HOMEWORK 3 - CSC 2515

ANEEK DAS

Q1) Given:

The pixels are distributed Bernoulli given the class $z = k$. with parameter $\theta_{k,j}$.

So, $p(x_j^{(i)} | z^{(i)} = k) \sim \text{Bern}(\theta_{k,j})$

where j represents the feature / pixel number.

Now, $p(x^{(i)} | z^{(i)} = k) = \prod_{j=1}^{D} p(x_j^{(i)} | z^{(i)} = k)$

$$\Rightarrow \boxed{p(x^{(i)} | z^{(i)} = k) = \prod_{j=1}^{D} \theta_{k,j}^{x_j^{(i)}} (1 - \theta_{k,j})^{1 - x_j^{(i)}}}$$

Now, the conjugate prior to this is given by the Beta dist.

$$p(\theta_{k,j}) \propto \theta_{k,j}^{a-1} (1 - \theta_{k,j})^{b-1}$$

So, $\boxed{p(\theta) = \prod_{j=1}^{D} \prod_{k=1}^{K} \theta_{k,j}^{a-1} \cdot (1 - \theta_{k,j})^{b-1}}$

The classes (k) are distributed Multinomial (π)

So, $\boxed{p(z^{(i)} = k) = \pi_k}$

And the prior for the classes is given by the Dirichlet dist. (conjugate prior to multinomial)

$$p(\pi) \sim \text{Dir}(a_1, a_2, \ldots, a_k).$$

Now, we will use a symmetric Dirichlet prior (a_k is same)

So, lets represent it by 'α' (alpha)

$$\boxed{p(\pi) = \prod_{k=1}^{K} (\pi_k)^{\alpha - 1}}$$

and $\boxed{\sum_{k=1}^{K} \pi_k = 1}$

1.1. Derive the M step update rules for θ and π.

Now, we have to maximize the objective function

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

where $r_k^{(i)}$ represents responsibility.

subject to the constraint

$$\sum_{k=1}^{K}\pi_k = 1$$

So, we build the Lagrangian as:

objective function $-\ \lambda$ (constraint) ; where 'λ' is the lagrange multiplier.

So,

$$\ell(\text{parameters}) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\left[\log P(z^{(i)}=k) + \log p(x^{(i)} \mid z^{(i)}=k)\right] + \log p(\pi) + \log p(\theta) - \lambda\left[\sum_{k=1}^{K}\pi_k - 1\right]$$

Substituting the values we get:

$$\ell(\text{parameters}) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\log\pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\cdot\log\left[\prod_{j=1}^{D}\theta_{k,j}^{x_j^{(i)}}\cdot(1-\theta_{k,j})^{1-x_j^{(i)}}\right] +$$

$$\log\left[\prod_{k=1}^{K}\pi_k^{\alpha-1}\right] + \log\left[\prod_{j=1}^{D}\prod_{k=1}^{K}\theta_{k,j}^{a-1}\cdot(1-\theta_{k,j})^{b-1}\right] - \lambda\left[\sum_{k=1}^{K}\pi_k - 1\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\log\pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\left[\sum_{j=1}^{D}\log\left\{\theta_{k,j}^{x_j^{(i)}}\cdot(1-\theta_{k,j})^{1-x_j^{(i)}}\right\}\right] +$$

$$\sum_{k=1}^{K}\log\pi_k^{\alpha-1} + \sum_{j=1}^{D}\sum_{k=1}^{K}\log\left\{\theta_{k,j}^{a-1}(1-\theta_{k,j})^{b-1}\right\} - \lambda\left[\sum_{k=1}^{K}\pi_k - 1\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\log\pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\left[\sum_{j=1}^{D} x_j^{(i)}\log\theta_{k,j} + (1-x_j^{(i)})\cdot\log(1-\theta_{k,j})\right] +$$

$$(\alpha-1)\sum_{k=1}^{K}\log\pi_k + \sum_{j=1}^{D}\sum_{k=1}^{K}\left[(a-1)\log\theta_{k,j} + (b-1)\log(1-\theta_{k,j})\right] - \lambda\left[\sum_{k=1}^{K}\pi_k - 1\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{D} r_k^{(i)} x_j^{(i)} \log \theta_{k,j} + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{D} r_k^{(i)} (1-x_j^{(i)}) \log(1-\theta_{k,j})$$

$$+ (\alpha-1)\sum_{k=1}^{K} \log \pi_k + (a-1)\sum_{j=1}^{D}\sum_{k=1}^{K} \log \theta_{k,j} + (b-1)\sum_{j=1}^{D}\sum_{k=1}^{K} \log(1-\theta_{k,j}) - \lambda\left[\sum_{k=1}^{K}\pi_k - 1\right]$$

Now, we take each of the terms and find the partial derivative w.r.t. π_k and $\theta_{k,j}$ [this is for convenience]

<u>term 1</u>

$$t_1 = \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log \pi_k \qquad \frac{\partial t_1}{\partial \theta_{k,j}} = 0 \qquad \frac{\partial t_1}{\partial \pi_k} = \frac{1}{\pi_k}\sum_{i=1}^{N} r_k^{(i)}$$

<u>term 2</u>

$$t_2 = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{D} r_k^{(i)} x_j^{(i)} \log \theta_{k,j}$$

$$\frac{\partial t_2}{\partial \pi_k} = 0$$

$$\frac{\partial t_2}{\partial \theta_{k,j}} = \sum_{i=1}^{N} \frac{r_k^{(i)} x_j^{(i)}}{\theta_{k,j}}$$

<u>term 3</u>

$$t_3 = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{D} r_k^{(i)} (1-x_j^{(i)}).\log(1-\theta_{k,j})$$

$$\frac{\partial t_3}{\partial \pi_k} = 0.$$

$$\frac{\partial t_3}{\partial \theta_{k,j}} = -\frac{1}{1-\theta_{k,j}}\sum_{i=1}^{N} r_k^{(i)}(1-x_j^{(i)}).$$

<u>term 4</u>

$$t_4 = (\alpha-1)\sum_{k=1}^{K} \log \pi_k$$

$$\frac{\partial t_4}{\partial \pi_k} = \frac{\alpha-1}{\pi_k}$$

$$\frac{\partial t_4}{\partial \theta_{k,j}} = 0$$

term 5

$$t_5 = (a-1)\sum_{j=1}^{D}\sum_{k=1}^{K} \log \theta_{k,j}$$

$$\frac{\partial t_5}{\partial \pi_k} = 0.$$

$$\frac{\partial t_5}{\partial \theta_{k,j}} = \frac{(a-1)}{\theta_{k,j}}.$$

term 6

$$t_6 = (b-1)\sum_{j=1}^{D}\sum_{k=1}^{K} \log(1-\theta_{k,j})$$

$$\frac{\partial t_6}{\partial \theta_{k,j}} = -\frac{(b-1)}{1-\theta_{k,j}}$$

term 7

$$t_7 = -\lambda\left[\sum_{k=1}^{K}\pi_k - 1\right]$$

$$\frac{\partial t_7}{\partial \pi_k} = -\lambda \qquad \frac{\partial t_7}{\partial \theta_{k,j}} = 0.$$

Now, that we have found the partial derivatives for the individual terms, we can equate each to 0 to get the maximum value of the parameters.

first we find $\frac{\partial l}{\partial \theta_{k,j}} = \sum_{z=1}^{7} \frac{\partial t_z}{\partial \theta_{k,j}} = 0$

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

$$\Rightarrow \frac{1}{\theta_{k,j}}\left[\sum_{i=1}^{N} v_k^{(i)} x_j^{(i)}\right] + \frac{(a-1)}{\theta_{k,j}} = \frac{1}{1-\theta_{k,j}}\sum_{i=1}^{N} v_k^{(i)}(1-x_j^{(i)}) + \frac{1}{1-\theta_{k,j}}(b-1)$$

$$\Rightarrow \frac{1}{\theta_{k,j}}\left[\left\{\sum_{i=1}^{N} v_k^{(i)} x_j^{(i)}\right\} + (a-1)\right] = \frac{1}{1-\theta_{k,j}}\left[\left\{\sum_{i=1}^{N} v_k^{(i)}(1-x_j^{(i)})\right\} + (b-1)\right]$$

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

Now, we know that if $\frac{x}{1-x} = \frac{a}{b}$. where $\theta_{k,j} = x$

'a' represents numerator of expression above

'b' represents denominator of expression above

$$\Rightarrow bx = a - ax$$

$$\Rightarrow (a+b)x = a.$$

$$\Rightarrow x = \frac{a}{a+b}.$$

So, we get.

$$\boxed{\theta_{k,j} = \frac{(a-1) + \sum_{i=1}^{N} v_k^{(i)} x_j^{(i)}}{\left[(a-1) + \sum_{i=1}^{N} v_k^{(i)} x_j^{(i)}\right] + \left[(b-1) + \sum_{i=1}^{N} v_k^{(i)} \left(1-x_j^{(i)}\right)\right]}}$$

Now, we need to find update for π_k and our lagrange multiplier λ.

$$\frac{\partial l}{\partial \pi_k} = \sum_{z=1}^{7} \frac{\partial t_z}{\partial \pi_k} = 0.$$

So,

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

$$\Rightarrow \frac{1}{\pi_k}\left[(\alpha-1) + \sum_{i=1}^{N} v_k^{(i)}\right] = \lambda$$

$$\Rightarrow \frac{1}{\lambda}\left[(\alpha-1) + \sum_{i=1}^{N} v_k^{(i)}\right] = \pi_k.$$

Now, as previously stated $\sum_{k=1}^{K} \pi_k = 1$, So, suming both sides over all values of k.

$$\frac{1}{\lambda} \sum_{k=1}^{K}\left[\alpha - 1 + \sum_{i=1}^{N} v_k^{(i)}\right] = 1$$

$$\Rightarrow \boxed{\lambda = K(\alpha - 1) + \sum_{k=1}^{K}\sum_{i=1}^{N} v_k^{(i)}}$$

Now, putting the value of λ, we get.

$$\boxed{\pi_k = \frac{(\alpha - 1) + \sum_{i=1}^{N} v_k^{(i)}}{K(\alpha-1) + \sum_{k=1}^{K}\sum_{i=1}^{N} v_k^{(i)}}}$$

.2. Please find code in mixture.py.

Output for mixture.print_part_1_values().

pi[0] 0.085

pi[1] 0.13

theta[0,239] 0.6427106227106227

theta[3,298] 0.4657361249584458

2.1 Calculate the posterior Prob. dist. $p(z|x)$

we know, from the previous answer,

$$p(x^{(i)} | z^{(i)} = k) = \prod_{j=1}^{D} \theta_{k,j}^{x_j^{(i)}} (1 - \theta_{k,j})^{1 - x_j^{(i)}}$$

and, $p(z^{(i)} = k) = \pi_k$.

From Bayes Rule we know,

$$p(z^{(i)} = k | x^{(i)}) = \frac{p(x^{(i)} | z^{(i)} = k) \cdot p(z^{(i)} = k)}{p(x^{(i)})} \quad \text{——} (1)$$

Now, $p(x)$ can be expressed as the sum of probabilities of x occuring for each of the classes k. So,

$$p(x^{(i)}) = \sum_{k=1}^{K} p(x^{(i)} \mid z^{(i)} = k) \cdot p(z^{(i)} = k)$$

$$= \sum_{k=1}^{K} \left[\prod_{j=1}^{D} \theta_{k,j}^{x_j^{(i)}} \cdot (1 - \theta_{k,j})^{1 - x_j^{(i)}} \right] \cdot \pi_k .$$

we will be removing superscript '(i)' to account for all x. Now, substituting the values in our Bayes Rule equation, we get:

$$p(z = k \mid x) = \frac{p(x \mid z = k) \cdot p(z = k)}{p(x)}$$

$$\boxed{p(z = k \mid x) = \frac{\left[\prod_{j=1}^{D} \theta_{k,j}^{x_j} \cdot (1 - \theta_{k,j})^{1 - x_j} \right] \cdot \pi_k}{\sum_{k=1}^{K} \left[\left\{ \prod_{j=1}^{D} \theta_{k,j}^{x_j} \cdot (1 - \theta_{k,j})^{1 - x_j} \right\} \cdot \pi_k \right]}}$$

2.2. Please find code in mixture.py

2.3. Please find code in mixture.py

Output for running mixture.print_part_2_values()

R[0,2] 2.028044728466474 e-27

R[1,0] 1.1176809102154152 e-39

P[0,183] 0.4389191263694054

P[2,628] 0.42264903589063 24

3.1. Conceptual Questions.

we know that the update value of $\theta_{k,j}$ is

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

Now, if $a=b=2$ we get

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

Now, if $x_j^{(i)}$ is 0 for all samples in the training set, then

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

So, it still leaves some prob. that $x_j^{(i)}$ would be 1 in the test set.

But if $a=b=1$. and $x_j^{(i)}$ is 0 for all samples in the training set, then

$\boxed{\theta_{k,j} = 0}$ [putting $a=b=1$ & $x_j^{(i)}=0$ in eqn. (1)]

So, now,

$$p(x^{(i)} \mid z^{(i)}=k) = \prod_{j=1}^{D} (\theta_{k,j})^{x_j^{(i)}} \cdot (1-\theta_{k,j})^{1-x_j^{(i)}}$$

As, $\theta_{k,j} = 0$

$$\boxed{p(x^{(i)} \mid z^{(i)} = k) = 0}$$

So, the prob. of $x^{(i)}=1$ for all classes 'k' is 0. So, it would assign 0 prob to images in test set.

3.2. A model can learn from 2 sources: 1. the data and the corresponding labels,
2. The latent parameters and their prior probabilities. In case of the
model part 1, it has access to the labels and can learn from that.
Even though the model from Part 2 partially observes the data, it uses
the knowledge of $\theta_{k,j}$ & π_k to make predictions. So, model Part 2
can still get higher avg. log probabilities.

3.3. A higher log probability corresponds to a more confident prediction.
So, a higher log probability of 1's means that the model is more
confident in predicting 1's than 8's. So, the model <u>will not</u> generate
more 1's if sampled.