4.1 Maximum likelihood estimation

(a)

$$P(Y=y) = \frac{\text{count}(Y=y)}{T}$$

$$P(X=x \mid Y=y) = \frac{\text{count}(X=x, Y=y)}{\sum_{x'} \text{count}(X=x', Y=y)}$$

$$P(Z=z \mid Y=y) = \frac{\text{count}(Y=y, Z=z)}{\sum_{z'} \text{count}(Y=y, Z=z')}$$

(b)

$$P(Z=z) = \frac{\text{count}(Z=z)}{T}$$

$$P(Y=y \mid Z=z) = \frac{\text{count}(Y=y, Z=z)}{\sum_{y'} \text{count}(Y=y', Z=z)}$$

$$P(X=x \mid Y=y) = \frac{\text{count}(X=x, Y=y)}{\sum_{x'} \text{count}(X=x', Y=y)}$$

(c)

in (a):

$$P(X=x, Y=y, Z=z) = P(Y=y)P(X=x \mid Y=y)P(Z=z \mid Y=y)$$

$$= \frac{\cancel{\text{count}(Y=y)}}{T} \frac{\text{count}(X=x, Y=y)}{\cancel{\sum_{x'} \text{count}(X=x', Y=y)}} \frac{\text{count}(Y=y, Z=z)}{\sum_{z'} \text{count}(Y=y, Z=z')}$$

$$= \frac{\text{count}(X=x, Y=y)\,\text{count}(Y=y, Z=z)}{T\,\text{count}(Y=y)}$$

in (b):

$$P(X=x, Y=y, Z=z) = P(Z=z)\,P(Y=y \mid Z=z)\,P(X=x \mid Y=y)$$

$$= \frac{\cancel{count(Z=z)}}{T} \frac{count(Y=y, Z=z)}{\cancel{\sum_{y'} count(Y=y', Z=z)}} \frac{count(X=x, Y=y)}{\sum_{x'} count(X=x', Y=y)}$$

$$= \frac{count(X=x, Y=y)\, count(Y=y, Z=z)}{T\, count(Y=y)}$$

therefore:

they have same joint distribution over X, Y and Z.

(d)

No. Because

$$P(\vec{F}=\vec{f} \mid \vec{C}=\vec{c}) = \frac{P(\vec{F}=\vec{f}, \vec{C}=\vec{c})}{P(\vec{C}=\vec{c})}$$

$$= \frac{\sum_{(X,Y,Z)\setminus\{\vec{F},\vec{C}\}} P(X=x, Y=y, Z=z)}{\sum_{(X,Y,Z)\setminus\{\vec{C}\}} P(X=x, Y=y, Z=z)}$$

Because in part(c) we prove that $P(X=x, Y=y, Z=z)$ maybe same. The $P(\vec{F}=\vec{f} \mid \vec{C}=\vec{c})$ maybe no relevant to DAGs.

4.2 Markov modeling

(a)

$$P_u(S) = \prod_{l=1}^{L} P_1(I_l)$$

Log-likelihood:

$$\mathcal{L}(S) = \sum_{l=1}^{L} \log P_1(I_l)$$

$$= \sum_{\pi} \text{count}(\pi) \log P_1(\pi)$$

Using the answer of HW1, when $\frac{\text{count}(\pi)}{P_1(\pi)} = \text{Const}$, then $\mathcal{L}$ get largest number:

I	a	b	c	d	
$P_1(I)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	

(b)

The BN is as follows:

$$I_1 \to I_2 \to I_3 \to \cdots \to I_{16}$$

Because all variables are visible, we can use ML to get the answer:

$P_1(T_1) = \text{count}(T_1)$

$$P_2(T'|T) = \frac{\text{count}(T_l, T_{l-1})}{\sum'_{T'_l} \text{count}(T'_l, T_{l-1})} \quad (l = 2, \cdots, 16)$$

$P'_2(T'\|T)$	a	b	c	d
a	aa $\frac{1}{2}$	ab $\frac{1}{4}$	ac 0	ad $\frac{1}{4}$
b	ba 0	bb $\frac{3}{4}$	bc $\frac{1}{4}$	bd 0
c	ca 0	cb 0	cc $\frac{2}{3}$	cd $\frac{1}{3}$
d	da $\frac{1}{4}$	db 0	dc $\frac{1}{4}$	dd $\frac{1}{2}$

(c)

The possible form is contained in the (b) table, Obviously the T_2, T_3 underlined string are 0 in the table

We may find that in S, T_1, T_2, T_3, the number of every letter is same. Therefore, is the number of letters is n, then obviously:

$$P_u(T) = \left(\frac{1}{n}\right)^L$$

Therefore, when n increases, then $P_u(T)$ decreases.

In conclusion,

$$P_u(S) < P_u(T_1)$$

$$P_u(S) = P_u(T_2)$$

$$P_u(S) < P_u(T_3)$$

$P_B(T)$ is relevant to the kinds of two-letter string. If there are more kinds of it, the $P_B(T)$ will be smaller.

In conclusion.

$$P_B(T_1) > P_B(S)$$

$$P_B(T_2) > P_B(S)$$

$$P_B(T_3) > P_B(T_2)$$

Because the factor of $P_u(S)$, $P_u(T_2)$ is $\frac{1}{4}$, and factor of $P_B(S)$, $P_u(T_2)$ is bigger than $\frac{1}{4}$.

Therefore,

$P_u(S) < P_B(S)$

$P_u(T_2) < P_B(T_2)$

Simillarly, the factor of $P_u(T_1)$ and $P_u(T_3)$ is 1, and the factor of $P_B(T_1)$ and $P_B(T_3)$ is $\frac{1}{2}$. Therefore,

$P_u(T_1) < P_B(T_1)$

$P_u(T_3) < P_B(T_3)$