第八次作业

Ex 8.1

Ex. 8.1 Let $r(y)$ and $q(y)$ be probability density functions. Jensen's inequality states that for a random variable X and a convex function $\phi(x)$, $\mathrm{E}[\phi(X)] \geq \phi[\mathrm{E}(X)]$. Use Jensen's inequality to show that

$$\mathrm{E}_q \log[r(Y)/q(Y)] \tag{8.61}$$

is maximized as a function of $r(y)$ when $r(y) = q(y)$. Hence show that $R(\theta, \theta) \geq R(\theta', \theta)$ as stated below equation (8.46).

Notice that $\log(x)$ is a concave function, thus $-\log(x)$ is convex. so using Jensen's inequality, we have that :

$$\begin{aligned} E_q[-\log[r(Y)/q(Y)]] &\geq -\log[E_q[r(Y)/q(Y)]] \\ &= -\log\left[\int \frac{r(y)}{q(y)} q(y)dy\right] \\ &= -\log\left[\int r(y)dy\right] \\ &= -\log(1) \\ &= 0 \end{aligned}$$

So, the inequality sign takes equality if and only if $r(Y) = q(Y)$.

For equation (8.46), we have:

$$\begin{aligned} R(\theta', \theta) - R(\theta, \theta) &= E[\ell_1(\theta; \mathbf{Z}^m|\mathbf{Z})|\mathbf{Z}, \theta] - E[\ell(\theta; \mathbf{Z}^m|\mathbf{Z})|\mathbf{Z}, \theta] \\ &= E_{\Pr(\mathbf{Z}^m|\mathbf{Z},\theta)}\left(\log \frac{\Pr(\mathbf{Z}^m|\mathbf{Z}, \theta')}{\Pr(\mathbf{Z}^m|\mathbf{Z}, \theta)}\right) \\ &\leq 0. \end{aligned}$$

请推导基展开模型的后验估计的均值与方差（equation (8.27)）

The second ingredient we need is a prior distribution. Distributions on functions are fairly complex entities: one approach is to use a Gaussian process prior in which we specify the prior covariance between any two function values $\mu(x)$ and $\mu(x')$ (Wahba, 1990; Neal, 1996).

Here we take a simpler route: by considering a finite B-spline basis for $\mu(x)$, we can instead provide a prior for the coefficients β, and this implicitly defines a prior for $\mu(x)$. We choose a Gaussian prior centered at zero

$$\beta \sim N(0, \tau \mathbf{\Sigma}) \tag{8.25}$$

with the choices of the prior correlation matrix $\mathbf{\Sigma}$ and variance τ to be discussed below. The implicit process prior for $\mu(x)$ is hence Gaussian, with covariance kernel

$$\begin{aligned} K(x, x') &= \operatorname{cov}[\mu(x), \mu(x')] \\ &= \tau \cdot h(x)^T \mathbf{\Sigma} h(x'). \end{aligned} \tag{8.26}$$

The posterior distribution for β is also Gaussian, with mean and covariance

$$\begin{aligned} \mathrm{E}(\beta|\mathbf{Z}) &= \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\mathbf{\Sigma}^{-1}\right)^{-1}\mathbf{H}^T\mathbf{y}, \\ \operatorname{cov}(\beta|\mathbf{Z}) &= \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\mathbf{\Sigma}^{-1}\right)^{-1}\sigma^2, \end{aligned} \tag{8.27}$$

Consider that:

$$\Pr[\beta|\mathbf{Z}] = \Pr[\beta|y,x] \approx \Pr[\beta]\Pr[y|x,\beta] = \mathcal{N}(0,\tau\Sigma)\mathcal{N}_y(\mu_\beta(x),\sigma_\epsilon^2) = \mathcal{N}(0,\tau\Sigma)\mathcal{N}_y((H^\top H)^{-1}Hy,(H^\top H)^{-1}\sigma^2)$$

We then use the following conclusion directly

进入正题。假设两个独立随机变量$\mathrm{x} \sim \mathrm{N}(\mu_\mathrm{x}, \sigma_\mathrm{x}^2)$，$\mathrm{y} \sim \mathrm{N}(\mu_\mathrm{y}, \sigma_\mathrm{y}^2)$，则它们的乘积符合高斯概率密度函数的形式：

$$(\mathrm{x}, \mathrm{y}) \sim \mathrm{N}\left(\frac{\mu_\mathrm{y}\sigma_\mathrm{x}^2 + \mu_\mathrm{x}\sigma_\mathrm{y}^2}{\sigma_\mathrm{x}^2 + \sigma_\mathrm{y}^2}, \frac{1}{1/\sigma_\mathrm{x}^2 + 1/\sigma_\mathrm{y}^2}\right)$$

具体的推导方式，可以通过$\mathrm{p(x)p(y)}$乘积获得：

$$\mathrm{p(x)p(y)} = \frac{1}{2\pi^2\sigma_\mathrm{x}\sigma_\mathrm{y}}\exp\left(-\frac{\sigma_\mathrm{y}^2(\mathrm{x}-\mu_\mathrm{x})^2 + \sigma_\mathrm{x}^2(\mathrm{x}-\mu_\mathrm{y})^2}{2\sigma_\mathrm{x}^2\sigma_\mathrm{y}^2}\right)$$

and we get:

$$
\begin{aligned}
\mathrm{E}(\beta|\mathbf{Z}) &= \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\mathbf{\Sigma}^{-1}\right)^{-1}\mathbf{H}^T\mathbf{y}, \\
\mathrm{cov}(\beta|\mathbf{Z}) &= \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\mathbf{\Sigma}^{-1}\right)^{-1}\sigma^2,
\end{aligned}
$$

如何将二分量混合高斯模型拓展到三个分量的混合高斯模型，请给出实现算法。

Assuming that there's a hiden parameter $\tau = [\tau, \tau_2, \tau_3]$ and $\tau_1 + \tau_2 + \tau_3 = 1$

Likelihood Function

The aim is to estimate the unknown parameters representing the mixing value between the Gaussians and the means and covariances of each:

$$
\theta = \left(\boldsymbol{\tau}, \boldsymbol{\mu}1, \boldsymbol{\mu}2, \Sigma_1, \Sigma_2\right),
$$

where the incomplete-data likelihood function is

$$
L(\theta; \mathbf{x}) = \prod_{i=1}^{n}\sum_{j=1}^{3}\tau_j\, f(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j)
$$

E-Step

This E step corresponds with setting up this function for Q:

$$
\begin{aligned}
Q(\theta \mid \theta^{(t)}) &= \mathrm{E}_{\mathbf{Z}|\mathbf{X}=\mathbf{x};\theta^{(t)}}\left[\log L(\theta; \mathbf{x}, \mathbf{Z})\right] \\
&= \mathbb{E}_{\mathbf{Z}|\mathbf{X}=\mathbf{x};\theta^{(t)}}\left[\log \prod_{i=1}^{n} L(\theta; \mathbf{x}_i, Z_i)\right] \\
&= \mathrm{E}_{\mathbf{Z}|\mathbf{X}=\mathbf{x};\theta^{(t)}}[\sum_{i=1}^{n}\log L(\theta; \mathbf{x}_i, Z_i)] \\
&= \sum_{i=1}^{n}\mathbb{E}_{Z_i|X_i=x_i;\theta^{(t)}}\left[\log L(\theta; \mathbf{x}_i, Z_i)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{3} P(Z_i = j \mid X_i = \mathbf{x}_i; \theta^{(t)}) \log L(\theta_j; \mathbf{x}_i, j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{3} T_{j,i}^{(t)}\left[\log \tau_j - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_j) - \frac{d}{2}\log(2\pi)\right].
\end{aligned}
$$

M-Step

To begin, consider $\boldsymbol{\tau}$, which has the constraint $\tau_1 + \tau_2 = 1$:

$$
\begin{aligned}
\boldsymbol{\tau}^{(t+1)} &= \underset{\boldsymbol{\tau}}{\arg\max}\, Q(\theta \mid \theta^{(t)}) \\
&= \underset{\boldsymbol{\tau}}{\arg\max} \left\{ \left[\sum_{i=1}^{n} T_{1,i}^{(t)}\right] \log \tau_1 + \left[\sum_{i=1}^{n} T_{2,i}^{(t)}\right] \log \tau_2 + \left[\sum_{i=1}^{n} T_{3,i}^{(t)}\right] \log \tau_3 \right\}.
\end{aligned}
$$

And for gaussian model 1($f(\mathbf{x}_1; \boldsymbol{\mu}_1, \Sigma_1) \sim \mathcal{N}(\mu_1, \Sigma_1)$), we do that.

$$
\boldsymbol{\mu}_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} T_{1,i}^{(t)}}
$$

$$
\Sigma_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^\top}{\sum_{i=1}^{n} T_{1,i}^{(t)}}
$$

And for the other two models, we use the same method to compute the new parameters.

Reference

https://en.wikipedia.org/wiki/Expectation–maximization_algorithm