IML 第六次作业

习题 7.4

欲防止 $p=\prod_{i=1}^{d} P(x_i \mid c)$ 计算时出现下溢，可以在计算时取对数，即计算

$$\log p=\sum_{i=1}^{d} \log P(x_i \mid c)$$

此时又可能发生上溢，可以在计算每一个 $\log P(x_i \mid c)$ 时，先除以 d，随后再相加。这样在实践中可减少溢出的出现，即使出现了，也容易定位溢出步骤。

习题 7.5

最小化分类错误率的贝叶斯最优分类器为

$$h^*(x)=\underset{c\in\mathcal{Y}}{\arg\max} P(c \mid x)=\underset{c\in\mathcal{Y}}{\arg\max} P(x \mid c)P(c)$$

数据满足高斯分布时，有

$$\begin{aligned} h^*(x) &= \underset{c\in\mathcal{Y}}{\arg\max}\left\{\log\left[\frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu_c)^\top \Sigma^{-1}(x-\mu_c)}\right]+\log P(c)\right\} \\ &= \underset{c\in\mathcal{Y}}{\arg\max}\left[-\frac{1}{2}(x-\mu_c)^\top \Sigma^{-1}(x-\mu_c)+\log P(c)\right] \\ &= \underset{c\in\mathcal{Y}}{\arg\max}\left[x^\top \Sigma^{-1}\mu_c-\frac{1}{2}\mu_c^\top \Sigma^{-1}\mu_c+\log P(c)\right] \end{aligned}$$

记

$$f(c)=x^\top \Sigma^{-1}\mu_c-\frac{1}{2}\mu_c^\top \Sigma^{-1}\mu_c+\log P(c)$$

则在二分类任务中，贝叶斯决策边界为

$$g(x)=f(1)-f(0)=x^\top \Sigma^{-1}(\mu_1-\mu_0)-\frac{1}{2}(\mu_1+\mu_0)^\top \Sigma^{-1}(\mu_1-\mu_0)+\log\frac{P(1)}{P(0)}$$

对于线性判别分析，参考书 3.39 式，可得投影界面为

$$w=(\Sigma_0+\Sigma_1)^{-1}(\mu_1-\mu_0)$$

当两类方差相同时，可化为

$$w=\frac{1}{2}\Sigma^{-1}(\mu_1-\mu_0)$$

两类在投影面连线的中点为

$$\frac{1}{2}(\mu_1+\mu_0)^\top w=\frac{1}{4}(\mu_1+\mu_0)^\top \Sigma^{-1}(\mu_1-\mu_0)$$

则线性判别分析的决策边界为

$$\tilde{g}(x)=x^\top \Sigma^{-1}(\mu_1-\mu_0)-\frac{1}{2}(\mu_1+\mu_0)^\top \Sigma^{-1}(\mu_1-\mu_0)$$

因为假设同先验，所以 $\log\frac{P(1)}{P(0)}=0$，所以 $g(x)=\tilde{g}(x)$，证毕。

作业 7.3

欲证明收敛，即证明 EM 算法每次迭代得到的 Θ^t 满足

$$P(X \mid \Theta^{t+1}) \geq P(X \mid \Theta^t)$$

由于 $\ln P(X \mid \Theta) = \ln P(X, Z \mid \Theta) - \ln P(Z \mid X, \Theta)$，等式两边同时求关于 $Z \mid X, \Theta^t$ 的期望，有

$$\mathbb{E}_{Z|X,\Theta^t}[\ln P(X \mid \Theta)] = \mathbb{E}_{Z|X,\Theta^t}[\ln P(X, Z \mid \Theta)] - \mathbb{E}_{Z|X,\Theta^t}[\ln P(Z \mid X, \Theta)]$$

因为 $\ln P(X \mid \Theta)$ 与 Z 无关，所以

$$\mathbb{E}_{Z|X,\Theta^t}[\ln P(X \mid \Theta)] = \int_Z P\left(Z \mid X, \Theta^t\right) \ln P(X \mid \Theta) \mathrm{d}Z = \ln P(X \mid \Theta)$$

而 $\mathbb{E}_{Z|X,\Theta^t}[\ln P(X, Z \mid \Theta)] = Q\left(\Theta, \Theta^t\right)$ ，因为

$$\Theta^{t+1} = \arg\max_{\Theta} Q\left(\Theta, \Theta^t\right)$$

必然有

$$Q\left(\Theta^{t+1}, \Theta^t\right) \geq Q\left(\Theta, \Theta^t\right)$$

令 $\Theta = \Theta^t$，则

$$Q\left(\Theta^{t+1}, \Theta^t\right) \geq Q\left(\Theta^t, \Theta^t\right)$$

下面只需证

$$\mathbb{E}_{Z|X,\Theta^t}\left[\ln P\left(Z \mid X, \Theta^{t+1}\right)\right] \leq \mathbb{E}_{Z|X,\Theta^t}\left[\ln P\left(Z \mid X, \Theta^t\right)\right]$$

即证

$$\mathbb{E}_{Z|X,\Theta^t}\left[\ln \frac{P\left(Z \mid X, \Theta^{t+1}\right)}{P\left(Z \mid X, \Theta^t\right)}\right] = -D_{KL}\left[P\left(Z \mid X, \Theta^t\right) \| P\left(Z \mid X, \Theta^{t+1}\right)\right] \leq 0$$

根据 KL 散度的性质知上式成立，所以 $\{P(X \mid \Theta^n)\}$ 单调递增有上界，收敛性得证。

作业 7.4

假设模型为 $\lambda = [A,B,\pi]$，则

$$P(x_{n+1} \mid x_1,\ldots,x_n) = \frac{P(x_{n+1},x_n \mid x_1,\ldots,x_{n-1})}{P(x_n \mid x_1,\ldots,x_{n-1})}$$

记 $P(n) = P(x_n \mid x_1,\ldots,x_{n-1}),P(1) = P(x_1)$，则

$$P(x_{n+1},x_n,\ldots,x_1) = \prod_{i=1}^{n+1} P(i)$$

又因为已知

$$P(x_{n+1},x_n,\ldots,x_1) = \sum_{y_t} \alpha(y_t)\beta(y_t)$$

所以只需要利用上述两式不断递推，即可求出 $P(n+1)$。

作业 14.1

（1）$p(D,\mu,\lambda) = p(\mu,\lambda)p(D|\mu,\lambda) = p(\mu,\lambda)\prod_{i=1}^{m} p(x_i|\mu,\lambda) = \frac{1}{\sqrt{2\pi(\kappa_0\lambda)^{-1}}}\cdot$ $\exp\left\{-\frac{1}{2(\kappa_0\lambda)^{-1}}(\mu-\mu_0)^2\right\}\cdot\frac{1}{\Gamma(a_0)}b_0^{a_0}\lambda^{a_0-1}\exp\{-b_0\lambda\}\cdot\left(\frac{\lambda}{2\pi}\right)^{\frac{m}{2}}\exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{m}(x_i-\mu)^2\right\}$

（2）由变分推断的推导知：

$$\sum_{i=1}^{m}\log P(x_i) = \log P(x) \geq \mathbb{E}[\log P(Z,x)] - \mathbb{E}[\log q(Z)]$$

所以证据下界为 $\mathbb{E}[\log P(Z,x)] - \mathbb{E}[\log q(Z)] = \mathbb{E}_q[\log P(\lambda)] + \mathbb{E}_q[\log P(\mu|\lambda)] + \mathbb{E}_q[\log P(x|\mu,\lambda)] - \mathbb{E}_q[\log q(\lambda)] - \mathbb{E}_q[\log q(\mu)]$，证明如下：

变分推断的目标是

$$q^*(Z) = \underset{q(Z)}{\arg\min} D_{KL}[q(Z)||P(Z|x)]$$

其中

$$D_{KL}[q(Z)||p(Z|x)] = \mathbb{E}[\log q(Z)] - \mathbb{E}[\log P(Z,x)] + \log P(x)$$

由于 $D_{KL} \geq 0$，所以

$$\log P(x) \geq \mathbb{E}[\log P(Z,x)] - \mathbb{E}[\log q(Z)]$$

不等号右边就是下界。

（3）通过最大化 L 来最小化 $KL(q||P)$，偏导

$$\frac{\partial L}{\partial q_\lambda(\mu)} = \mathbb{E}_\lambda[\log P(\mu|\lambda)] + \mathbb{E}_\lambda[\log P(D|\mu,\lambda)] - \log q(\mu) = 0$$

则

$$\begin{aligned}\log q^*(\mu) &= -\frac{\mathbb{E}\lambda\kappa_0}{2}(\mu-\mu_0)^2 - \frac{\mathbb{E}\lambda}{2}\sum_{i=1}^{m}(x_i-\mu)^2 \\ &= -\frac{\mathbb{E}\lambda}{2}\left[(\kappa_0+m)\mu^2 + \sum_{i=1}^{m}x_i^2 - 2\mu(\kappa_0\mu_0+m\bar{x})\right] \\ &= -\frac{\mathbb{E}\lambda}{2}\left[(\kappa_0+m)\left(\mu-\frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m}\right)^2 + \sum_{i=1}^{m}x_i^2 - \frac{(\kappa_0\mu_0+m\bar{x})^2}{\kappa_0+m}\right]\end{aligned}$$

后两项不影响 $q(\mu)$ 的分布，所以

$$q(\mu) \sim \mathcal{N}\left(\mu \mid \frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m}, [(\kappa_0+m)\mathbb{E}\lambda]^{-1}\right)$$

又

$$\frac{\partial L}{\partial q_\mu(\lambda)} = \mathbb{E}_\mu[\log P(D|\lambda,\mu)] + \mathbb{E}[\log(\mu|\lambda)] + \mathbb{E}_\mu[\log P(\lambda)] - \log q(\lambda) = 0$$

所以

$$\begin{aligned}\log q^*(\lambda) &= -\frac{\lambda}{2}\mathbb{E}_\mu[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^m(x_i-\mu)^2]+(a_0-1)\log\lambda-b_0\lambda+\frac{m+1}{2}\log\lambda\\ &= (a_0+\frac{m-1}{2})\log\lambda-(b_0+\frac{1}{2}\mathbb{E}_\mu[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^m(x_i-\mu)^2])\lambda\end{aligned}$$

即

$$q^*(\lambda)\sim \mathrm{Gam}(\lambda|a_0+\frac{m-1}{2},b_0+\frac{1}{2}\mathbb{E}_\mu[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^m(x_i-\mu)^2])$$

所以 $q^*(\mu,\lambda)\sim\mathcal{N}(\mu|\frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m},[(\kappa_0+m)\mathbb{E}\lambda]^{-1})\mathrm{Gam}(\lambda|a_0+\frac{m-1}{2},b_0+\frac{1}{2}\mathbb{E}_\mu[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^m(x_i-\mu)^2])$。

在无先验的情况下 $\mu_0=a_0=b_0=\kappa_0=0$，且

$$\mathbb{E}\lambda=\frac{a_0+\frac{m-1}{2}}{b_0+\frac{1}{2}\mathbb{E}_\mu[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^m(x_i-\mu)^2]}$$

$$\mathbb{E}\mu^2=\bar{x}^2+\frac{1}{m\mathbb{E}\lambda}\qquad \mathbb{E}\mu=\mu_m=\bar{x}$$

联立解得

$$\mathbb{E}\lambda=\frac{1}{\mathrm{Var}X}$$

代回含 $\mathbb{E}\lambda$ 的式子得到 λ^*,b^*，即可得

$$P(\mu,\lambda|D)\sim\mathcal{N}(\mu|\bar{x}^2,\lambda^*)\mathrm{Gam}(\lambda|\frac{m-1}{2},b^*)$$

作业 14.2

条件随机场的预测问题是给定条件随机场 $P(Y\mid X)$ 和输入序列 (观测序列) x, 求条件概率最大的输出序列 (标记序列) y^*, 即对观测序列进行标注。

由 $P_w(y|x)=\frac{\exp(w\cdot F(y,x))}{Z_w(x)}$ 可得：

$$\begin{aligned}y^* &= \arg\max_y P_w(y\mid x)\\ &= \arg\max_y\frac{\exp(w\cdot F(y,x))}{Z_w(x)}\\ &= \arg\max_y\exp(w\cdot F(y,x))\\ &= \arg\max_y(w\cdot F(y,x))\end{aligned}$$

于是, 问题转化为求非规范化概率最大的最优路径问题

$$\max_y(w\cdot F(y,x))$$

这里, 路径表示标记序列。其中,

$$
\begin{aligned}
& w=\left(w_1, w_2, \cdots, w_K\right)^{\top} \\
& F(y, x)=\left(f_1(y, x), f_2(y, x), \cdots, f_K(y, x)\right)^{\top} \\
& f_k(y, x)=\sum_{i=1}^n f_k\left(y_{i-1}, y_i, x, i\right), \quad k=1,2, \cdots, K
\end{aligned}
$$

为了求解最优路径，目标函数写成如下形式：

$$
\max _y \sum_{i=1}^n w \cdot F_i\left(y_{i-1}, y_i, x\right)
$$

其中局部特征向量

$$
F_i\left(y_{i-1}, y_i, x\right)=\left(f_1\left(y_{i-1}, y_i, x, i\right), f_2\left(y_{i-1}, y_i, x, i\right), \cdots, f_K\left(y_{i-1}, y_i, x, i\right)\right)^{\top}
$$

下面是用维特比算法解决此问题的步骤。

首先求出位置 1 的各个标记 $j=1,2, \cdots, m$ 的非规范化概率

$$
\delta_1(j)=w \cdot F_1\left(y_0=\text{start}, y_1=j, x\right), \quad j=1,2, \cdots, m
$$

由递推公式, 求出到位置 i 的各个标记 $l=1,2, \cdots, m$ 的非规范化概率的最大值, 同时记录非规范化概率最大值的路径

$$
\delta_i(l)=\max _{1 \leqslant j \leqslant m}\left\{\delta_{i-1}(j)+w \cdot F_i\left(y_{i-1}=j, y_i=l, x\right)\right\}, \quad l=1,2, \cdots, m
$$

$$
\Psi_i(l)=\arg \max _{1 \leqslant j \leqslant m}\left\{\delta_{i-1}(j)+w \cdot F_i\left(y_{i-1}=j, y_i=l, x\right)\right\}, \quad l=1,2, \cdots, m
$$

直到 $i=n$ 时终止，这时求得非规范化概率的最大值为

$$
\max _y(w \cdot F(y, x))=\max _{1 \leqslant j \leqslant m} \delta_n(j)
$$

及最优路径的终点

$$
y_n^*=\arg \max _{1 \leqslant j \leqslant m} \delta_n(j)
$$

由此最优路径终点返回

$$
y_i^*=\Psi_{i+1}\left(y_{i+1}^*\right), \quad i=n-1, n-2, \cdots, 1
$$

求得最优路径 $y^*=\left(y_1^*, y_2^*, \cdots, y_n^*\right)^{\top}$。