EEE – 8803

JEET KIRAN PAWANI
GTid- 903397407

PGM ASSIGNMENT – 4

1] Naive Bayes model with class C
Discrete evidence variables $X_1 \ldots X_n$.

CPD s $P(C=c) = \theta_c$ $\quad P(X_i = x \mid C=c) = \theta_{x_i|c}$ $\quad i = 1 \ldots n$

$x_i \in Val(X_i)$ $\quad c \in Val(C)$

Dataset $D = \{x[1], \ldots x[m]\}$ where $x[m]$ is a complete assignment of evidence variables $X_1 \ldots X_n$.

If parameters are assigned uniformly.

$$\theta_c^0 = \frac{1}{|Val(C)|} \qquad \theta_{x_i|c}^0 = \frac{1}{|Val(X_i)|} \text{ for all } x_i, c$$

TP – EM algo converges in one iteration.

Now lets start with step 1

$$\theta_c^1 = \frac{1}{m} \sum_m P(c \mid x(m), \theta^0) \rightarrow (2)$$

Each step of EM is given by $d(y|i) = P(y \mid x(i), \theta^{(t-1)})$ where $\theta = \theta^0$

Then new parameter is calculated by

$$q^t(y) = \frac{1}{n} \sum_{i=1}^{n} d(y|i) \rightarrow (1)$$

(1) being used to drive (2)

$$P(c \mid x(m), \theta^0) = \frac{P(c, x(m), \theta^0)}{P(x(m), \theta^0)}$$

$$= \frac{P(x(m) \mid c, \theta^0)\, P(c, \theta^0)}{P(x(m) \mid \theta^0)\, P(\theta^0)}$$

$$= \frac{P(x(m) \mid c, \theta^0)\, P(c \mid \theta^0)\, P(\theta^0)}{P(x(m) \mid \theta^0)\, P(\theta^0)}$$

$$P(c/x(m), \theta^0) = \frac{P(x[m]/c, \theta^0)\, P(c/\theta^0)}{P(x(m)/\theta^0)}$$

$$\Rightarrow \theta_c' = \frac{1}{m} \sum_m P(c/x(m), \theta^0)$$

$$= \frac{1}{m} \sum_m \frac{P(x(m)/c, \theta^0)\, P(c/\theta^0)}{P(x(m)/\theta^0)}$$

↑ adding c' term here

$$= \frac{1}{m} \sum_m \left[\frac{P(x(m)/c, \theta^0)\, P(c/\theta^0)}{\sum_{c' \in Val(c)} P(x(m)/c', \theta^0)\, P(c'/\theta^0)} \right]$$

we know $x(m)$ ranges from $\{x(1) \ldots x(n)\}$

bringing it to product form

$$= \frac{1}{m} \sum_m \left[\frac{P(c/\theta^0) \prod_{i=1}^{n} P(x_i(m)/c, \theta^0)}{\sum_{c' \in Val(c)} P(c'/\theta^0) \prod_{i=1}^{n} P(x_i(m)/c', \theta^0)} \right]$$

Given $P(c=c) = \theta_c$ $\quad P(x_i = a / c = c) = \theta_{x_i/c}$

$\Rightarrow P(c/\theta_c^0) = \frac{1}{Val(c)}$, $\quad P(x_i = x / c = c, \theta^0) = \frac{1}{Val(x_i)}$

$$\theta_c' = \frac{1}{m} \sum_m \left[\frac{\frac{1}{|Val(c)|} \prod_{i=1}^{n} \frac{1}{|Val(x_i)|}}{\sum_{c' \in Val(c)} \frac{1}{|Val(c)|} \prod_{i=1}^{n} \frac{1}{|Val(x_i)|}} \right]$$

$\prod_{i=1}^{n} \frac{1}{|Val(x_i)|}$ is independent of m

also $\sum_{c \in G} Val(c) \frac{1}{|Val(c)|} = 1$

$$\theta_c' = \frac{1}{m} \sum_m \left[\frac{1}{|Val(c)|} \right]$$ (2)

$$= \frac{1}{m} \times \frac{m}{|Val(c)|} = \frac{1}{|Val(c)|}$$

$$\boxed{\theta_c' = \frac{1}{|Val(c)|}}$$

we know $q_j^{\theta}(x'/y) = \frac{\sum_i c_i x_j^i = x \; d(y/i)}{\sum_i d(y/i)}$

Using this as we have $x_j^i = x$ so indicator function used.

$$\theta'_{x_i/c} = \frac{\sum_m P(c/x(m), \theta^0) \; 1(x_i(m) = x)}{\sum_m P(c/x(m), \theta^0)}$$

we know $P(c/x(m), \theta^0) = \frac{1}{|Val(c)|}$

$$= \frac{\sum_m \frac{1}{|Val(c)|} \; 1(x_i(m) = x)}{\sum_m \frac{1}{|Val(c)|}} \Rightarrow \frac{m}{|Val(c)|}$$ as c exists of m

$$= \frac{\sum_m \frac{1}{|Val(c)|} \; 1(x_i(m) = x)}{\frac{m}{|Val(c)|}}$$

$$= \frac{\sum_m 1(x_i(m) = x)}{m}$$ → this is indicator function when $x_i(m) = x$

as we know all are equally distributed

$\Rightarrow$ This can be written as —

$$\theta'_{x_i/c} = \frac{\sum_m 1(x_i(m) = x)}{m}$$

$\sum_m 1(x_i(m) = x)$ is number of times $x_i(m) = x$ occurs indicator.

This is for the step 1 now we try to generalize using induction.

To prove $\theta^t_{c=c} = \frac{1}{|Val(c)|}$

$$\theta^t_{x_i/c} = \frac{\sum_m 1(x_i(m) = x)}{m}$$

Now we try to prove θ^{t+1}_c as $\frac{1}{Val(c)}$

Similar to previous steps

$$\theta^{t+1}_{c=c} = \frac{1}{m} \sum P(c/x(m), \theta^t)$$

From (2) we can write this as

$$= \frac{1}{m} \sum_m \frac{P(x(m)/c, \theta^t) P(c/\theta^t)}{P(x(m)/\theta^t)} \rightarrow \text{adding } c'$$

$$= \frac{1}{m} \sum_m \frac{P(x(m)/c, \theta^t) P(c/\theta^t)}{\sum_{c' \in Val(c)} P(x(m)/c', \theta^t) P(c'/\theta^t)}$$

$x(m)$ writing as prod of $\prod_{i=1}^{\hat{}} x_i(m)$

$$\theta^{t+1}_{c=c} = \frac{1}{m} \sum_m \frac{P(c/\theta^t) \prod_{i=1}^{\hat{}} P(x_i(m)/c, \theta^t)}{\sum_{c' \in Val(c)} P(c'/\theta^t) \prod_{i=1}^{\hat{}} P(x_i(m)/c', \theta^t)}$$

$$P(x_i(m)/c, \theta^t) = P(x_i(m)/c', \theta^t) = \frac{1}{|Log(x_i)|}$$

$$P(c/\theta^t) = P(c'/\theta^t) = \frac{1}{|Val(c)|}$$

$$\theta_{c=c}^{t+1} = \frac{1}{m} \sum_{m} \frac{1}{|Val(c)|} \prod_{i=1}^{n} \frac{1}{|Val(x_i)|}$$

$$\sum_{c' \in Val(c)} \prod_{i=1}^{n} \frac{1}{|Val(x_i)|} \cdot \frac{1}{|Val(c)|} \Rightarrow 1$$

$$= \frac{1}{m} \sum_{m} \frac{1}{|Val(c)|} = \frac{1}{m} \cdot m \cdot \frac{1}{|Val(c)|} = \frac{1}{|Val(c)|}$$

$$\boxed{\theta_{c=c}^{t+1} = \frac{1}{|Val(c)|}}$$

$$\theta_{x_i=x/c=c}^{t+1} = \frac{\sum_{m} P(c/x(m), \theta^t) \; 1\{x_i(m)=x\} \cdot \frac{1}{|Val(c)|}}{\sum_{m} P(c/x(m), \theta^t)}$$

(from prev) $P(c/x(m), \theta^t) = \frac{1}{|Val(c)|}$

$$= \frac{\sum_{m} \frac{1}{|Val(c)|} \; 1(x_i(m)=x)}{\sum_{m} \frac{1}{|Val(c)|}}$$

$$\boxed{\theta_{x_i=x/c=c}^{t+1} = \frac{\sum_{m} 1(x_i(m)=x)}{m}}$$

Thus, by induction we proved that when uniformly initialized at 0^{th} step, EM algo converges in 1 step.

$$\theta_{c=c}^{t} = \frac{1}{|Val(c)|}$$

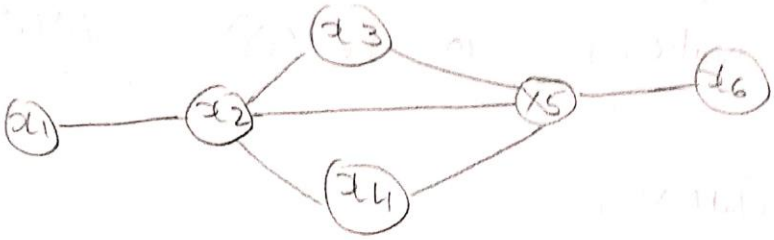
$$\theta_{x_i=x/c=c}^{t} = \frac{\sum_{m} 1(x(m)=x)}{m}$$

2] Dataset $\{x^n \quad n=1 \dots N\}$

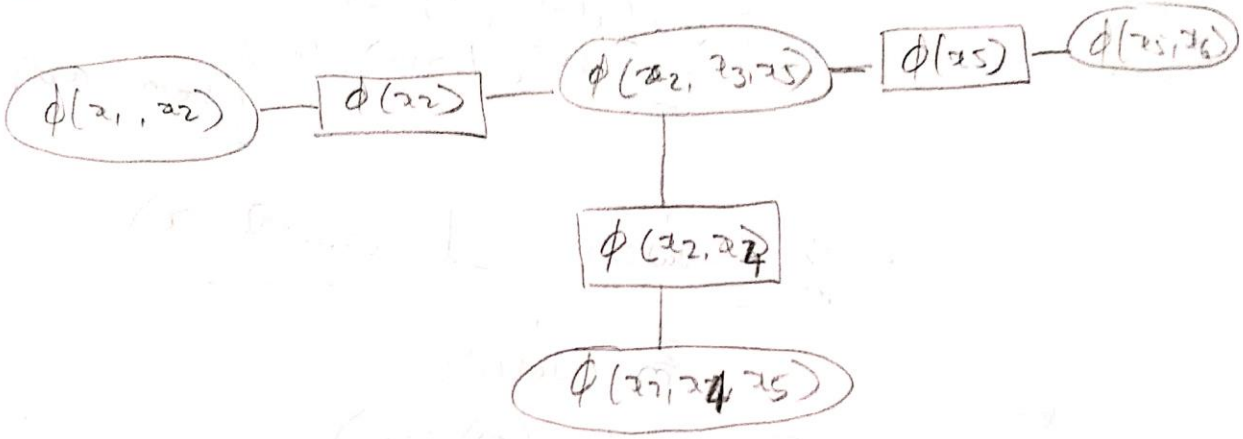
$x = \{x_1 \dots x_6\}$

empirical distribution $\hat{p}(x)$

$$p(x_1 \dots x_6) = \frac{1}{2} \phi_1(x_1, x_2)\ \phi_2(x_2, x_3, x_5)\ \phi_3(x_2, x_4, x_5)\ \phi_4(x_5, x_6)$$



Junction tree



from the above junction tree

$$p(x_1 \dots x_6) = \frac{\phi(x_1, x_2)\ \phi(x_2, x_3, x_5)\ \phi(x_2, x_4, x_5)\ \phi(x_5, x_6)}{\phi(x_2)\ \phi(x_5)\ \phi(x_2, x_4)}$$

$$\frac{\phi(x_1, x_2)}{\phi(x_2)} = \phi(x_1 / x_2) \qquad \frac{p(x_2, x_4, x_5)}{p(x_2, x_4)} = p(x_5 / x_2, x_4)$$

$$\frac{p(x_5, x_6)}{p(x_5)} = p(x_6 / x_5)$$

$p(x_1 \dots x_6) = p(x_1/x_2)\, p(x_2 x_3 x_5)\, p(x_4/x_2 x_5)$
$\rightarrow (1)\ p(x_6/x_5)$

Since we have identified clique potentials, the normalization constant $(z) = 1$

As we are joining bayesian networks or clusters the clique potentials are independent

Taking log likelihood -

$$L = \sum_n \log\left(p(x_1^n/x_2^n)\right) + \log p(x_2^n x_3^n x_5^n) + \log p(x_4^n/x_2^n, x_5^n) + \log p(x_6^n/x_5^n)$$

from (1)

$\phi(x_1, x_2) = p(x_1/x_2)$

$\phi(x_2 x_3 x_5) = p(x_2 x_3 x_5)$

$\phi(x_2 x_4 x_5) = p(x_4/x_2 x_5)$

$\phi(x_5 x_6) = p(x_6/x_5)$

Since all terms in dataset are independent as clique potentials are independent. The max. likelihood solution in this bayesian network can be simply setting each factor to its empirical distributions

$\phi(x_1 x_2) = \varepsilon(x_1/x_2)$

$\phi(x_2, x_3, x_5) = \varepsilon(x_2, x_3, x_5)$

$\phi(x_2, x_4, x_5) = \varepsilon(x_4/x_2, x_5)$

$\phi(x_5, x_6) = \varepsilon(x_6/x_5)$

3] N observations $x = x^1, \ldots x^N$

the log likelihood of belief network to generate x is

$$\log p(x) = \sum_{n=1}^{N} \sum_{i=1}^{K} \log p(x_i^n / pa(x_i^n))$$

$$\theta_s^i(t) = p(x_i = s / pa(x_i) = t)$$

representing prob that var x_i is in state s given parents of x_i are of states t.

Using Lagrangian

$$L = \sum_{n=1}^{N} \sum_{i=1}^{k} \log p(x_i^n / pa(x_i^n)) + \sum_{i=1}^{K} \sum_{t}^{i} \lambda_t^i \left(1 - \sum \theta_s^i(t^i)\right)$$

To show

$$\theta_s^i(t^i) = \frac{\sum_{n=1}^{N} \mathbb{I}[x_i^n = s] \, \mathbb{I}[pa(x_j^n) = t^j]}{\sum_{n=1}^{N} \sum_{s} \mathbb{I}[x_j^n = s] \, \mathbb{I}[pa(x_j^n) = t^j]}$$

$\Rightarrow$ We can write $\theta_s^i(t)$ in indicator function form

ie, $\theta_s^i(t) = p(x_i / pa(x_i)) \, \mathbb{I}[x_i = s] \, \mathbb{I}[pa(x_i) = t]$

$$\log p(x) = \sum_{n=1}^{N} \sum_{i=1}^{K} \log(x_i^n / pa(x_i^n))$$

Since we are using a discrete variable, the normalization constant $= 1$

$$\Rightarrow \sum_{s} \theta_s^i(t^i) = \sum_{s} p(x_i = s / pa(x_i) = t^i) = 1 \longrightarrow (1)$$

The log likelihood function in the form of θ_s^i can be written as $\approx \log(x_i^n / pa(x_i^n))$

$$L = \sum_{i=1}^{K} \sum_{n=1}^{N} \prod (x_i^n = s] \prod (pa(x_i^n) = t^i) \log \theta_s^i(t) + \sum_{i=1}^{K} \sum_{t^i} \lambda^i(t^i) \left(1 - \sum_s \theta_s^i(t^i)\right)$$

Now we differentiate wrt $\theta_s^i(t^i)$ and set the equations to 0.

$$\frac{\partial L}{\partial \theta_s^i(t^i)} = 0$$

$$\Rightarrow \frac{\partial L}{\partial \theta_s^j(t^j)} = \sum_{n=1}^{N} \prod (x_j^n = s) \prod (pa(x_j^n) = t^j) \frac{1}{\theta_s^j(t^j)} - \lambda^j(t^j)$$

$$\frac{\partial (\text{constant}) = 0}{\partial \theta_s^j(t^j)}$$

$$\Rightarrow \sum_{n=1} \prod (x_i^n = s) \prod (pa(x_j^n) = t^j) \frac{1}{\theta_s^j(t^j)} - \lambda^j(t^j) = 0$$

$$\Rightarrow \lambda^j(t^j) = \sum_{n=1} \prod (x_j^n = s) \prod (pa(x_j^n) = t^j) \frac{1}{\theta_s^j(t^j)}$$

or rearranging

$$\theta_s^j(t^j) = \frac{1}{\lambda^j(t^j)} \sum_{n=1} \prod (x_i^n = s) \prod (pa(x_j^n) = t^j)$$

here $\lambda^j(t^j)$ is the normalizing constant

$$\Rightarrow \lambda^j(t^j) = 1 \Rightarrow \lambda^j(t^j) = \sum_s \theta_s^j(t^j)$$

$$\rightarrow \theta_s^{ij}(t^i) = \frac{\sum_{n=1}^{N} \mathbb{I}(x_j^n = s)\, \mathbb{I}(pa(x_j^n) = t^i)}{\sum_{s} \theta_s^{ij}(t^i)}$$

$$= \frac{\sum_{n=1}^{N} \mathbb{I}(x_j^n = s)\, \mathbb{I}(pa(x_j^n) = t^i)}{\sum_{s} p(x_j = s \mid pa(x_j) = t^i)}$$

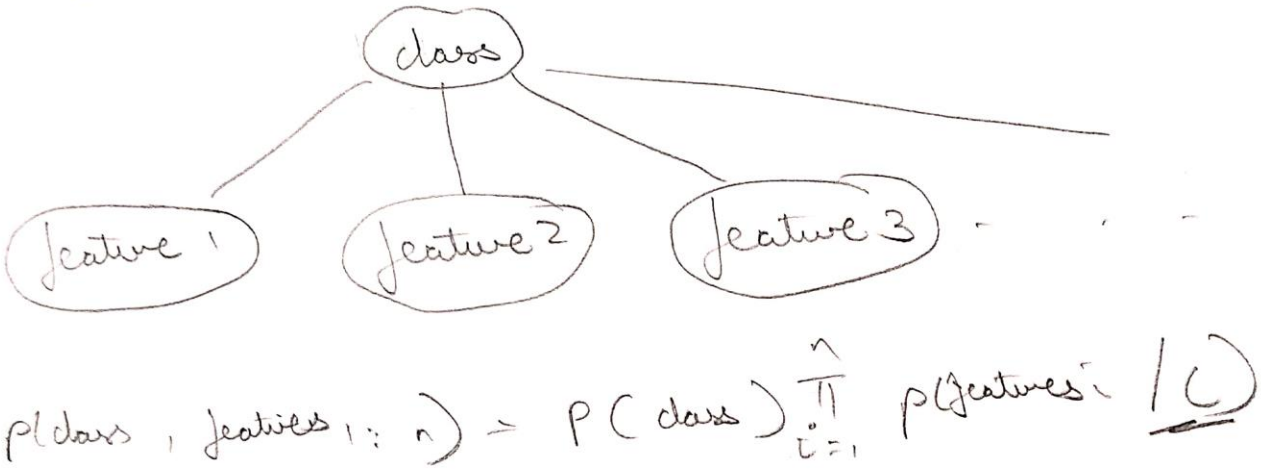
$$= \frac{\sum_{n=1}^{N} \mathbb{I}(x_j^n = s)\, \mathbb{I}(pa(x_j^n) = t^i)}{\sum_{s} \mathbb{I}(x_j = s)\, \mathbb{I}(pa(x_j) = t^i))}$$

written in form of indicator functions

taking sum all all x!

$$\theta_s^{ij}(t^i) = \frac{\sum_{n=1}^{N} \mathbb{I}(x_j^n = s)\, \mathbb{I}(pa(x_j^n) = t^i)}{\sum_{n=1}^{N} \sum_{s} \mathbb{I}(x_j^n = s)\, \mathbb{I}(pa(x_j^n) = t^i))}$$

4] Graphical Bayesian model for Naive Bayes setup



$$p(\text{class}, \text{features}_{1:n}) = P(\text{class}) \prod_{i=1}^{n} p(\text{features}_i \mid C)$$

```
Input Test Data

xd =

     1     0     0     1     1     1     1     0

Probability of Input test data being about politics is

ProbOfPolitics =

    0.8306

fx >>
```

First I load all the Sequences for Politics and Sports.
Find mean of each row.
Then load the test datapoint.
Then I use the for formula probPolitics * (meanPolitics.^(testdatapoints) .* (1-meanPolitics).^(1-testdatapoints))

Probability of Sequence being Politics is = 0.8306

5)

```
Probability of a drum unit problem given the evidence is:
First value for No and second value for yes  using Joint probability found by me
3.810115e-01 No
6.189885e-01 Yes
First value for No and second value for yes  using multpot function
3.810115e-01 No
6.189885e-01 Yes
```

In this Question, I first make the network in the form of Brml array and then keep random probabilities to all.
I have set a tolerance value of 0.0001 and number of iterations of EM to 30
It converges after almost 15-17 iterations.

Now we have new potentials of Each data. I have self found joint probability and then even used the multipot function of brml library. Both answers are matching.
The condition given is quality of paper is bad , but no bad smell or wrinkled paper. I have set that according, and used the condpot function of brml library and obtained the values of whether it is a drum problem or not.

Probability that it is drum problem is 0.6189885
Probability that it is not a drum problem 0.3810115

6 Structure learning of tree augmented Naive Bayes

Class variable C and feature variables $X_1 \ldots X_n$

a) Class variable C is connected to X_i's and also, X_i's have one additional parent. As per the hint given by TA, this can be ~~maximized~~ solved by maximizing the BIC score.

$$\Delta(G) = score_{BIC}(G : D) - score_{BIC}(G' : D)$$

we have been given one parent as C and one additional parent say X_j where j is anything except the node itself.

This can be maximized by

$$\Delta(G) = Score_B(G : D) - score(G' : D)$$

G' is naive bayes graph with only C as parent

G is a graph with parents C and X_j as stated above

Lets define weight $w_{j \to i}$ as weight of the edge which converts X_j to X_i

$$w_{j \to i} = FamScore(X_i | X_j, C : D) - famScore(X_i | C : D)$$

we know $FamScore(X_i | C : D)$

$$= m\left(\hat{I}_P(X_i; C) - \hat{H}(X)\right] - \frac{\log m}{2}\left(|Val(X) - 1| \, |Val(C)|\right)$$

On substituting we get

$$w_{x_j \to x_i} = m\left[\hat{I}_p(x_i : x_j, C) - \hat{H}(x_i) - \frac{\log m}{2}\left(|Val(x_i)-1||Val(x_j)-1||C|\right)\right]$$

$$= m\left(\hat{I}_p(x_i : C) - \hat{H}(x_i) + \frac{\log m}{2}\left(|Val(x_i)-1||C|\right)\right.$$

$$\left. - \hat{I}_p(x_i ; C)\right]$$

$$= m\left(\hat{I}_p(x_i : x_j, C) - \hat{I}_p(x_i ; C)\right] - \frac{\log m}{2}(2 \cdot 2 \cdot 2) + \frac{\log m}{2}(2 \cdot 2)$$

as all x_i, x_j, C are binary (2 possibilities)

$$\Rightarrow w_{j \to i} = m\left[\hat{I}_p(x_i ; x_j, C) - \hat{I}_p(x_i ; C)\right] - \log m$$

Now, we need to prove $w_{j \to i} = w_{i \to j}$

$$I_p(x_i ; x_j, C) - I_p(x_i ; C)$$

$$= H(x_i) + H(x_j, C) - H(x_i, x_j, C)$$

$$- H(x_i) - H(C) + H(x_i, C)$$

adding $H(x_j)$

$$= H(x_j) + H(x_j, C) - H(x_i, x_j, C)$$

$$- H(x_j) - H(C) + H(x_i, C)$$

$$= I(x_j ; x_i, C) - I(x_j ; C)$$

$$\Rightarrow \boxed{w_{i \to j} = w_{j \to i}}$$

Thus, an undirected graph with x_i nodes and edge weights $w_{j \to i}$ connecting to other x_i nodes can be made.

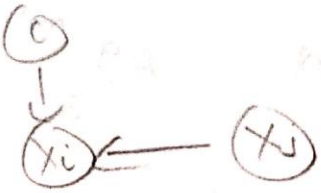
We can now choose x_i to be root and by using maximum spanning tree algo, the graph can be connected to a directed graph and then finally a edge can be added from C to x_i.

b) we must allow more flexibility for each x_i we need to decide whether to add C as parent or not. So, we must consider all incoming connections to node x_i.

these 3 possibilities can be

a) only C as parent

b) C and x_j as parent



c) only x_j as parent

maximize
ΔG = score with parents
- score without parents

Case ① only C as parent

$$\Delta(G) = \text{famScore}(x_i / C : D) - \text{famScore}(x_i : D)$$

$$= m\left(\hat{I}_P(x_i, C) - \hat{I}(x_i)_D\right) - \log(m) + \frac{\log m}{2}$$

$$\boxed{\Delta(G) = m\left(\hat{I}_P(x_i, C)\right) - \frac{\log|m|}{2}}$$

Case ②

x_j and C as parents

$$\Delta(G) = \text{famScore}(x_i / x_j, C : D) - \text{famScore}(x_i : D)$$

$$= m\left[\hat{I}_P(x_i ; x_j, C)\right] - 2\log|m| + \frac{\log(m)}{2}$$

$$\boxed{\Delta(G) = m\left[\hat{I}_P(x_i ; x_j, C)\right] - 1.5 \log(m)}$$

Case 3 only C as parent

$$\Delta G = \text{famScore}(x_i / C : D) - \text{famScore}(x_i : D)$$

$$= m\left(\hat{I}_P(x_i ; C)\right) - \log|m| + \frac{\log(m)}{2}$$

$$\boxed{\Delta G = m\left(\hat{I}_P(x_i ; C)\right) - \frac{\log|m|}{2}}$$

So, now we consider a graph which connects node C to X_i and all $X_j \to X_i$ for all nodes X_i of graph. (4)

⇒ We have a fully connected directed graph which if X_i taken as roots, has C as parent as well as all X_j's are parents

weights can be set as

$w_{C \to i} = \Delta(G')_c$

$w_{j \to i}$ = maximum of ($\Delta(G)$ when X_j and C are parents) or ΔG when only X_j is parent)

⇒ This is clearly seen as a directed graph. so we can run the maximum weighted directed spanning tree algo to get a Bayesian network. Due to regularization in BIC scoring, there are possibility of negative scores as well. We need to 0 them out. Thus a graph with maximum difference in BIC scores is created as this is a sum of weights of the graph.

Thus after finding the max. directed spanning graph F, the parent set of each feature X_i in the resulting network can be defined as

Parent in graph wrt X_i = $\begin{cases} \{C\} & \text{if } C \rightarrow X_i \in F \\ \{X_j, C\} & \text{if } X_j \rightarrow X_i \in F \text{ and } \Delta\{X_j, X_i, C\} \text{ scores} \\ \{X_j\} & \text{if } X_j \rightarrow X_i \in F \text{ and } \Delta\{X_j, C\} \text{ scores} \\ \phi, & \text{if none of the above are satisfied.} \end{cases}$

7 and 8

(1) sequence $v_1 \dots v_T, v_t \in \{1, \dots V\}$ is generated by markov chain. Single chain of length T, we have

$$p(v_1 \dots v_T) = p(v_1) \prod_{t=1}^{T-1} p(v_{t+1} / v_t)$$

The sequence of visible variables is given by

$$v = (v_1 \dots v_T)$$

For a single markov chain labelled by h,

$$p(v/h) = p(v_1/h) \prod_{t=1}^{T-1} p(v_{t+1} / v_t, h) \rightarrow (1)$$

In total, there are a set of H such markov chains ($h = 1 \dots H$)

Distr of visible variables is

$$p(v) = \sum_{h=1}^{H} p(v/h)\, p(h) \rightarrow (2)$$

Set of training variables v^n, $n = 1 \dots N$

v^n independently drawn.

The energy function of $p(v^n)$ is given as:

$$\sum_n \langle \log p(v^n) \rangle_{p(h/v^n)}$$

from (2)

$$= \sum_n \langle \log p(v^n/h)\, p(h) \rangle_{p(h/v^n)}$$

from (1)

$$\sum_n \langle \log p(v^n) \rangle_{p^{old}(h/v^n)} = \sum_n \sum_{t=1}^{T-1} \langle \log (v_{t+1}^h / v_t^n, h) \rangle_{p^{old}(h/v^n)}$$

in log π changes to Σ

$$+ \sum_n \langle \log p(v_1^n / h) \rangle_{p^{old}(h/v^n)}$$

$$+ \sum_{t=1}^{T-1} \sum_n \langle \log p(h) \rangle_{p^{old}(h/v^n)}$$

This needs to be optimized and is possible by differentiating wrt transitions

$p(v_{t+1} = i \ / v_t = j, h) =$ let this be θ_{ij}^h

$i =$ new state

$j =$ old state

Initial distribution is given by $p(v_1/h)$ and $p(h)$ is prior, so these become 0 or give no contribution to transitions.

The contribution of transition of energy with lagrange term is

$$L = \sum_n \sum_{t=1}^{T-1} \sum_{i,j} \sum_h p^{old}(h/v^n) \prod (v_{t+1}^h = i \& v_t^h = j) \log \theta_{ij}^h$$

$$+ \sum_h \sum_j \lambda_j \sum_i (1 - \theta_{ij}^h)$$

states

lagrange

Differentiating wrt θ_{ij}^h and setting to 0

$$\Rightarrow \sum_n \sum_{t=1}^{T-1} p^{old}(h/v^n) \, \mathbb{I}[v_t^n = j, v_{t+1}^n = i] \frac{1}{\theta_{ij}^h} - \lambda_j = 0$$

$$\lambda_j = \frac{\sum_n \sum_{t=1}^{T-1} p^{old}(h/v^n) \, \mathbb{I}[v_t^n = j, v_{t+1}^n = i]}{\theta_{ij}^h}$$

On rearranging

$$\theta_{ij}^h = \frac{\sum_n \sum_{t=1}^{T-1} p^{old}(h/v^n) \, \mathbb{I}[v_t^n = j, v_{t+1}^n = i]}{\lambda_j}$$

we have used normalization, as this is discrete,

$$\Rightarrow \sum_i \theta_{ij}^h = 1$$

$$\Rightarrow \boxed{\theta_{ij}^h = \frac{\sum_n \sum_{t=1}^{T-1} p^{old}(h/v^n) \, \mathbb{I}[v_t^n = j, v_{t+1}^n = i]}{\sum_i \sum_n \sum_{t=1}^{T-1} p^{old}(h/v^n) \, \mathbb{I}[v_t^n = j, v_{t+1}^n = i]}}$$

This gives us the weighted numbers of j to i transitions across all the sequences

Also, initial distribution $p(v_1 = i/h)$ can θ_i and the optimum energy wrt initial distri θ_i is given by

By following previous steps

$$\theta_i = \frac{\sum_n p^{old}(h/v^n) \prod [v_i^n = i]}{\sum_i \sum_n p^{old}(h/v^n) \prod [v_i^n = i]}$$

The optimal prior $p(h)$ is given by

$$p(h) = \frac{\sum_n p^{old}(h/v^n)}{\sum_h \sum_n p^{old}(h/v^n)}$$

8)
Sequence belongs to Group 1
Sequence is: CATAGGCATTCTATGTGCTG
Sequence is: CCAGTTACGGACGCCGAAAG
Sequence is: CGGCCGCGCCTCCGGGAACG
Sequence is: ACATGAACTACATAGTATAA
Sequence is: GTTGGTCAGCACACGGACTG
Sequence is: CACTACGGCTACCTGGGCAA
Sequence is: CGGTCCGTCCGAGGCACTCG
Sequence is: CACCATCACCCTTGCTAAGG
Sequence is: CAAATGCCTCACGCGTCTCA
Sequence is: GCCAAGCAGGGTCTCAACTT
Sequence is: CATGGACTGCTCCACAAAGG

Sequence belongs to Group 2
Sequence is: TGGAACCTTAAAAAAAAAAA
Sequence is: GTCTCCTGCCCTCTCTGAAC
Sequence is: GTGCCTGGACCTGAAAAGCC
Sequence is: AAAGTGCTCTGAAAACTCAC
Sequence is: CCTCCCCTCCCCTTTCCTGC
Sequence is: TAAGTGTCCTCTGCTCCTAA
Sequence is: AAAGAACTCCCCTCCCTGCC
Sequence is: AAAAAAACGAAAAACCTAAG
Sequence is: GCGTAAAAAAAGTCCTGGGT

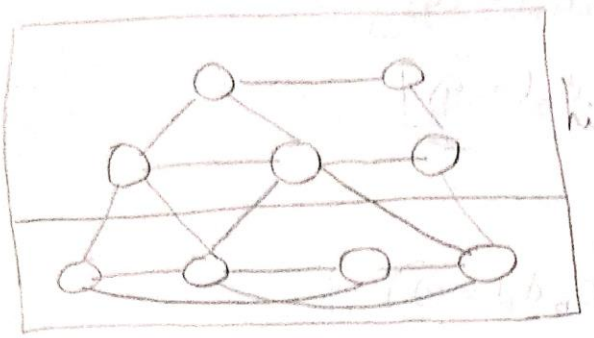
```
Sequence is:
CATAGGCATTCTATGTGCTG
Sequence belongs to Group 1
Sequence is:
CCAGTTACGGACGCCGAAAG
Sequence belongs to Group 1
Sequence is:
TGGAACCTTAAAAAAAAAAA
Sequence belongs to Group 2
Sequence is:
GTCTCCTGCCCTCTCTGAAC
Sequence belongs to Group 2
Sequence is:
GTGCCTGGACCTGAAAAGCC
Sequence belongs to Group 2
Sequence is:
CGGCCGCGCCTCCGGGAACG
Sequence belongs to Group 1
Sequence is:
AAAGTGCTCTGAAAACTCAC
Sequence belongs to Group 2
Sequence is:
ACATGAACTACATAGTATAA
Sequence belongs to Group 1
Sequence is:
GTTGGTCAGCACACGGACTG
Sequence belongs to Group 1
Sequence is:
CCTCCCCTCCCCTTTCCTGC
Sequence belongs to Group 2
Sequence is:
CACTACGGCTACCTGGGCAA
Sequence belongs to Group 1
Sequence is:
CGGTCCGTCCGAGGCACTCG
Sequence belongs to Group 1
Sequence is:
TAAGTGTCCTCTGCTCCTAA
Sequence belongs to Group 2
Sequence is:
CACCATCACCCTTGCTAAGG
Sequence belongs to Group 1
```

```
Sequence is:
GCCAAGCAGGGTCTCAACTT
Sequence belongs to Group 1
Sequence is:
CATGGACTGCTCCACAAAGG
Sequence belongs to Group 1
Sequence is:
AAAAAAACGAAAAACCTAAG
Sequence belongs to Group 2
Sequence is:
GCGTAAAAAAAGTCCTGGGT
Sequence belongs to Group 2
Log Likelihood is

loglikelihood =

 -483.6352
```

In this question I have just loaded the sequences.mat file. The visible nodes, A, C,G , T are set as 1,2 3 4 respectively. I then call the mixMArkov function with iterations of 50.

Then in this function we get a value phgv if that is above 0.5 then the sequence belongs to cluster 2 else it belongs to cluster one. Results are attached as above.

9



5 hidden units , 4 visible units

$$P(x,y) = \frac{1}{Z} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)$$

$\phi_k(x,y)$ are pairwise potentials for Boltzmann machines

θ_k are weights of BMs

To show $\frac{\partial \log P(x)}{\partial \theta_L} = \sum_y \phi_L(x,y) P(y/x) - \sum_{x,y} \phi_L(x,y) P(x,y)$

$\Rightarrow$ we know Z's normalization constant

$$\Rightarrow P(x,y) = \frac{\exp\left(\sum_k \theta_k \phi_k(x,y)\right)}{\sum_{x,y} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)}$$

$$\Rightarrow Z = \sum_{x,y} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)$$

to get $P(x)$ we have to marginalize $P(x,y)$

$$P(x) = \sum_y P(x,y)$$

$$P(x) = \sum_y \left[\frac{\exp\left(\sum_k \theta_k \phi_k(x,y)\right)}{\sum_{x,y} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)} \right]$$

$$P(x) = \frac{\sum_y \exp\left(\sum_k \theta_k \phi_k(x,y)\right)}{\sum_{x,y} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)}$$

- taking log

$$\log P(x) = \log\left[\frac{\sum_y \exp\left(\sum_k \theta_k \phi_k(x,y)\right)}{\sum_{x,y} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)}\right]$$

$$= \log\left[\sum_y \exp\left(\sum_k \theta_k \phi_k(x,y)\right)\right] - \log\left[\sum_{x,y} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)\right]$$

differentiating wrt θ_k

$$\frac{\partial \log P(x)}{\partial \theta_k} = \frac{\partial}{\partial \theta_k}\left[\log\left(\sum_y \exp\left(\sum_k \theta_k \phi_k(x,y)\right)\right) - \log\left(\sum_{x,y} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)\right)\right]$$

$$= \sum_y \frac{1}{\sum_y \exp\left(\sum_k \theta_k \phi_k(x,y)\right)} \exp\left(\sum_k \theta_k \phi_k(x,y)\right) \phi_k(x,y) \quad \left(\sum_y \exp\left(\sum_k \theta_k \phi_k(x,y)\right) \to P(x)\right)$$

$$- \sum_{x,y} \frac{\exp\left(\sum_k \theta_k \phi_k(x,y)\right)}{\sum_{x,y} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)} \phi_k(x,y) \quad \left(\exp\left(\sum_k \theta_k \phi_k(x,y)\right) \to P(x,y)\right)$$

$$\frac{\partial e^{2x}}{\partial x} = 2e^{2x}$$

$$P(x,y) = \frac{1}{z} \exp\left(\sum_k \theta_k \phi_k(x,y)\right)$$

$$P(x) = \frac{1}{z} \sum_y \exp\left(\sum_k \theta_k \phi_k(x,y)\right)$$

$$\frac{\partial \log P(x)}{\partial \theta_k} = \sum_y \frac{P(x,y)}{P(x)} \phi_k(x,y) - \sum_{x,y} P(x,y)\, \phi_k(x,y)$$

$$\frac{\partial \log P(x)}{\partial \theta_k} = \sum_y P(y/x)\, \phi_k(x,y) - \sum_{x,y} P(x,y)\, \phi_k(x,y)$$

$$\frac{\partial \log P(x)}{\partial \theta_i} = \sum_y P(y/x)\, \phi_i(x,y) - \sum_{x,y} P(x,y)\, \phi_i(x,y)$$