Name: Yuhan Shao
Student Number: 1002281327
UTORid: shaoyuha

# CSC411 Assignment 6

## PART I:

**1.** • Derive the M-step update rules for $\Theta$ and $\pi$ by setting partial derivative of the following equation to 0.

$$\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\left[\log \Pr(z^{(i)}=k) + \log p(\mathbf{x}^{(i)} \mid z^{(i)}=k)\right] + \log p(\boldsymbol{\pi}) + \log p(\boldsymbol{\Theta})$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\left[\log(\pi_k) + \log p(\vec{x}^{(i)} \mid z^{(i)}=k)\right] + \log p(\vec{\pi}) + \log p(\vec{\theta}) \qquad (*)$$

• Consider partial derivative w.r.t. $\pi_k$.

→ Ignore terms in $(*)$ not influence $\pi_k$, i.e: wants to maximize:

$$\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log P(z^{(i)}=k) + \log p(\vec{\pi})$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log \pi_k + \log p(\vec{\pi})$$

$$\propto \left(\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log \pi_k\right) + \left(\log \prod_{k=1}^{K} \pi_k^{a_k-1}\right) = \left(\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log \pi_k\right) + \left(\sum_{k=1}^{K}(a_k-1)\log \pi_k\right) \qquad (**)$$

→ Since $(**)$ subject to the constraint $\sum_k \pi_k = 1$, therefore, we can use Lagrangian to compute max and set the partial derivative to zero.

• Let $L = \left(\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log \pi_k\right) + \left(\sum_{k=1}^{K}(a_k-1)\log \pi_k\right) + \lambda\left(1 - \sum_{k=1}^{K}\pi_k\right)$

Then, $\frac{\partial L}{\partial \pi_k} = \frac{\sum_{i=1}^{N} r_k^{(i)}}{\pi_k} + \frac{(a_k-1)}{\pi_k} - \lambda.$

• Let $\frac{\partial L}{\partial \pi_k} = 0$, we have:

$$\lambda = \frac{\left(\sum_{i=1}^{N} r_k^{(i)}\right) + (a_k-1)}{\pi_k} \quad \text{for } \forall k. \quad \text{i.e: } \pi_k = \frac{\left(\sum_{i=1}^{N} r_k^{(i)}\right) + (a_k-1)}{\lambda} \qquad (***)$$

• Plug $(***)$ into the constraint $\sum_{k=1}^{K}\pi_k = 1$, we have:

$$\sum_{k=1}^{K}\pi_k = \sum_{k=1}^{K}\frac{\left(\sum_{i=1}^{N} r_k^{(i)}\right) + (a_k-1)}{\lambda} = \frac{\sum_{k=1}^{K}\left(\left(\sum_{i=1}^{N} r_k^{(i)}\right) + (a_k-1)\right)}{\lambda} = 1$$

$$\Rightarrow \sum_{k=1}^{K}\left[\left(\sum_{i=1}^{N} r_k^{(i)}\right) + (a_k-1)\right] = \lambda$$

→ Therefore, we have:

$$\frac{\left(\sum_{i=1}^{N} r_k^{(i)}\right) + (a_k-1)}{\pi_k} = \sum_{k'=1}^{K}\left[\left(\sum_{i=1}^{N} r_{k'}^{(i)}\right) + (a_{k'}-1)\right]$$

→ Therefore, $\boxed{\pi_k \leftarrow \frac{\left(\sum_{i=1}^{N} r_k^{(i)}\right) + (a_k-1)}{\sum_{k'=1}^{K}\left[\left(\sum_{i=1}^{N} r_{k'}^{(i)}\right) + (a_{k'}-1)\right]}}$

• Consider partial derivative w.r.t $\theta_{k,j}$

→ Ignore terms in $(*)$ not influence $\pi_k$, i.e: wants to maximize:

Let $L = \sum_{i=1}^{N} r_k^{(i)} \log P(z^{(i)}=k) + \log p(\vec{\theta})$

$$= \sum_{i=1}^{N} r_k^{(i)} \log \text{Bernoulli}(\theta_{k,j}) + \log p(\vec{\theta})$$

$$\propto \sum_{i=1}^{N} r_k^{(i)} \log\left(\theta_{k,j}^{x_j^{(i)}} (1-\theta_{k,j})^{1-x_j^{(i)}}\right) + \log\left(\theta_{k,j}^{a-1}(1-\theta_{k,j})^{b-1}\right)$$

$$= \sum_{i=1}^{N} r_k^{(i)}\left(x_j^{(i)}\log\theta_{k,j} + (1-x_j^{(i)})\log(1-\theta_{k,j})\right) + (a-1)\log\theta_{k,j} + (b-1)\log(1-\theta_{k,j})$$

→ Then, $\frac{\partial L}{\partial \theta_{k,j}} = \frac{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}}{\theta_{k,j}} - \frac{\sum_{i=1}^{N} r_k^{(i)}(1-x_j^{(i)})}{1-\theta_{k,j}} + \frac{a-1}{\theta_{k,j}} - \frac{b-1}{1-\theta_{k,j}}$

• Let $\frac{\partial L}{\partial \theta_{k,j}} = 0$, we have:

$$\frac{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}}{\theta_{k,j}} - \frac{\sum_{i=1}^{N} r_k^{(i)}(1-x_j^{(i)})}{1-\theta_{k,j}} + \frac{a-1}{\theta_{k,j}} - \frac{b-1}{1-\theta_{k,j}} = 0$$

$$(1-\theta_{k,j})\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} - \theta_{k,j}\sum_{i=1}^{N} r_k^{(i)}(1-x_j^{(i)}) + (1-\theta_{k,j})(a-1) + \theta_{k,j}(b-1) = 0.$$

$$\left(\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}\right) + (a-1) = \theta_{k,j}\left(\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} + \sum_{i=1}^{N} r_k^{(i)}(1-x_j^{(i)}) + (a-1) + (b-1)\right)$$

$$\left(\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}\right) + (a-1) = \theta_{k,j}\left(\sum_{i=1}^{N} r_k^{(i)} + a + b - 2\right)$$

→ Therefore, $\boxed{\theta_{k,j} \leftarrow \frac{\left(\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}\right) + (a-1)}{\left(\sum_{i=1}^{N} r_k^{(i)}\right) + (a+b-2)}}$

2.

```
The theta update seems OK.
The pi update seems OK.
pi[0] 0.08499999999999992
pi[1] 0.12999999999999999
theta[0, 239] 0.6427106227106232
theta[3, 298] 0.46573612495845823
```

PART 2:

1.

$$\Pr(z=k \mid X_{obs}) = \frac{\Pr(z=k)\, P(X_{obs} \mid z=k)}{\sum_{k'=1}^{K} P(z=k')\, P(X_{obs} \mid z^{(i)}=k')}$$

$$= \frac{\Pr(z=k) \prod_{j=1}^{D} P(m_j^{(i)}, x_j^{(i)} \mid z=k)}{\sum_{k'=1}^{K} P(z=k') \prod_{j'=1}^{D} P(m_{j'}^{(i)}, x_{j'}^{(i)} \mid z^{(i)}=k')}, \text{ where } m_j^{(i)} = \begin{cases} 1, & \text{if observed} \\ 0, & \text{otherwise} \end{cases}$$

$$= \frac{\pi_k \prod_{j=1}^{D} \left(\theta_{k,j}^{m_j^{(i)} \cdot x_j^{(i)}}\right)\left((1-\theta_{k,j})^{m_j^{(i)}(1-x_j^{(i)})}\right)}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j'=1}^{D} \left(\theta_{k',j'}^{m_{j'}^{(i)} \cdot x_{j'}^{(i)}}\right)\left((1-\theta_{k',j'})^{m_{j'}^{(i)}(1-x_{j'}^{(i)})}\right)}$$

2.

```
The E-step seems OK.
R[0, 2] 0.17488951492117286
R[1, 0] 0.6885376761092291
P[0, 183] 0.6516151998131036
P[2, 628] 0.4740801724913303
```

## PART 3:

1. • By part 1, we have:

$$\theta_{kj} \leftarrow \frac{(\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}) + (a-1)}{(\sum_{i=1}^{N} r_k^{(i)}) + (a+b-2)}$$

• Substitute $a=b=1$:

$$\theta_{kj} \leftarrow \frac{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}}{\sum_{i=1}^{N} r_k^{(i)}} \qquad (*)$$

• We can observe from $(*)$ that:

→ In training set, if a pixel is always 0, then after training $\theta_{k,j}$, the $\theta$ of this pixel will be 1.

→ Therefore, if the pixel is 1 in the test image, then MAP learning algorithm cannot estimate it well.

2. • By previous clarification, we know, there are 10 different digits classes.

→ There are too small number of components, therefore, it is hard to train a good model to get different writing styles.

• However, in Part 2, we have 100 number of components. In this case, we can train a better model and get higher average log probabilities through learning more written styles and get a better result for image completion.

3. • $\boxed{\text{CLAIM}}$: No, it does not mean that the model thinks 1's are far more common than 8's.

• Since we just use the top half of the image in the model to predict the digit, there may have some mis-classification.

→ Digit 8 is similar to 9 when only consider the top half of the digit, therefore, the probability to predict a correct 8 is lower.

→ However, for digit 1, there is no other digit whose top half is similar to 1, therefore, 1 can get higher average log probability.