Chuong Vu

COMP 5800

Final Take Home

Prof. Byung Kim

**FINAL Take Home TEST**

1. Suppose a dishonest dealer has two coins, one fair and the other biased. The biased coin has heads with probability of ¼ in each coin toss. Assume that the dealer never switches the coins. Which coin is more likely to have been used when coin tosses produced HTTTHHHTTTTHTHHTT?

**Given:**

**Solve:**

Let X be the output HTTTHHHTTTTHTHHTT. To find the answer, we need to find

and

Where:

**Based on the Bayes Theory, we got**

From the distribution probabilities, Pr(Fair|X) > Pr(Bias|X). It shows that the fair coin have used more than the biased coin. Moreover, Pr(Head|X) = 7/17 and Pr(Tail|X) = 10/17 are closer to the Pr(Head|Fair) and Pr(Tail|Fair) than the Pr(Head|Bias) and Pr(Tail|Bias). **So it is more likely that the used coin is the Fair one.**

1. A two-state HMM is constructed from the measurements shown below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | G | T | C | Mean length |
| P | 2/5 | 2/5 | 1/10 | 1/10 | 10 |
| Y | 3/10 | 3/10 | 1/5 | 1/5 | 10 |

The mean length indicates the average time that the HMM stays in the state. Decode the most likely sequence of states (P/Y) for sequence GGCT.

**Answer**

Let probability matrix of state P/Y call PY:

|  |  |  |
| --- | --- | --- |
| PY Matrix | P | Y |
| P | 0.9 | 0.1 |
| Y | 0.1 | 0.9 |

Let probability matrix of measurements call S:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S Matrix | A | G | T | C |
| P | 0.4 | 0.4 | 0.1 | 0.1 |
| Y | 0.3 | 0.3 | 0.2 | 0.2 |

Since we have 2 states and no given probability for them, probability of each state that can be started initially is ½ or 0.5. Let probability matrix of starting state call s:

|  |  |  |
| --- | --- | --- |
| s Matrix | P | Y |
|  | 0.5 | 0.5 |

Given the sequence GGCT, the observation as: O = (O0,O1,O2,O3) = (G,G,C,T)

So the state sequence will be X=(X0,X1,X2,X3) of length of 4. Therefore, the probability of state X given the observation value is:

Pr(X0,O0) = sx0 \* Sx0 Pr(X1,O1) = PYx0-x1 \* Sx1 Pr(X2,O2) = PYx1-x2 \* Sx2

Pr(X3,O3) = PYx2-x3 \* Sx3 Pr(X,O) = Pr(X0,O0) \* P(rX1,O1) \* Pr(X2,O2) \* Pr(X3,O3)

We have the given observation sequence in length of four and there is two states involve in each sequence. Total possible number of states is 24 or 16. They are:

PPPP, PPPY, PPYP, PPYY, PYPP, PYPY, PYYP, PYYY, YPPP, YPPY, YPYP, YPYY, YYPP, YYPY, YYYP, YYYY. Using the above formula, we can calculate the probability of each state. For example, for state PPPP, the probability is:

Pr(X,O) = (0.5 \* 0.4) \* (0.9 \* 0.4) \* (0.9 \* 0.1) \* (0.9 \* 0.1) = 0.0005832

Follow the above calculation, the probability of states table is:

|  |  |  |  |
| --- | --- | --- | --- |
| States | Pr(X,O) | Calculated Probability | Normalized Probability |
| PPPP | 0.5\*0.4\*0.9\*0.4\*0.9\*0.1\*0.9\*0.1 | 0.0005832 | 0.215202952 |
| PPPY | 0.5\*0.4\*0.9\*0.4\*0.9\*0.1\*0.1\*0.2 | 0. 0001296 | 0.047822878 |
| PPYP | 0.5\*0.4\*0.9\*0.4\*0.1\*0.2\*0.1\*0.1 | 0.0000144 | 0.005313653 |
| PPYY | 0.5\*0.4\*0.9\*0.4\*0.1\*0.2\*0.9\*0.2 | 0. 0002592 | 0.095645756 |
| PYPP | 0.5\*0.4\*0.1\*0.3\*0.1\*0.1\*0.9\*0.1 | 0.0000054 | 0.00199262 |
| PYPY | 0.5\*0.4\*0.1\*0.3\*0.1\*0.1\*0.1\*0.2 | 0.0000012 | 0.000442804 |
| PYYP | 0.5\*0.4\*0.1\*0.3\*0.9\*0.2\*0.1\*0.1 | 0.0000108 | 0.00398524 |
| PYYY | 0.5\*0.4\*0.1\*0.3\*0.9\*0.2\*0.9\*0.2 | 0.0001944 | 0.071734317 |
| YPPP | 0.5\*0.3\*0.1\*0.4\*0.9\*0.1\*0.9\*0.1 | 0.0000486 | 0.017933579 |
| YPPY | 0.5\*0.3\*0.1\*0.4\*0.9\*0.1\*0.1\*0.2 | 0.0000108 | 0.00398524 |
| YPYP | 0.5\*0.3\*0.1\*0.4\*0.1\*0.2\*0.1\*0.1 | 0.0000012 | 0.000442804 |
| YPYY | 0.5\*0.3\*0.1\*0.4\*0.1\*0.2\*0.9\*0.2 | 0.0000216 | 0.00797048 |
| YYPP | 0.5\*0.3\*0.9\*0.3\*0.1\*0.1\*0.9\*0.1 | 0.0000364 | 0.013431734 |
| YYPY | 0.5\*0.3\*0.9\*0.3\*0.1\*0.1\*0.1\*0.2 | 0.0000081 | 0.00298893 |
| YYYP | 0.5\*0.3\*0.9\*0.3\*0.9\*0.2\*0.1\*0.1 | 0.0000729 | 0.026900369 |
| YYYY | 0.5\*0.3\*0.9\*0.3\*0.9\*0.2\*0.9\*0.2 | 0.0013122 | 0.484206642 |

From the table, two states that are highlighted in red are PPPP and YYYY. The normalized probability value of YYYY is highest and the value of PPPP is next lower highest.

Since the probability of starting state is 50% to both of P and Y, chance to start either in P or Y is split evenly. The sequence of state (P/Y) for sequence GGCT is YYYY and PPPP. To confirm which state it is really going to, the follow table is the calculated of the probability of each state.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pr/Column | 0 | 1 | 2 | 3 |
| Pr(P) | 0.44214 | 0.394317 | 0.303801 | 0.285203 |
| Pr(Y) | 0.55786 | 0.605683 | 0.696199 | 0.714797 |

From the above table, Pr(Y) in all states is higher than Pr(P). The most likely sequence of states (P/Y) for sequence GGCT is YYYY.

1. Hierarchical clustering

In HW3, PCA was performed for 20 AAs as outlined in [www.cs.uml.edu/~kim/580/pca\_AA.pdf](http://www.cs.uml.edu/~kim/580/pca_AA.pdf). In 2.6.2 of the article, the hierarchical clustering is performed for the 20 AAs. Repeat the hierarchical clustering for the 20 AAs with eight features, and compare the resulting clustering tree. Perform the clustering with the PLUTO package (<http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview>) and compare the resulting clustering trees.

**Answer:**

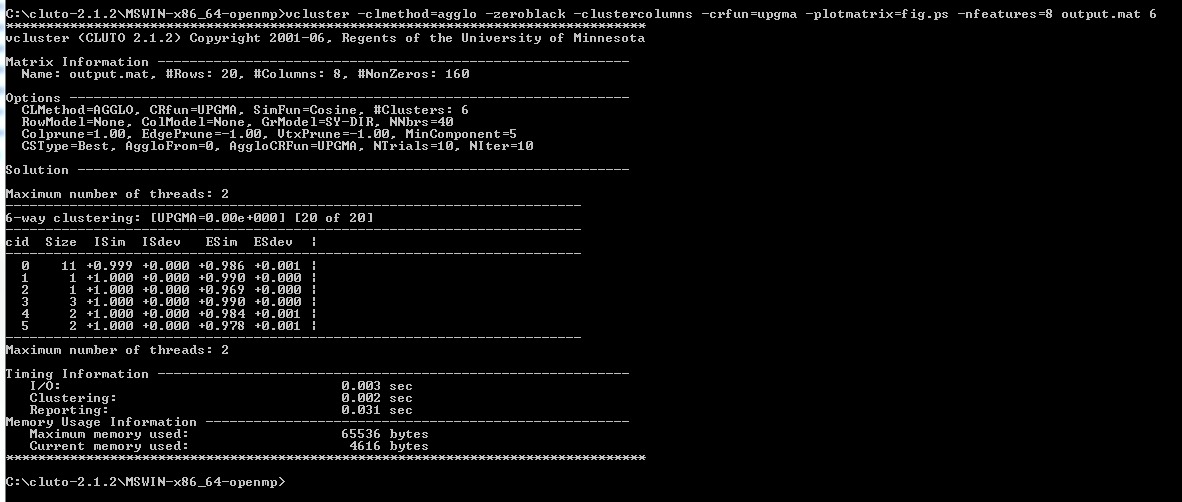
By using the cluto package, the command I use to do the Hierachical clustering is

vcluster -clmethod=agglo -zeroblack -clustercolumns -crfun=upgma -plotmatrix=fig.ps -nfeatures=8 output.mat 6

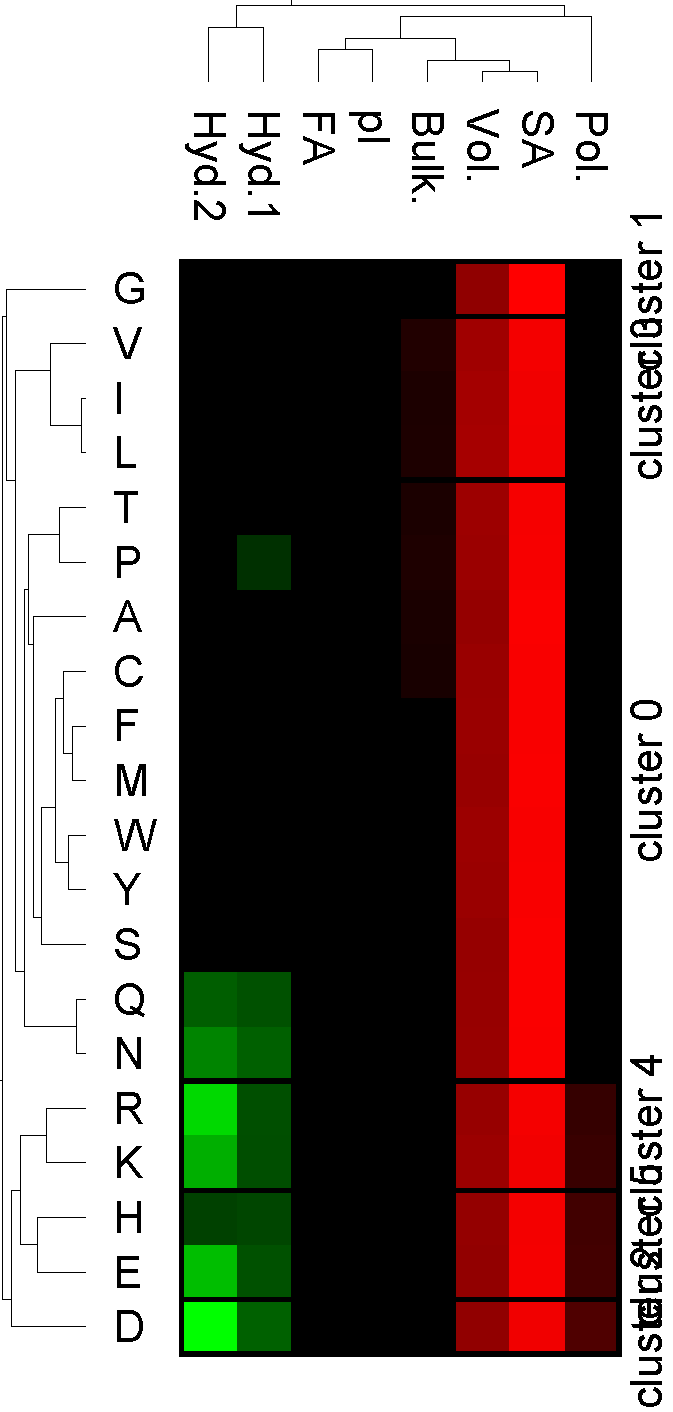
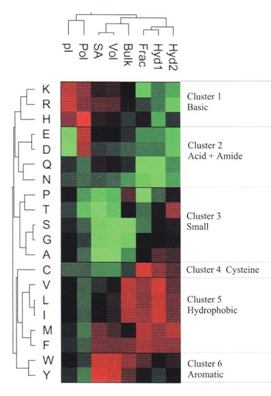
where:

-clmethod=agglo 🡺 method use is agglomerative

-crfun=upgma 🡺 criterion function is UPGMA also mean average

And similarity function by default is Cosine

Output from cluto: From 03\_pca.ppt



My output is totally different from the lecture. Here is the compare:

|  |  |
| --- | --- |
| My output (Counting Cluster from 1 to 6) | From Lecture |
| * L and I are very similar and clustered at the beginning, and then it combined with V later on. * In the means time, T, P, A, C, F, M, W, Y, S, Q, N combined together (cluster 1) and join to the VIL later on, which is much more than the right picture. * G is alone in Custer 2, and D alone in cluster 3. * H clustered with E in Cluster 6 * RK combined in clustered 5 * Overall:   Cluster 1: T, P, A, C, F, M, W, Y, S, Q, N  Cluster 2: G  Cluster 3: D  Cluster 4: V, I, L  Cluster 5: R, K  Cluster 6: H, E | * L and I are very similar and clustered at the beginning, and then it combined with V later on. * It joined with M, F and combined together in Cluster 5. * C is alone in Cluster 4 because Cysteine has an unusual role in protein structure because of its potential to form disulfide bonds between pairs of cysteine residues. * EDQN combined in Cluster 2 * Overall:   Cluster 1: K, R, H  Cluster 2: E, D, Q, N  Cluster 3: P, T, S, G, A  Cluster 4: C  Cluster 5: V, L, I, M, F  Cluster 6: W, Y |

So, I have tried several hierarchical cluster method using cluto, however the output is still not the same ask the question. So what I guess is maybe I have a different data than the question so the output turned out complete difference but still, the AA is clustered based on the similar properties and then all of them will combined to a larger clusters.

Several output but not same as the question:

|  |  |
| --- | --- |
| C:\Users\ChuongVu\Downloads\cluto-2.1.2\MSWIN-x86_64\fig1.png  vcluster -fulltree -clustercolumns -plotmatrix=fig1.ps -nfeatures=8 output.mat 6 | C:\Users\ChuongVu\Downloads\cluto-2.1.2\MSWIN-x86_64\fig2.png  vcluster -clmethod=agglo -clustercolumns -plotmatrix=fig2.ps -nfeatures=8 output.mat 6 |
|  |  |
| C:\Users\ChuongVu\Downloads\cluto-2.1.2\MSWIN-x86_64\fig3.png  vcluster -sim=corr -clmethod=agglo -clustercolumns -plotmatrix=fig3.ps -nfeatures=8 output.mat 6 | C:\Users\ChuongVu\Downloads\cluto-2.1.2\MSWIN-x86_64\fig4.png  vcluster -clmethod=agglo -zeroblack -plotmatrix=fig4.ps -nfeatures=8 output.mat 6 |