MA060129, Bayesian ML Homework 2: Theoretical Problems

Nikita.Balabin@skoltech.ru

November 7, 2022

Problem 1 [24 pts]

We observe the iid dataset $\{(x_n, y_n)\}_{n=1}^{N}, y_n \in \mathbb{R}, x_n \in \mathbb{R}^D, \forall n$. We would like to solve the regression task with the mixture of linear regression models. We consider K components and assign each observation (x_n, y_n) to one of K components using a binary random variable z_{nk}. Each component is described as following:

$$
\begin{aligned}
&p(z_{nk} = 1) = \pi_k, \\
&p(y_n, w \mid x_n, z_{nk} = 1) = \mathcal{N}\left(y_n \mid x_n^T w, \beta_k\right) \mathcal{N}\left(w \mid \mu_k, \alpha I\right), \ w \in \mathbb{R}^D.
\end{aligned}
$$

Consider the model with latent variables $w, \{z_{nk}\}_{n,k=1}^{N,K}$ and parameters $\alpha, \{\pi_k, \beta_k, \mu_k,\}_{k=1}^{K}$ and derive EM updates.

(a) (12 points) Derive Expectation step of the EM-algorithm for the given model: find posteriors and expectations which is necessary for M-step.

(b) (12 points) Derive Maximization step of the EM-algorithm for the given model: take the gradients of the M-step objective and find new values of the parameters.

Note: You are required to demonstrate all the relevant calculations.

Recap:

$$
\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)
$$

Solution

Let θ denotes adaptive parameters $\alpha, \{\pi_k, \beta_k, \mu_k,\}_{k=1}^{K}$.

(a)

$$
p(w, z \mid y, x, \theta) = p(w \mid x, y, z, \theta)\, p(z \mid \theta)
$$

Since w_k independent

$$
p(w \mid x, y, z, \theta) = \prod_{k=1}^{K} p(w_k \mid x, y, z, \theta)
$$

w_k depends only on x_n, y_n where $z_{nk} = 1$ (denote it as X_k, Y_k) $\Rightarrow$

$$
p(w_k \mid x, y, z, \theta) = p(w_k \mid X_k, Y_k, \theta) - \textit{usual bayesian linear regression} \Rightarrow
$$

posterior $w_k = \mathcal{N}\left(w_k \mid \hat{\mu_k}, \hat{\Sigma_k}\right)$ is given by:

$$
\hat{\mu_k} = \hat{\Sigma_k}\left(\alpha^{-1}\mu_k + \beta_k^{-1} X_k^T Y_k\right), \ \hat{\Sigma_k}^{-1} = \alpha^{-1} I + \beta_k^{-1} X_k^T X_k
$$

$$
p(z \mid \theta) = \prod_{n,k: z_{nk}=1} \pi_k
$$

$$
\mathsf{E}_q \log p(y \mid x, w, z, \theta) = \mathsf{E}_z \mathsf{E}_w \log p(y \mid x, w, z, \theta)
$$

I don't know how to proceed analytically :,)

Problem 2 [26 pts]

We would like to build a small recommendation system for the video service of Russian comedy films. We observe a matrix $M \in \mathbb{R}^{D\times d}$, where $M_{ij} \in \mathbb{R}$ is the rating from user i for the movie j. We would like to decompose it on the latent vectors of the user preferences $U_i \in \mathbb{R}^k$ and movie description $V_j \in \mathbb{R}^k$:

$$p\left(M_{ij} \mid U, V\right) = \mathcal{N}\left(M_{ij} \mid U_i^T V_j, \tau^2\right).$$

Also, for these comedy films we will use normal distribution prior with zero mean, since we are unbiased critics:

$$p\left(U_{ik}\right) = \mathcal{N}\left(U_{ik} \mid 0, \sigma_k^2\right)$$
$$p\left(V_{jk}\right) = \mathcal{N}\left(V_{jk} \mid 0, \beta_k^2\right)$$

We would like to find the mean-field approximation of the posterior $p(U, V \mid M) \approx q(U)q(V)$

(a) (12 points) Derive mean-field update for $q(U)$.

(b) (12 points) Derive mean-field update for $q(V)$.

(c) (2 points) Provide full iterative process.

Note: You are required to demonstrate all the relevant calculations.

Solution

$$\begin{aligned}
\mathbf{ln}\ p\left(M_{ij}, U, V\right) = \mathbf{ln}\ p\left(M_{ij} \mid U_i, V_j\right) + \mathbf{ln}\ p\left(U_i \mid 0, \sigma_i^2 I\right) + \mathbf{ln}\ p\left(V_j \mid 0, \beta_j^2 I\right) = \\
= \mathbf{ln}\ \frac{1}{\tau\sqrt{2\pi}} - \frac{1}{2\tau^2}\left(M_{ij} - U_i^T V_j\right)^2 + \\
\frac{k}{2}\mathbf{ln}\ \frac{1}{\sigma_i^2 2\pi} - \frac{1}{2\sigma_i^2} U_i^T U_i + \\
+\frac{k}{2}\mathbf{ln}\ \frac{1}{\beta_j^2 2\pi} - \frac{1}{2\beta_j^2} V_j^T V_j
\end{aligned}$$

(a)

$$\begin{aligned}
\mathbf{ln}\ q\left(U_i\right) &= \mathsf{E}_{V_j}\left[-\frac{1}{2\tau^2}\left(M_{ij} - U_i^T V_j\right)^2 - \frac{1}{2\sigma_i^2} U_i^T U_i\right] + const1 = \\
&= \mathsf{E}_{V_j}\left[\frac{Mij}{2\tau^2}\left(U_i^T V_j + V_j^T U_i\right) - \frac{1}{2\tau^2} U_i^T V_j V_j^T U_i - \frac{1}{2\sigma_i^2} U_i^T U_i\right] + const2 = \\
&= \mathsf{E}_{V_j}\left[\frac{Mij}{2\tau^2}\left(U_i^T V_j + V_j^T U_i\right) - \frac{1}{2} U_i^T \left(\frac{1}{\tau^2} V_j V_j^T + \frac{1}{\sigma_i^2} I\right) U_i\right] + const2 = \\
&= \frac{1}{2}\mathsf{E}_{V_j}\left[-\left(U_i - \frac{Mij}{\tau^2}\Sigma V_j\right)^T \Sigma^{-1}\left(U_i - \frac{Mij}{\tau^2}\Sigma V_j\right)\right] + const4,
\end{aligned}$$

where $\Sigma = \left(\frac{1}{\tau^2} V_j V_j^T + \frac{1}{\sigma_i^2} I\right)^{-1}$. Hence,

$$q\left(U_i\right) = \mathcal{N}\left(U_i \mid \frac{Mij}{\tau^2}\mathsf{E}_{V_j}\left[\Sigma V_j\right], \mathsf{E}_{V_j}\left[\Sigma\right]\right)$$

(b)

$$\begin{aligned}
\mathbf{ln}\ q\left(V_j\right) &= \mathsf{E}_{U_i}\left[-\frac{1}{2\tau^2}\left(M_{ij} - U_i^T V_j\right)^2 - \frac{1}{2\beta_j^2} V_j^T V_j\right] + const1 = \\
&= \mathsf{E}_{U_i}\left[-\frac{1}{2\tau^2}\left(M_{ij} - V_j^T U_i\right)^2 - \frac{1}{2\beta_j^2} V_j^T V_j\right] + const1
\end{aligned}$$

Similarly (a):

$$q\left(V_j\right) = \mathcal{N}\left(V_j \mid \frac{Mij}{\tau^2}\mathsf{E}_{U_i}\left[\Sigma U_i\right], \mathsf{E}_{U_i}\left[\Sigma\right]\right),$$

where $\Sigma = \left(\frac{1}{\tau^2} U_i U_i^T + \frac{1}{\beta_j^2} I\right)^{-1}$.

(c)
   (a) Initialize mean and covariance matrix of any (e.g. $q(U_1)$) vector to some arbitrary value.
   (b) Using U_1 calculate means and covariance matrices of $q(V_j)\,,\ j = 1..d$.
   (c) Using V_1 calculate means and covariance matrices of $q(U_i)\,,\ i = 1..D$.
   (d) Repeat the last two steps until neither of the values change much.