# Bayesian Model Machine Learning Midterm exam

Ben Zhu
bjz2107

1) $x_1 \ldots x_n$ with each $x \in \{0,1,2,3 \ldots\}$

$x_i \overset{iid}{\sim} p(x|\pi,r)$

$\pi \sim Beta(a,b)$

$$p(x|\pi,r) = \binom{x+r-1}{x} \pi^x (1-\pi)^r$$

a) Posterior distribution of $\pi$, $p(\pi|X)$

ok I recognize this as a binomial ... ok some sort of strange binomial.

$$p(\pi|x,r) = \frac{p(x|\pi,r)\, p(\pi)}{\int p(x|\pi,r)\, p(\pi)}$$

$$p(X|\pi,r) = \prod_{i=1}^{n} \binom{x_i+r-1}{x_i} \pi^{x_i} (1-\pi)^r$$

$$p(\pi|X,r) = \left( \prod_{i=1}^{n} \binom{x_i+r-1}{x_i} \pi^{x_i} (1-\pi)^r \right) \left( \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1} \right)$$

$$\propto \left[ \pi^x (1-\pi)^r \right] \left[ \pi^{a-1} (1-\pi)^{b-1} \right]$$

$$\propto \left[ \pi^{x+a-1} (1-\pi)^{r+b-1} \right]$$

which will be proportional to a $B(a', b')$

where $a' = a + \sum_{i=1}^{N} x_i$ $\quad b' = b + r$ ← where this may or not be a constant

B) predictive distribution

$$p(\hat{x}|x) = \int p(\hat{x}|\theta)\, p(\theta|x)\, d\theta$$

for our problem

$$p(x_{n+1}|x_{old}) = \int p(x_{n+1}|\pi, r)\, p(\pi|x, r)\, d\pi$$

$\pi | x, r \to$ Beta

$x | \pi, r \to$ Binomial

this is a $p(x|\pi r)$     this is what we found in part I

$$= \int \left[\binom{x_{n+1}+r-1}{x_{n+1}} \pi^{x_{n+1}}(1-\pi)^r\right]\left[\frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')}\pi^{a'-1}(1-\pi)^{b'-1}\right] d\pi$$

where $a' = a + \sum_{i=1}^{N} x_i \quad b' = b + r$

$$= \binom{x_{n+1}+r-1}{x_{n+1}}\left(\frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')}\right)\int \pi^{x_{n+1}}(1-\pi)^r\, \pi^{a'-1}(1-\pi)^{b'-1}\, d\pi$$

$$= \binom{x_{n+1}+r-1}{x_{n+1}}\left(\frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')}\right)\int \pi^{x_{n+1}+a'-1}(1-\pi)^{r+b'-1}\, d\pi$$

and we know this is Beta in form, so no need to "actually" integrate

$$= \binom{x_{n+1}+r-1}{x_n}\left(\frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')}\right)\left(\frac{\Gamma(x_n+a')\,\Gamma(r+b')}{\Gamma(x_{n+1}+a'+r+b')}\right)$$

this is the

$p(x_{n+1}|\text{old } X)$

2) EM algorithm

Bayesian linear regression

Data set D $\{(x_i, y_i)\}_{i=1}^N$ $x \in R^d$ $y \in R$

The Model

$y_i \overset{iid}{\sim} Normal(x_i^T w, \alpha^{-1})$ $w \sim Normal(0, \lambda^{-1} I)$ $\lambda \sim Gamma(a, b)$

want $p(y, \lambda | x)$

$p(w|x,y) = \frac{p(y|w,x)\,p(w)}{p(y|w)} = Normal(\mu, \Sigma)$

but not really relevant

Integrate out $w$ to learn about $\lambda$

previously $p(y, w | x) = \int p(y, w, \alpha | x)\, d\alpha$

hard direction is start with desired marginal distribution and introduce a variable that gives marginal distribution

new easy direction, start with expanded model and try to maximize other original

$\boxed{\text{point estimate of } \lambda}$

once you have

$$p(y, \lambda | x) = \int p(y, w, \lambda | x)\, dw$$

let's assume $\alpha \sim Gamma(c, d)$

$$p(y, \lambda | x) = \int p(y, w, \lambda, \alpha | x)\, d\alpha\, dw$$

$$\ln p(y, \lambda | x) = \iint q(w, \alpha) \ln \frac{p(y, \lambda, w, \alpha | x)}{q(w, \alpha)}\, d\alpha\, d\lambda + \iint KL$$

will set $q(\alpha, w) = p(\alpha, w | y, \lambda, x)$

Bayes

$$p(\alpha, w | y, \lambda, x) \propto p(y | \alpha, w, \lambda)\, p(\alpha)\, p(w)$$

$$p(\alpha | y, w, \lambda, x) = \prod^N p(y | w, \lambda, \alpha, x)\, p(\alpha)$$

Gamma

$$p(w | y, \alpha, \lambda, x) = p(y | w, \lambda, \alpha, x)\, p(w)$$

Normal

$$\mathcal{L}(\lambda) = E_q[\ln p(y, w, \alpha, \lambda | x) - \ln q(\alpha, w)]$$

expected value will be a Normal and a Gamma so will be constants

$$\mathcal{L}_t(\lambda) = E_q[\ln p(y, w, \alpha, \lambda | x)]$$

$$\mathcal{L}_t(\lambda) = E_q[\cancel{\ln p(y | w, \alpha, \lambda, x)\, p(\alpha)\, p(w)\, p(\lambda)}]$$

E step continued ->

E step

$p(\alpha, w \mid \lambda, y, x) = p(y, w \mid \lambda, x, \alpha)\, p(\alpha) = p(y \mid \lambda, w, x, \alpha)\, p(\alpha)\, p(\lambda)$

$p(\alpha \mid y, w, x)$ $\quad p(w \mid \lambda) = \text{Normal}$

$p(\alpha, w \mid \lambda, y, x) = p(\alpha \mid y, \lambda, x)\, p(w \mid \lambda)$

Gamma( ) Normal $(\mu, \Sigma)$

$\text{Gamma}\left(c + \frac{N}{2},\ d + \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i^T w)^2\right)$

$\mu = \Sigma\left(\sum_{i=1} \alpha y_i x_i\right)$

$\Sigma = \left(\lambda I + \alpha \sum_{i=1}^{N} x_i x_i^T\right)^{-1}$

$\mathcal{L}_t(\lambda) = \mathbb{E}_q[\ln p(y, w, \lambda, \alpha \mid x)]$

$= \mathbb{E}_q[\ln p(y, w \mid \lambda, x, \alpha)\, p(\alpha)\, p(w)]$

$= \mathbb{E}_{q(\alpha, w)}\ \left(\lambda I \cdot \alpha \sum_{i=1}^{N} x_i x_i^T\right)^{-1} \left(\sum \alpha y_i x_i\right)$

I don't know how to actually go through and fully take the Expectation

but following from the notes,

$p(\alpha)$

$\mathcal{L}_t(\lambda) = \left(\frac{\mathbb{E}_q[\alpha]}{2} \sum_{i=1}^{N}(x_i x_i^T) + \lambda I\right)\left(\sum_{i=1}^{N} \mathbb{E}_q[\alpha]\, y_i x_i\right)$

$= \mathbb{E}_q \ln\left[(2\pi\alpha^{-1})^{-\frac{1}{2}} \exp\left[\frac{1}{2\alpha}\sum(y_i - x_i^T w)^2\right]\right] + \mathbb{E}_q \ln\left[2\pi^{-\frac{d}{2}} \exp\left[\frac{\lambda}{2I}(w^T w)\right]\right]$

$+ \mathbb{E}_q \ln\left[\alpha^{a-1} e^{-b\alpha}\right]$

constants are whatever

$= -\frac{\mathbb{E}_q[\alpha]}{2}\sum_{i=1}^{N}(y_i - x_i^T w)^2 + \mathbb{E}_q\left[\frac{\lambda}{2I}(w^T w)\right] + \mathbb{E}_q[-b\alpha]$

M step, take $\nabla_\lambda = 0$ solve for a $\lambda$

$0 = \lambda \mathbb{E}_q[w^T w]\frac{1}{2I} + \mathbb{E}_q[\alpha](-b)$

$$\boxed{\lambda_t = \frac{-b\,\mathbb{E}_q[\alpha]}{\mathbb{E}_q[w^T w]}}$$

pseudocode

1) Initialize $\lambda$

2) for every iteration, calculate

a) $\mathbb{E}_q[\alpha]$ $\quad \mathbb{E}_q[w]$, which is the expectation of a gamma, and a Normal that have been calculated previously →

E step

$\mathbb{E}_q[w] = \Sigma\left(\sum_{i=1}^{N} \mathbb{E}_q[\alpha]\, y_i x_i\right)$

$\mathbb{E}_q[\alpha] = \frac{c}{d}$

b) M step Calculate $\lambda_t = \frac{-b\,\mathbb{E}_q[\alpha]}{\mathbb{E}_q[w^T w]}$

3) calculate $\ln p_t(y, \lambda \mid w)$ from $\ln p(y, \lambda, w, \alpha \mid x)$ until they have a $\Delta$ sufficiently small

3) Variational inference $\quad D\{x_i, y_i\}_{i=1}^{N}$

$y_i \overset{iid}{\sim} N(x_i^T w, \alpha^{-1}) \quad w \sim \text{Norm}(0, \lambda^{-1} I) \quad \alpha \sim \text{Gamma}(a,b) \quad \lambda \sim \text{Gamma}(e,f)$

want to learn $q(w,\alpha,\lambda) = p(w,\alpha,\lambda \mid y, x)$

optimal $q(\lambda)$ $\quad$ using $q(w,\alpha,\lambda) = q(w)\, q(\alpha)\, q(\lambda)$

$$q(\alpha) \propto \exp\{E_{q(w,\lambda)}[\ln p(y,\alpha,w,\lambda \mid x)]\}$$

not relevant to $\lambda$

$$\propto \exp\{E_{q(w,\lambda)}[\ln p(y \mid \alpha, w, \lambda, x) + \ln p(\alpha) + \ln p(w) + \ln p(\lambda)]\}$$

$$\propto \exp\{E_{q(w,\lambda)}[\ln p(y \mid \alpha, w, \lambda, x)] + E_{q(w,\lambda)} \ln p(\alpha)\}$$

$$\propto \exp\{E_{q(w,\lambda)}[\ln p(y \mid x, \alpha, w, \lambda)] + \ln p(\alpha)\}$$

$$\propto \exp\{E_q[\sum_{i=1}^{N} \tfrac{1}{2}\ln\alpha - \tfrac{1}{2}\ln 2\pi - \tfrac{\alpha}{2}(y_i - x_i^T w)^2] + (a-1)\ln\alpha - b\alpha\}$$

$$\propto \left(\frac{\alpha}{2\pi}\right)^{\frac{N}{2}} \exp\left\{\frac{-\alpha}{2} E_{q(w,\lambda)}[y_i^2 - 2y_i x_i^T w + w^T x_i x_i^T w)] + (a-1)\ln\alpha - b\alpha\right\}$$

$$\propto \alpha^{\frac{N}{2}+a-1} \exp\left\{-\frac{\alpha}{2}\left(\sum_{i=1}^{N} y_i^2 - 2y_i x_i^T E_q[w] + E_q[\text{trace}(x_i x_i^T w w^T)]\right) - b\alpha\right\}$$

note $\quad$ actually this expansion doesn't really matter

this is clearly a gamma

$$\propto \alpha^{a+\frac{N}{2}-1} e^{[-\alpha[b + \frac{1}{2}\sum_{i=1}^{N} E_{q(w,\lambda)}[(y_i - x^T w)^2]]]}$$

$q(\alpha) \sim \text{Gamma}(a', b')$ where $a' = a + \frac{N}{2}$

$b' = b + \frac{1}{2}\sum_{i=1}^{N} E_{q(w,\lambda)}[(y_i - x^T w)^2]$

optimal

$$q(\lambda) \propto \exp\{E_{q(w,\alpha)}[\cancel{\ln p(y \mid \alpha, w, x)} + \ln p(\lambda) + \ln p(w \mid \lambda) + \cancel{\ln p(\alpha)}]\}$$

not actually dependent on $\lambda$... $\quad$ not relevant

$$\propto \exp\{E_{q(w,\alpha)}[\ln p(w \mid \lambda)] + \ln p(\lambda)\}$$

$$\propto \underbrace{\lambda^{e-1} e^{-f\lambda}}_{p(\lambda)} \cdot \exp\{E_{q(w,\alpha)}[\sum_{i=1}^{N} \tfrac{1}{2}\ln\tfrac{\lambda}{2\pi} - \tfrac{\lambda}{2}(w^T w)]\}$$

$$\propto \lambda^{e-1} e^{-f\lambda} \lambda^{\frac{1}{2}} \cdot e^{-\lambda E_{q(w,\alpha)}[w^T w]} = \lambda^{e+\frac{1}{2}} e^{-\lambda[f + \frac{1}{2}E_q[w^T w]]}$$

Definitely a Gamma

$q(\lambda) \sim \text{Gamma}(e', f')$ where $e' = e + \frac{1}{2}$

$f' = f + \frac{1}{2}E_{q(w,\alpha)}[w^T w]$

optimal

$$q(w) \propto \exp\left\{ E_{q(\lambda,\alpha)} \left[ \ln p(y|w,\alpha,x) + \ln p(\lambda) + \ln p(w|\lambda) + \ln p(\alpha) \right) \right.$$

doesn't depend on $w$     doesn't depend

$$\propto \exp\left\{ \mathbb{E}_{q(\alpha,\lambda)} \left[ \sum_{i=1}^{N} \frac{1}{2} \ln \frac{\alpha}{2\pi} - \frac{\alpha}{2}(y_i - x_i^T w)^2 \right] \right] - \ln p(w|\lambda) \Big\}$$

$$\propto \exp\left\{ E_{q(\alpha,\lambda)} \left[ \sum_{i=1}^{N} \cancel{\frac{1}{2}\ln\alpha} - \frac{\alpha}{2}(y_i - x_i^T w)^2 \right] - \left[ \cancel{\frac{1}{2}\ln\frac{\lambda}{2\pi}} - \frac{\lambda}{2}(w^T w) \right] \right\}$$

$$\propto \exp\left\{ E_{q(\alpha,\lambda)} \sum \left[ -\frac{\alpha}{2}(y_i - x_i^T w)^2 \right] - \frac{\lambda}{2}(w^T w) \right\}$$

$$\propto e^{\left( -\frac{E_{q(\alpha,\lambda)}[\alpha]}{2} \sum_{i=1}^{N} (y_i - x_i^T w)^2 - \frac{\lambda}{2} w^T w \right)}$$

This is definitely a normal

$$q(w) = N(w \mid \mu^*, \Sigma^*) \quad \text{where} \quad \Sigma^* = \left( E_{q(\alpha,\lambda)}[\lambda] I + E_{q(\alpha,\lambda)}[\alpha] \sum_{i=1}^{N} (x_i x_i^T) \right)^{-1}$$

$$\mu^* = \Sigma^* \left( E_{q(\alpha,\lambda)}[\alpha] \sum_{i=1}^{N} y_i x_i \right)$$

psuedo code

VI for Bayesian linear regression

You have your setup Data $D$, $\{(x_i, y_i)\}_{i=1}^{N}$ and $q(\lambda) \sim \text{Gamma}(e, f)$

$q(w) \sim N(w \mid \mu, \Sigma)$

$q(\alpha) \sim \text{Gamma}(\alpha \mid a, b)$

1) initialize $a$ and $b$ to d dimensions, and $e, f, \mu, \Sigma$.

2) for every iteration,

update $q(\lambda)$ with $e_{t+1} = e_t + \frac{1}{2}$, $f_{t+1} = f_t + \frac{1}{2}[\mu_t \mu_t^T]$

$q(\alpha)$ with $a_{t+1} = a_t + \frac{N}{2}$, $b_{t+1} = b_t + \frac{1}{2} \sum_{i=1}^{N} (y_i - x_i^T \mu_t)^2$

$q(w)$ with $\Sigma_{t+1} = \left[ \left( \frac{e_{t+1}}{f_{t+1}} \right) I + \left( \frac{a_{t+1}}{b_{t+1}} \right) \sum_{i=1}^{N} (x_i x_i^T) \right]^{-1}$ and

$$\mu_{t+1} = \Sigma_{t+1} \left[ \left( \frac{a_{t+1}}{b_{t+1}} \right) \sum_{i=1}^{N} y_i x_i \right]$$