Brendan Neal 1001160236|nealbre1 2018-11-21

CSC411 Homework 6

Part 1: Learning the Parameters

Question 1: M-Step Update Rules

Note: assume symmetric Dirichlet as per instructions, with A $= a_1 = a_2 = \cdots = a_k$

Final Answers:

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

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

Derivation:

Objective function for reference

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

Note that $\sum \pi_k = 1$, so Lagrange multipliers must be used for maximization. Maximizing under this constraint is equal to maximizing:

$$\left(\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\left(\log P\left(z^{(i)} = k\right) + \log p\left(x^{(i)} | z^{(i)} = k\right)\right)\right) + \log p(\pi) + \log p(\Theta) - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

Definitions of probabilities

$$P\left(z^{(i)} = k\right) = \pi_k$$

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

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

$$p(\Theta) = \prod_{j=1}^{D}\prod_{k=1}^{K} p\left(\theta_{k,j}\right) = \prod_{j=1}^{D}\prod_{k=1}^{K} \left(\theta_{k,j}\right)^{a-1}\left(1 - \theta_{k,j}\right)^{b-1}$$

Applying log to each probability

$$\log P\left(z^{(i)} = k\right) = \log \pi_k$$

$$\log p\left(x^{(i)} \middle| z^{(i)} = k\right) = \sum_{j=1}^{D} x_j^{(i)} \log \theta_{k,j} + \sum_{j=1}^{D} \left(1 - x_j^{(i)}\right) \log\left(1 - \theta_{k,j}\right)$$

$$\log p(\pi) = (A - 1) \sum_{k=1}^{K} \log \pi_k$$

$$\log p(\Theta) = (a - 1) \sum_{j=1}^{D} \sum_{k=1}^{K} \log \theta_{k,j} + (b - 1) \sum_{j=1}^{D} \sum_{k=1}^{K} \log(1 - \theta_{k,j})$$

Rearranging objective function:

$$\sum_{i=1}^{N} \sum_{k=1}^{K} r_k^{(i)} \log P\left(z^{(i)} = k\right) + \sum_{i=1}^{N} \sum_{k=1}^{K} r_k^{(i)} \log p\left(x^{(i)} | z^{(i)} = k\right) + \log p(\pi) + \log p(\Theta) + \lambda \left(1 - \sum_{k=1}^{K} \pi_k\right)$$

Subbing in log probabilities

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

Simplifying multiplication inside summation

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

Deriving with respect to $\pi_k, \theta_{k,j}, \lambda$

Term	**Derivative WRT:**		
	$\boldsymbol{\lambda}$	$\boldsymbol{\pi_k}$	$\boldsymbol{\theta_{k,j}}$
$\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log \pi_k$	0	$\frac{1}{\pi_k}\sum_{i=1}^{N} r_k^{(i)}$	0
$\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{D} r_k^{(i)} x_j^{(i)} \log \theta_{k,j}$	0	0	$\sum_{i=1}^{N} \frac{r_k^{(i)} x_j^{(i)}}{\theta_{k,j}}$
$\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{D} r_k^{(i)}\left(1 - x_j^{(i)}\right) \log\left(1 - \theta_{k,j}\right)$	0	0	$-\sum_{i=1}^{N} \frac{r_k^{(i)}\left(1 - x_j^{(i)}\right)}{1 - \theta_{k,j}}$
$(A-1)\sum_{k=1}^{K} \log \pi_k$	0	$\frac{A-1}{\pi_k}$	0
$(a-1)\sum_{j=1}^{D}\sum_{k=1}^{K} \log \theta_{k,j}$	0	0	$\frac{a-1}{\theta_{k,j}}$
$(b-1)\sum_{j=1}^{D}\sum_{k=1}^{K} \log(1 - \theta_{k,j})$	0	0	$-\frac{(b-1)}{\left(1 - \theta_{k,j}\right)}$
$\lambda\left(1 - \sum_{k=1}^{K} \pi_k\right)$	$1 - \sum_{k=1}^{K} \pi_k$	$-\lambda$	0

Derivative WRT:	
π_k	$\frac{1}{\pi_k}\sum_{i=1}^{N} r_k^{(i)} + \frac{A-1}{\pi_k} - \lambda$
$\theta_{k,j}$	$\sum_{i=1}^{N} \frac{r_k^{(i)} x_j^{(i)}}{\theta_{k,j}} - \sum_{i=1}^{N} \frac{r_k^{(i)}\left(1 - x_j^{(i)}\right)}{1 - \theta_{k,j}} + \frac{(a-1)}{\theta_{k,j}} - \frac{(b-1)}{\left(1 - \theta_{k,j}\right)}$
λ	$1 - \sum_{k=1}^{K} \pi_k$

Maximizing π_k

$$\left[\frac{1}{\pi_k}\sum_{i=1}^{N} r_k^{(i)} + \frac{A-1}{\pi_k} - \lambda = 0\right] \Rightarrow \left[\frac{1}{\pi_k}\left(\sum_{i=1}^{N} r_k^{(i)} + A - 1\right) = \lambda\right] \Rightarrow \left[\pi_k = \frac{1}{\lambda}\left(\sum_{i=1}^{N} r_k^{(i)} + A - 1\right)\right](1)$$

$$\left[1 - \sum_{k=1}^{K} \pi_k = 0\right] \Rightarrow \left[\sum_{k=1}^{K} \pi_k = 1\right](2)$$

$$[(1) \wedge (2)] \Rightarrow \left[\frac{1}{\lambda}\sum_{k=1}^{K}\left(\sum_{i=1}^{N} r_k^{(i)} + A - 1\right) = 1\right] \Rightarrow \left[\lambda = \sum_{k=1}^{K}\left(\sum_{i=1}^{N} r_k^{(i)} + A - 1\right)\right](3)$$

$$[(3) \wedge (1)] \Rightarrow \left[\pi_k = \frac{\sum_{i=1}^{N} r_k^{(i)} + A - 1}{\sum_{k=1}^{K}\left(\sum_{i=1}^{N} r_k^{(i)} + A - 1\right)}\right] \Rightarrow \left[\pi_k = \frac{\left(\sum_{i=1}^{N} r_k^{(i)}\right) + A - 1}{\left(\sum_{k=1}^{K}\sum_{i=1}^{N} r_k^{(i)}\right) + KA - K}\right]$$

Maximizing $\theta_{k,j}$

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

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

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

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

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

Lemma 1:

$$\left[\frac{X}{1-X}=\frac{A}{B}\right]\Leftrightarrow\left[X=\frac{A}{A+B}\right]$$

Proof:

$$\left[\frac{X}{1-X}=\frac{A}{B}\right]\Leftrightarrow\left[X=\frac{A}{B}(1-X)\right]\Leftrightarrow\left[X=\frac{A}{B}-\frac{A}{B}X\right]$$

$$\Leftrightarrow\left[X+\frac{A}{B}X=\frac{A}{B}\right]\Leftrightarrow\left[X\left(1+\frac{A}{B}\right)=\frac{A}{B}\right]\Leftrightarrow\left[X\left(\frac{A+B}{B}\right)=\frac{A}{B}\right]$$

$$\Leftrightarrow\left[X=\frac{A}{B}*\frac{B}{A+B}\right]\Leftrightarrow\left[X=\frac{A}{A+B}\right]$$

Question 2: Output of Part 1

```
pi[0] 0.085

pi[1] 0.13

theta[0, 239] 0.642710622711

theta[3, 298] 0.465736124958
```

Part 2: Posterior Inference

Question 1: Derivation of Posterior Distribution

Final Answer:

$$p(z=k|x)=\frac{\pi_k \times \prod_{j=1}^{D}\left(\theta_{k,j}\right)^{x_j}\left(1-\theta_{k,j}\right)^{(1-x_j)}}{\sum_{i=1}^{K}\pi_i\prod_{j=1}^{D}\left(\theta_{i,j}\right)^{x_j}\left(1-\theta_{i,j}\right)^{(1-x_j)}}$$

Derivation:

Bayes Rule

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

Definitions of the above terms

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

$$p(z=k)=\pi_k$$

Can use marginal probability to compute $p(x)$

$$p(x)=\sum_{k=1}^{K}p(x,z=k)=\sum_{k=1}^{K}p(z=k)p(x|z=k)$$

$$\Rightarrow p(x)=\sum_{k=1}^{K}\pi_k\prod_{j=1}^{D}\left(\theta_{k,j}\right)^{x_j}\left(1-\theta_{k,j}\right)^{(1-x_j)}$$

Subbing in terms to original equation:

$$p(z=k|x)=\frac{\pi_k\prod_{j=1}^{D}\left(\theta_{k,j}\right)^{x_j}\left(1-\theta_{k,j}\right)^{(1-x_j)}}{\sum_{i=1}^{K}\pi_i\prod_{j=1}^{D}\left(\theta_{i,j}\right)^{x_j}\left(1-\theta_{i,j}\right)^{(1-x_j)}}$$

Question 3: Part 2 Output

```
R[0, 2] 2.02804472847e-27

R[1, 0] 1.11768091022e-39

P[0, 183] 0.438919126369

P[2, 628] 0.422649035891
```

Part 3: Conceptual Questions

Question 1

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

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

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

If pixel $x_j^{(i)} = 0$ for every training sample i in class k, then $\theta_{k,j} = 0$.

Therefore, the probability of observing $x_j^{(i)}$ will always be 0 – (see definition of $p(x^{(i)}|z = k)$, equation 2 – the product will cancel into 0)

Question 2

This makes sense because the model has access to learned latent variables (θ, π), which it can use to compensate for the missing information.

Question 3

This does NOT mean that the model believes there are more 1's than 8's.

The higher the log probability, the more confident the model is in its prediction.

If the log probability for 1 is higher than the log probability of 8, this means that the model is more confident about judging an image to be 1 than it would be to judge it as an 8.

Intuitively this makes sense as 1's are often near-vertical lines, whereas 8's have features that are present in other classes:

3 looks like right half of an 8

5's bottom looks like bottom of 8

6's bottom looks like bottom of 8

9's top looks like top of 8