Homework 3

Zihan Zhao

1001103708

1 Learning the parameters

1.1

Let $L = \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}[logPr(z^{(i)}=k) + logp(x^{(i)}|z^{(i)}=k)] + logp(\pi) + logp(\Theta)$

First π_k:
Let $\frac{\partial L}{\partial \pi_k} = 0$ to find out π_k.

$$\begin{aligned}
\frac{\partial L}{\partial \pi_k} &= \frac{\partial \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}[logPr(z^{(i)}=k) + logp(x^{(i)}|z^{(i)}=k)] + logp(\pi) + logp(\Theta)}{\partial \pi_k} \\
&= \frac{\partial \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}[logPr(z^{(i)}=k)]}{\partial \pi_k} + \frac{\partial \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}[logp(x^{(i)}|z^{(i)}=k)]}{\partial \pi_k} + \frac{\partial logp(\pi)}{\partial \pi_k} + \frac{\partial logp(\Theta)}{\partial \pi_k} \\
&= \frac{\partial \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}[logPr(z^{(i)}=k)]}{\partial \pi_k} + \frac{\partial 0}{\partial \pi_k} + \frac{\partial logp(\pi)}{\partial \pi_k} + \frac{\partial 0}{\partial \pi_k} \\
&= \frac{\partial \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}[logPr(z^{(i)}=k)]}{\partial \pi_k} + \frac{\partial logp(\pi)}{\partial \pi_k}
\end{aligned}$$

We know $logPr(z^{(i)}=k) = log(\pi_k)$ and
$logp(\pi) = log(A * \prod_{k=1}^{K} \pi_k^{a_k-1}) = log(A) + \sum_{k=1}^{K}(a_k-1)log(\pi_k)$ *where* $log(A)$ *is a constant*
Plug the two equations into $\frac{\partial L}{\partial \pi_k}$:

$$\begin{aligned}
\frac{\partial L}{\partial \pi_k} &= \frac{\partial \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} log(\pi_k)}{\partial \pi_k} + \frac{\partial log(A) + \sum_{k=1}^{K}(a_k-1)log(\pi_k)}{\partial \pi_k} \\
&= \frac{\partial \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} log(\pi_k)}{\partial \pi_k} + \frac{\partial \sum_{k=1}^{K}(a_k-1)log(\pi_k)}{\partial \pi_k} \\
&= \frac{\partial \{\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} log(\pi_k) + \sum_{k=1}^{K}(a_k-1)log(\pi_k)\}}{\partial \pi_k}
\end{aligned}$$

We know $\sum_{k=1}^{K}\pi_k = 1$,
so using Lagrange multipliers, let $L'(\pi_k, \lambda) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} log(\pi_k) + \sum_{k=1}^{K}(a_k-1)log(\pi_k) - \lambda(\sum_{k=1}^{K}\pi_k - 1)$. Then:

$$\begin{aligned}
\frac{\partial L'}{\partial \pi_k} &= \frac{\partial \{\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} log(\pi_k) + \sum_{k=1}^{K}(a_k-1)log(\pi_k) - \lambda(\sum_{k=1}^{K}\pi_k - 1)\}}{\partial \pi_k} \\
&= \frac{\sum_{i=1}^{N} r_k^{(i)}}{\pi_k} + \frac{a_k - 1}{\pi_k} - \lambda
\end{aligned}$$

Set it to zero:

$$\frac{\sum_{i=1}^N r_k^{(i)}}{\pi_k} + \frac{a_k - 1}{\pi_k} - \lambda = 0$$
$$\pi_k = \frac{\sum_{i=1}^N r_k^{(i)} + a_k - 1}{\lambda}$$

plug into equation $\sum_{k=1}^K \pi_k = 1$:

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

plug λ into equation $\pi_k = \frac{\sum_{i=1}^N r_k^{(i)} + a_k - 1}{\lambda}$:

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

Second $\theta_{k,j}$:

$$\frac{\partial L}{\partial \theta_{k,j}} = \frac{\partial \sum_{i=1}^N \sum_{k=1}^K r_k^{(i)}[logPr(z^{(i)} = k) + logp(x^{(i)}|z^{(i)} = k)] + logp(\pi) + logp(\Theta)}{\partial \theta_{k,j}}$$
$$= \frac{\partial \sum_{i=1}^N \sum_{k=1}^K r_k^{(i)}[logp(x^{(i)}|z^{(i)} = k)]}{\partial \theta_{k,j}} + \frac{\partial logp(\Theta)}{\partial \theta_{k,j}}$$

We know $logp(x^{(i)}|z^{(i)} = k) = log\prod_{j=1}^D(\theta_{k,j}^{x_j^{(i)}}(1-\theta_{k,j})^{1-x_j^{(i)}}) = \sum_{j=1}^D(x_j^{(i)}log(\theta_{k,j}) + (1-x_j^{(i)})log(1-\theta_{k,j}))$ and $logp(\Theta) = log\prod_{k=1}^K\prod_{j=1}^D \theta_{k,j}^{a-1}(1-\theta_{k,j})^{b-1} = \sum_{k=1}^K\sum_{j=1}^D\{(a-1)log(\theta_{k,j}) + (b-1)log(1-\theta_{k,j})\}$. Plug them into equation:

$$= \frac{\partial \sum_{i=1}^N \sum_{k=1}^K r_k^{(i)}[\sum_{j=1}^D(x_j^{(i)}log(\theta_{k,j}) + (1 - x_j^{(i)})log(1 - \theta_{k,j}))]}{\partial \theta_{k,j}} +$$
$$\frac{\partial \sum_{k=1}^K \sum_{j=1}^D\{(a-1)log(\theta_{k,j}) + (b-1)log(1-\theta_{k,j})\}}{\partial \theta_{k,j}}$$
$$= \sum_{i=1}^N (\frac{r_k^{(i)} x_j^{(i)}}{\theta_{k,j}} + \frac{r_k^{(i)}(1 - x_j^{(i)})}{\theta_{k,j} - 1}) + \frac{a-1}{\theta_{k,j}} + \frac{b-1}{\theta_{k,j} - 1}$$
$$= \frac{\sum_{i=1}^N (r_k^{(i)}(\theta_{k,j} - x_j^{(i)})) + (a+b-2)\theta_{k,j} + 1 - a}{\theta_{k,j}(1-\theta_{k,j})}$$

set it to zero:

$$\frac{\sum_{i=1}^{N}(r_k^{(i)}(\theta_{k,j} - x_j^{(i)})) + (a+b-2)\theta_{k,j} + 1 - a}{\theta_{k,j}(1-\theta_{k,j})} = 0$$
$$\sum_{i=1}^{N}(r_k^{(i)}(\theta_{k,j} - x_j^{(i)})) + (a+b-2)\theta_{k,j} + 1 - a = 0$$
$$(\sum_{i=1}^{N} r_k^{(i)} + a + b - 2)\theta_{k,j} - \sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} + 1 - a = 0$$

So we got

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

1.2

Look at $mixture.py$
The results of running $mixture.print\_part\_1\_values()$:
$pi[0]$ 0.08499999999999992
$pi[1]$ 0.1299999999999999
$theta[0, 239]$ 0.6427106227106232
$theta[3, 298]$ 0.46573612495845823

2 Posterior inference

2.1

We use Bayes Rule:

$$Pr(z = k|x) = \frac{P(x_{obs}|z = k)P(z = k)}{\sum_{k=1}^{K}[P(x_obs|z = k)P(z = k)]}$$

We know:

$$P(x_{obs}|z = k) = \prod_{i=1}^{N}\prod_{j=1}^{D}(\theta_{k,j}^{x_j^{(i)}}(1-\theta_{k,j})^{1-x_j^{(i)}})^{m_j^{(i)}}$$

Plug it into $Pr(z = k|x)$:

$$Pr(z = k|x) = \frac{[\prod_{i=1}^{N}\prod_{j=1}^{D}(\theta_{k,j}^{x_j^{(i)}}(1-\theta_{k,j})^{1-x_j^{(i)}})^{m_j^{(i)}}]\pi_k}{\sum_{k=1}^{K}[[\prod_{i=1}^{N}\prod_{j=1}^{D}(\theta_{k,j}^{x_j^{(i)}}(1-\theta_{k,j})^{1-x_j^{(i)}})^{m_j^{(i)}}]\pi_k]}$$

Make it more clear

$$Pr(z = k|x) = \frac{\pi_k\prod_{i=1}^{N}\prod_{j=1}^{D}(\theta_{k,j}^{x_j^{(i)}}(1-\theta_{k,j})^{1-x_j^{(i)}})^{m_j^{(i)}}}{\sum_{k=1}^{K}\pi_k\prod_{i=1}^{N}\prod_{j=1}^{D}(\theta_{k,j}^{x_j^{(i)}}(1-\theta_{k,j})^{1-x_j^{(i)}})^{m_j^{(i)}}}$$

2.2

Look at *mixture.py*
The results of running $mixture.print\_part\_2\_values()$:
$R[0, 2]$ 0.17488951492117283
$R[1, 0]$ 0.6885376761092292
$P[0, 183]$ 0.6516151998131037
$P[2, 628]$ 0.4740801724913301

3 Conceptual questions

3.1

Let's assume a pixel is always 0 in training set and it is 1 in test set. If we choose a = b = 1. When updating $\theta_{k,j}$, plug $a = b = 1$ into Part 1 update equation. We got $\theta_{k,j} = \frac{a-1+\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}}{(\sum_{i=1}^{N} r_k^{(i)})+a+b+2} = \frac{1-1+\sum_{i=1}^{N} r_k^{(i)} * 0}{(\sum_{i=1}^{N} r_k^{(i)})+1+1-2} = \frac{0}{(\sum_{i=1}^{N} r_k^{(i)})} = 0$. So posterior prediction will always predict 0 for this pixel with the probability 100%. But the pixel is 1 in test set. So it is always an error.

3.2

Part 1 model only performs M-step and Part 2 model performs complete EM algorithm. Although Part 1 model has labels of data, it still only owns 10 digit class as the number of components. Only M-step can NOT handle the various shapes of the digit figures of the same digit. Part 2 model can handle various shapes for the same digit. For each iteration, after doing M-step, it does E-step to figure out the posterior prediction. It sufficiently maximize log-likelihood through training various shapes/writing styles of the same digit. The purpose of Part 2 model is to predict the completion of the unobserved part through looking for the features of raw data. But Part 1 is to predict from the labels of data. Part 2 log-likelihood is better than Part 1.

3.3

No. We know the ten digits probabilities are soft-related. The 1's average log probability is high since the shapes of many other digits are more similar to 1 than 8. Given a train/test data, its log probability of some digit is higher and will increase probability of 1 if it and 1 are similar. Such the digits who are more like 1 than 8. Therefore, the model gives higher log probability to 1 than 8.