CSC 411 Fall 2018

Machine Learning and Data Mining

Homework 6

Family name: Bhatia Given name: Pooja

Solution Part 1

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

$P(\theta_{k,j}) \propto \theta_{k,j}^{\alpha-1} (1-\theta_{k,j})^{b-1}$ — (2)

$$\theta'_k = \arg\max_{\theta_k} \left\{ \sum_i \{ r_k^{(i)} \log P(x^{(i)} | z^{(i)} = k) + \log P(\Theta) \right\}$$

$$\frac{\partial(l(\theta'_k))}{\partial \theta_{k,j}} = \frac{\partial \sum_i \left\{ r_k^{(i)} \log \prod_{j'} P(x_{j'}^{(i)} \mid \theta'_{kj'}) \right\} + \log \prod_{j'} P(\theta'_{kj'})}{\partial \theta'_{k,j}}$$

$$= \left\{ \frac{\partial \sum_{j \neq j'} (\cdot)}{\partial \theta'_{k,j}} = 0 \right\}$$

$$= \frac{\partial \sum_i \left\{ r_k^{(i)} \log \left(\theta'^{\,x_j^{(i)}}_{k,j} (1-\theta'_{k,j})^{(1-x_j^{(i)})} \right) \right\} + \log \theta'^{\,a-1}_{kj} (1-\theta'_{kj})^{b-1}}{\partial \theta'_{kj}}$$

(from eq 2)

#1 (a)

$$= \frac{\partial \sum_i r_k^{(i)} \left(x_j^{(i)} \log \theta'_{kj} + (1 - x_j^{(i)}) \log (1 - \theta'_{kj}) \right)}{\partial \theta'_k}$$

$$+ \frac{\partial (a-1) \log \theta'_{kj} + (B-1) \log (1 - \theta'_{kj})}{\partial \theta'_k}$$

$$= \frac{\left(\sum_i r_k^{(i)} x_j^{(i)} \right) + a - 1}{\theta'_{kj}} - \frac{\left(\sum_i r_k^{(i)} (1 - x_j^{(i)}) \right) + B - 1}{1 - \theta'_{kj}}$$

$$= \frac{(1 - \theta'_{kj}) \left(\sum_i r_k^{(i)} x_j^{(i)} \right) - \theta'_{kj} \left(\sum_i r_k^{(i)} (1 - x_j^{(i)}) \right)}{\theta'_{kj} (1 - \theta'_{kj})}$$

$$+ \frac{(1 - \theta'_{kj})(a-1) - \theta'_{kj}(B-1)}{\theta'_{kj} (1 - \theta'_{kj})}$$

$$= \frac{\left(\sum_i r_K^{(i)} x_j^{(i)}\right) - \theta'_{Kj}\left(\sum_i r_K^{(i)}\right) + (a-1) - \theta'_{Kj}(a+B-2)}{\theta'_{Kj}(1-\theta'_{Kj})}$$

Equating the above equation to zero

$$0 = \frac{\left(\sum_i r_K^{(i)} x_j^{(i)}\right) - \theta'_{Kj}\left(\sum_i r_K^{(i)}\right) + (a-1) - \theta'_{Kj}(a+B-2)}{\theta'_{Kj}(1-\theta'_{Kj})}$$

$$\theta'_{Kj}\left[\left(\sum_i r_K^{(i)}\right) + a + B - 2\right] = \left(\sum_i r_K^{(i)} x_j^{(i)}\right) + a - 1$$

$$\boxed{\theta'_{Kj} = \frac{\left(\sum_i r_K^{(i)} x_j^{(i)}\right) + a - 1}{\left(\sum_i r_K^{(i)}\right) + a - B - 2}}$$

Ans

For updating M-step for π Basically we need to maxmize

$$\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log Pr(Z^{(i)}=k) = \sum_{i=1}^{N}\sum_{i=1}^{K} r_K^{(i)} \log \pi_K$$

$$P(\pi) \propto \pi_1^{a-1} \cdot \cdot \cdot \cdot \cdot \cdot \pi_K^{aK-1} \hookrightarrow ①$$

but we are using symmetric Drichlet prior where all a_K are equal

$$h = \sum_{i=1}^{N}\sum_{k=1}^{K} r_K^{(i)} \log \pi_K + \lambda\left(1 - \sum_{K=1}^{K} \pi_K\right)$$

By applying Lagrangian

$$\frac{\partial h}{\partial \pi_K} = \frac{(a-1)\sum_{i=1}^{N} r_K^{(i)}}{\pi_K} - \lambda \qquad \text{(from ①)}$$

setting derivative to zero

$$\lambda = \frac{\left(\sum_{i=1}^{N} r_K^{(i)}\right) a-1}{\pi_K} \quad \text{for each } K$$

For this to be true π_K must be propotional to $\sum_{i=1}^{N} r_K^{(i)}$

$$\boxed{\pi_K = \frac{(a-1)\sum_{i=1}^{N} r_K^{(i)}}{\sum_{K'=1}^{K}\sum_{i=1}^{N} r_{K'}^{(i)}}}$$

Q(2)

output of mixture.print_part_1_values()

pi[0] 0.3761467889908 2565

pi[1] 0.6238532110091743

Solution Part 3

Q2

The model from Part 2 gets significantly higher average log probabilities on both the training and test sets compared to model from Part 1 because the mixture component has a nice form. ~~and~~ In the model from part 2 since some of the observables are only observed i.e X_{obs}, we are making predictions about rest of the observables. (i.e predictions about missing data given the observed data). The same can be achived by posterior predictive distribution, i.e. $p(z|X_{obs})$. Hence Prosterior inference is working better.

Q3

For Both training and test set, images of 1's are assigned far higher log-probability than images of 8's. Since log probability of 1's in higher, when we sample from this distribution we are likely to set more 1's than

8's. Hence Yes sample from distribution will generate far more 1's than 8's.