

Part 1

Question 1

2019年03月23日 C SC411 HW6

1. $\lambda = \sum_{i=1}^N \sum_{k=1}^K r_k^i [\log \Pr(z^i = k) + \log p(x^i | z^i = k)] + \log p(\pi) + \log p(\theta)$

$z \sim \text{Multinomial}(\pi)$

$x_i | z = k \sim \text{Bernoulli}(\theta_{k,j})$

$\theta_{k,j} \sim \text{Beta}(a, b)$ e.g. $p(\theta_{k,j}) \propto \theta_{k,j}^{a-1} (1-\theta_{k,j})^{b-1}$ prior over θ

$\pi \sim \text{Dirichlet}(c, \dots, c)$ e.g. $p(\pi) = \frac{c!}{\pi_k^{c-1}} \prod_k \pi_k^{c-1}$ prior over class

How to derive λ above?

$\log P(\theta, \pi | D) = \log P(D | \theta) P(\theta) = \log P(D | \theta) + \log P(\theta) + \log P(\pi)$

$P(D | \theta) = P(x^1, y^1) \dots P(x^N, y^N)$ assuming we knew labels

$= P(x^1 | y^1) P(y^1) \dots P(x^N | y^N) P(y^N)$

$\log P(D | \theta) = \sum_{i=1}^N [\log P(x^i | y^i) + \log P(y^i)] + \log P(\theta) + \log P(\pi)$

$= \sum_{i=1}^N \sum_{k=1}^K S(y^i = k) [\log P(x^i | y^i = k) + \log P(y^i = k)] + \log P(\theta) + \log P(\pi)$

So plug in $r_k^i = P(y^i = k | x^i)$ for $S(y^i = k)$

$r_k^i = P(y^i = k | x^i) = \frac{P(x^i | y^i = k) P(y^i = k)}{P(x^i)} = \frac{\prod_{j=1}^D \theta_{k,j}^{x_j^i} (1-\theta_{k,j})^{1-x_j^i} \cdot \pi_k}{\sum_{p=1}^K \prod_{j=1}^D \theta_{p,j}^{x_j^i} (1-\theta_{p,j})^{1-x_j^i} \pi_p}$

r_k^i is the responsibility of the k^{th} cluster for the i^{th} data point

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

r_k^i & π here are constants

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

Fact: $\sum_{k=1}^K r_k^i = 1$

Maximize λ subject to $\sum_{k=1}^K \pi_k = 1$ $\sum_{k=1}^K \pi_k - 1 = 0 \equiv g$

$\frac{d\lambda}{d\pi_k} = \sum_{i=1}^N \sum_{k=1}^K r_k^i \frac{1}{\pi_k} + \frac{c_k}{\pi_k} = \lambda + 1 = \lambda \frac{dg}{d\pi_k}$

$\frac{1}{\pi_k} \left(\sum_{i=1}^N r_k^i + c_k \right) = \lambda$

$\pi_k = \frac{c_k + \sum_{i=1}^N r_k^i}{\lambda}$

$\sum_{k=1}^K \frac{c_k + \sum_{i=1}^N r_k^i}{\lambda} = 1 \Rightarrow \lambda = \frac{\sum_{k=1}^K (c_k + \sum_{i=1}^N r_k^i)}{1}$

$\lambda = KC + \sum_{k=1}^K \sum_{i=1}^N r_k^i = KC + \sum_{i=1}^N \sum_{k=1}^K r_k^i = KC + \sum_{i=1}^N 1 = KC + N$

$\therefore \pi_k = \frac{(c-1) + \sum_{i=1}^N r_k^i}{K(c-1) + N}$

Note: as expected, $\sum_{k=1}^K \pi_k = 1$

$$\frac{dL}{d\Theta_{p,q}} = \sum_{i=1}^N r_p^i \left(\frac{x_q^i}{\Theta_{p,q}} + \frac{x_q^i - 1}{1 - \Theta_{p,q}} \right) + \frac{a-1}{\Theta_{p,q}} - \frac{b-1}{1 - \Theta_{p,q}}$$

$$0 = \frac{a-1 + \sum_{i=1}^N r_p^i x_q^i}{\Theta_{p,q}} + \frac{b-1 + \sum_{i=1}^N r_p^i (x_q^i - 1)}{1 - \Theta_{p,q}}$$

$$0 = (1 - \Theta_{p,q})(a-1 + f) + \Theta_{p,q}(1-b+g), \quad f = \sum_{i=1}^N r_p^i x_q^i, \quad g = \sum_{i=1}^N r_p^i (x_q^i - 1)$$

$$0 = (a+f-1) + (1-f-a)\Theta_{p,q} + (1+g-b)\Theta_{p,q}$$

$$0 = (a+f-1) + (1-f-a+b+g+1)\Theta_{p,q}$$

$$0 = \frac{1-a-f}{-b+g-f-a+2} = \Theta_{p,q}$$

$$0 = -b+g-f-a+2$$

$$0 = -b+g-f-a+2$$

$$0 = -b+g-f-a+2$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{1-a - \sum_{i=1}^N r_p^i x_q^i}{-(b+a) + \sum_{i=1}^N r_p^i + 2}$$

$$\Theta_{p,q} = \frac{\sum_{i=1}^N (r_p^i x_q^i) + (a-1)}{\left(\sum_{i=1}^N r_p^i \right) + (a+b)-2}$$

class
feature

How to vectorize?

$N \times D$

$$R = \begin{bmatrix} r^1 & r^2 & \dots & r^N \\ 1 & 1 & \dots & 1 \end{bmatrix} \quad X = \begin{bmatrix} x^1 & x^2 & \dots & x^D \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

Say $R = [r^1 r^2 \dots r^N]$ — each col. is responsibilities of k^{th} class of all training examples.

$$X = [x^1 x^2 \dots x^D]$$

— same idea.

$$R^T X = \begin{bmatrix} r^1 x^1 & r^1 x^2 & \dots & r^1 x^D \\ \vdots & \vdots & \ddots & \vdots \\ r^N x^1 & r^N x^2 & \dots & r^N x^D \\ k^1 x^1 & k^1 x^2 & \dots & k^1 x^D \\ \vdots & \vdots & \ddots & \vdots \\ k^N x^1 & k^N x^2 & \dots & k^N x^D \end{bmatrix}$$

Then divide each element of a row by $\sum_{i=1}^N r_p^i$.

Question 2

Part 1 values:

('pi[0]', 0.084999999999999923)

('pi[1]', 0.129999999999999987)

('theta[0, 239]', 0.64271062271062329)

('theta[3, 298]', 0.46573612495845823)

Part 2

Question 1

Part 2

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

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

↳ assumes fully observed though.

$$Pr(Z=K | X_{obs}) = \sum_{X_{hidden}} Pr(Z=K, X_{hidden} | X_{obs})$$

$$= \sum_{X_{hidden}} Pr(Z=K | X_{hidden}, X_{obs}) Pr(X_{hidden} | X_{obs})$$

↳ By independence.

$$= \sum_{X_{hidden}} Pr(Z=K | X_{hidden}, X_{obs}) Pr(X_{hidden})$$

$$= \sum_{X_{hidden}} \frac{Pr(X_{hidden}, X_{obs} | Z=K) Pr(Z=K)}{Pr(X_{hidden}, X_{obs})} \cdot Pr(X_{hidden})$$

↳ independence

$$= \sum_{X_{hidden}} \frac{Pr(X_{hidden} | Z=K) Pr(X_{obs} | Z=K) Pr(Z=K)}{Pr(X_{hidden}) Pr(X_{obs})} Pr(X_{hidden})$$

$$= \frac{Pr(X_{obs} | Z=K) Pr(Z=K)}{Pr(X_{obs})} \sum_{X_{hidden}} \frac{1}{Pr(X_{hidden} | Z=K)}$$

$$Pr(Z=K | X_{obs}) = \frac{Pr(X_{obs} | Z=K) \cdot Pr(Z=K)}{Pr(X_{obs})}$$

— e.g. we only multiply $\theta_{k,j}$ for x_j observed.

Question 2

Posterior Predictive Mean

Given Posterior: $p(z=k | x_{obs}, \hat{\theta})$. Define $x_u := x_{unknown}$

$$p(x_u | x_{obs}, \hat{\theta}) = \sum_{k=1}^K p(x_u, z=k | x_{obs}, \hat{\theta})$$

$$= \sum_{k=1}^K p(x_u | x_{obs}, z=k, \hat{\theta}) p(z=k | x_{obs}, \hat{\theta})$$

↓

By our model, $z=k$ totally gives distribution

$$= \sum_{k=1}^K p(x_u | z=k, \hat{\theta}) p(z=k | x_{obs}, \hat{\theta})$$

$$E[x_u] = 0 \cdot p(x_u=0 | x_{obs}, \hat{\theta}) + 1 \cdot p(x_u=1 | x_{obs}, \hat{\theta})$$

→ abuse of notation a bit. we are considering a specific pixel.

here, say feature d

$$\therefore E[x_u] = \sum_{k=1}^K p(x_u=1 | z=k, \hat{\theta}) p(z=k | x_{obs}, \hat{\theta})$$

$$= \sum_{k=1}^K \hat{\theta}_{k,d} \cdot p(z=k | x_{obs}, \hat{\theta})$$

e.g. Average the θ of the classes weighted by the ~~post~~ posterior

Vectorization:

$$\Theta = \begin{bmatrix} -\theta_0- \\ \vdots \\ -\theta_K- \end{bmatrix} \quad \text{posterior} = \begin{bmatrix} P(y^1=0) & \dots & P(y^1=K) \\ \vdots & & \vdots \\ P(y^N=0) & \dots & P(y^N=K) \end{bmatrix}$$

→ Need weighted avg of rows of this

$$\Theta^T \begin{bmatrix} P(y^1=0) \\ \vdots \\ P(y^N=K) \end{bmatrix} = E[x_u \text{ for } x^1] = \begin{bmatrix} 1 \\ \theta_0 \\ 1 \end{bmatrix} P(y^1=0) + \dots + \begin{bmatrix} 1 \\ \theta_K \\ 1 \end{bmatrix} P(y^1=K)$$

Question 3

Part 2 values:

('R[0, 2]', 0.1748895149211743)

('R[1, 0]', 0.68853767610923089)

('P[0, 183]', 0.74454557952318989)

('P[2, 628]', 0.45918083205837279)

Part 3

Output of average log probability by digit class for model trained by labels:

Training set

Average log-probability of a 0 image: -201.136

Average log-probability of a 1 image: -97.755

Average log-probability of a 2 image: -201.701

Average log-probability of a 3 image: -184.880

Average log-probability of a 4 image: -172.018

Average log-probability of a 5 image: -191.928

Average log-probability of a 6 image: -177.253

Average log-probability of a 7 image: -156.895

Average log-probability of a 8 image: -187.626

Average log-probability of a 9 image: -162.041

Test set

Average log-probability of a 0 image: -199.743

Average log-probability of a 1 image: -96.973

Average log-probability of a 2 image: -199.290

Average log-probability of a 3 image: -182.329

Average log-probability of a 4 image: -171.179

Average log-probability of a 5 image: -190.813

Average log-probability of a 6 image: -181.537

Average log-probability of a 7 image: -154.226

Average log-probability of a 8 image: -187.226

Average log-probability of a 9 image: -159.568

Part 3

1. Recall: $\theta_{p,q} \leftarrow \frac{(\sum_{i=1}^N r_p^i x_q^i) + a - 1}{(\sum_{i=1}^N r_p^i) + a + b - 2}$ if $x_q^i = 0 \forall i, \theta_{p,q} \leftarrow 0$.

If $a=b=1$, $\theta_{p,q} \leftarrow \frac{\sum_{i=1}^N r_p^i x_q^i}{\sum_{i=1}^N r_p^i}$

By our indep. assumption:

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

if $x_q^i = 1$, then it is clear $\forall k, P(x|z=k) = 0$.

Conceptually, based on the data, algo is 100% sure that for all classes, that pixel must be off. So it predicts 0 prob if by chance it is on.

2. The part 1 model only has 10 clusters for each label so it averages and misses out on variations in writing. This will be detrimental to similar looking numbers, e.g. 7, 7, 9, 8, 9

In contrast, part 2 has 100 clusters. It can form clusters to differentiate 7, 7, 9, 8, 9

3. The average log probability for each digit class is a report of how confidently the model predicts the correct class for examples in the class by averaging over all the examples in a class.

It is clear then that the model does NOT believe 1's are more common than 8's. It means it is easier to differentiate 1's than 8's. This makes sense because 7, 9, 3 for example look like 8.

The relative frequency of sampling a class is actually given by π . Thus, from the info given, it is unknown if one would sample more 1s or 8s.

Extra

Train_from_labels output:

('Training log-likelihood:', -172.07854196938328)

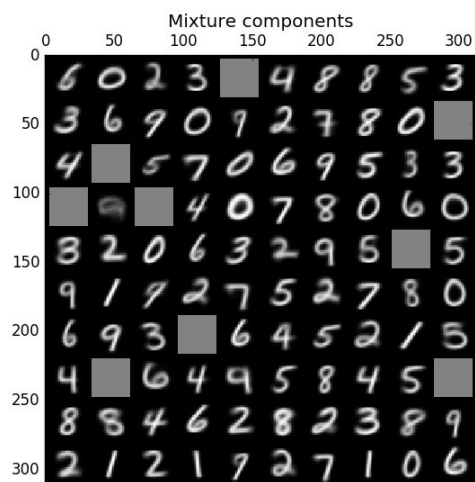
('Test log-likelihood:', -170.97538133747267)

Train_with_em output:

('Final training log-likelihood:', -138.14790916090735)

('Final test log-likelihood:', -138.52928834184047)

EM Mixtures:



Model Predictions (top half was observed):

