```

Problem B [17 points]: Probability of Emission Sequence for 6 Trained HMMs

Solution B.:





Problem C [10 points]: Learned State Transition and Output Emission Matrices of Supervised Hidden Markov Model

Solution 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
```

Problem D [15 points]: Learned State Transition and Output Emission Matrices of Unsupervised Hidden Markov Model

Solution D.:

```
...
Transition Matrix:
######################################################################
5.413e-06   1.342e-01   8.658e-01   2.379e-08
1.269e-01   3.610e-01   2.221e-02   4.899e-01
3.634e-01   6.366e-01   4.555e-06   3.907e-09
3.501e-02   1.027e-04   3.197e-01   6.452e-01


Observation Matrix:
######################################################################
1.362e-01   7.629e-04   1.634e-01   1.769e-01   6.810e-03   3.249e-01   8.314e-03   3.654e-02   9.327e-02   5.301e-02
2.355e-01   1.144e-01   1.697e-01   3.305e-07   1.571e-01   6.108e-15   1.349e-01   3.375e-13   1.884e-01   2.590e-05
1.178e-01   6.175e-02   2.302e-41   1.560e-01   1.620e-01   1.034e-01   1.120e-01   1.037e-02   1.403e-01   1.363e-01
7.573e-02   6.812e-02   7.632e-02   1.293e-01   8.978e-02   7.933e-02   3.900e-02   2.643e-01   1.047e-01   7.342e-02
```

Problem E [5 points]: Compare 2C and 2D

Solution E.: *The matrices from C, the supervised case, providesa more accurate representation of Ron's moods and how they affect his music choices. One simple reason C is more accurate is becuase it is supervised. A supervised HMM, given a sufficient amount of data will trivially be more accurate than an unsupervised HMM. Furthermore, the matrices from D has much more variance than the matrices from C. This is probably a bad sign. As we look closer we see that the probability of staying in state 0 or 2 in D is very small whereas all the values along the diagonal in C are reasonably big. This should not be the case becuase peoples' mood should tend to stay the same. There definetly shouldn't be a mood which you exit immediately, that's not really a mood. Furthermore, there really shouldn't be very small values in the transition matrix at all because there should be a reasonably big ¿.01 chance to get to a mood from any other mood . It shouldn't be almost impossible to get to a mood from a different mood. In C this is true, in D it isn't. Looking at the observation matrix we also see near 0 values in the D observation matrix and not in C. 0 values don't really make sense because you can listen to all sorts of music in all different moods. For example, Marvin's Room by Drake, a sad song, is good even when you're happy. To improve D we could get more data because unsupervised HMM will only get better with more data.*

Problem F [5 points]: Generating Emission Sequences

Solution F.:

```
Generated Emission
######################################################################
46476371053143577440
15424360747754274277
77757324300052524765
71254150552474572252
75252522470724200575


File #1:
Generated Emission
######################################################################
47750720034367025747
46366104020775647750
31210245344137074403
17564174422405624047
72450275254573730422


File #2:
Generated Emission
######################################################################
79342046956196554905
81901716177422626646
59848065207020631061
37769529233595055394
27733797706650338372


File #3:
Generated Emission
######################################################################
22421031422634164661
53006505415511121320
36102121415221513132
26141256054133446642
54036112341111316532


File #4:
Generated Emission
######################################################################
50522624416422566356
34336532252334525233
41012601162262631160
41133225062632063152
65206006331511123326


File #5:
Generated Emission
######################################################################
55002683656782641831
46062412624166166816
64446303188844226254
05315620880666382104
66066104048533650816
```

Problem G [3 points]: Sparsity of Trained A and O Matrices

Solution G.: *Both of the matrices are very sparse. This makes sense because some transitions will be very unlikely like if one state is prepositions and one is adjectives. Each state has one or two likely transitions. The sparsity means there is a consistent transition from certain states to others. For example state 2 almost always goes to state one. Looking at the wordcloud we see these are adjectives and nouns respectively. This clearly makes sense.*

Problem H [5 points]: Hidden States vs. Sample Emission Sentences from HMM

Solution H.: *As the number of hidden states increases the sentences become more and more coherent. The special case when there is one hidden state, we are essentially choosing states at random. This is because all the states are within the one hidden state. In general, as the number of hidden states increases, the training data liklihood increases. This is because the number of parameters increases. If the max training likelihood is desired you should have one hidden state for each observation.*

Problem I [5 points]: Analyzing Visualization of State

Solution I.: *State 2 represents adjectives. This differs from other states because it's the only one that contains mostly adjectives. We can see that it trasitions to nouns almost all of the time. United, executive, judicial, as well as many numbers such as one, two, and thirty are in this word cloud.*