7.4 实践中使用式(7.15)决定分类类别时，若数据的维数非常高，则概率连乘 $\prod_{i=1}^{d} P(x_i \mid c)$ 的结果通常会非常接近于 0 从而导致下溢. 试述防止下溢的可能方案.

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c)\prod_{i=1}^{d} P(x_i \mid c)\ , \qquad (7.15)$$

为防止下溢，可对上式取对数，将连乘变成连加.

即 $h_{nb}(x) = \arg\max_{c\in y} \log P(c) + \sum_{i=1}^{d} \log(P(x_i|c))$

若结果出现 $-\infty$ 溢出，可以在每次取对数时除以总次数，

即 $h_{nb}(x) = \arg\max_{c\in y} \dfrac{\log P(c)}{d+1} + \sum_{i=1}^{d} \dfrac{\log(P(x_i|c))}{d+1}$

7.5 试证明：二分类任务中两类数据满足高斯分布且方差相同时，线性判别分析产生贝叶斯最优分类器.

假设同先验；参见 3.4 节.

贝叶斯最优分类器 $h^*(x) = \arg\max_{c\in y} P(c|x)$

由贝叶斯定理，$P(c|x) = \dfrac{P(c,x)}{P(x)} = \dfrac{P(c)P(x|c)}{P(x)}$

若数据服从高斯分布，则有 $P(x|c) \sim N(\mu_c, \Sigma_c)$，$\mu_c$ 为均值向量，Σ_c 为方差矩阵.

$h^*(x) = \arg\max_{c\in y} P(c|x) \iff \arg\max_{c\in y} P(x|c)P(c) \iff \arg\max_{c\in y} \log(f(x|c)P(c))$

$= \arg\max_{c\in y} \log \dfrac{1}{\sqrt{(2\pi)^d|\Sigma_c|}} e^{-\frac{1}{2}(x-\mu_c)^T\Sigma_c^{-1}(x-\mu_c)} + \log(P(c))$

$= \arg\max_{c\in y} -\frac{1}{2}(x-\mu_c)^T\Sigma_c^{-1}(x-\mu_c) + \log(P(c))$

$= \arg\max_{c\in y} -\frac{1}{2}(x^T\Sigma_c^{-1}x - 2x^T\Sigma_c^{-1}\mu) + \log(P(c))$

对于二分类任务，设类别 1 数据的均值向量为 μ_1，类别 0 数据的均值向量为 μ_0，

两类数据满足高斯分布且方差 Σ 相同.

贝叶斯决策边界为 $g_1(x) = x^T\Sigma^{-1}\mu_1 - x^T\Sigma^{-1}\mu_0 - (\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0) + \log(\frac{P(1)}{P(0)})$

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

对于线性判别分析，其投影界面方向向量 $w \propto S_w^{-1}(\mu_0-\mu_1)$ ，$S_w=\Sigma_0+\Sigma_1$

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

两类数据方差 Σ 相同， $w=\frac{1}{2}(\Sigma_0+\Sigma_1)^{-1}(\mu_0-\mu_1) = \Sigma^{-1}(\mu_1-\mu_0)$

两类别在投影面连线的中点可为 $\frac{1}{2}(\mu_1+\mu_0)^T w = \frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)$

故线性判别分析决策边界 $g_2(x) = x^T\Sigma^{-1}(\mu_1-\mu_0) - \frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)$

又假设同先验，$\log(\frac{P(1)}{P(0)})=0$，故二者决策边界相同，得证.

- 证明EM算法的收敛性

设 $P(X|\theta)$ 为观测数据的似然函数，$\theta^{(i)}$ $(i=1,2,\cdots)$ 为EM算法得到的参数估计序列，

设 $L(\theta)=\log P(X|\theta)$， $L(\theta^{(i)})$ $(i=1,2,\cdots)$ 为对应的对数似然函数序列.

下证 $L(\theta^{(i)})$ 是单调递增的，即 $\log P(X|\theta^{(i+1)}) \geq \log P(X|\theta^{(i)})$.

$$P(X|\theta) = \frac{P(X,Z|\theta)}{P(Z|X,\theta)}$$

取对数有 $\log P(X|\theta) = \log(X,Z|\theta) - \log(Z|X,\theta)$

$$Q(\theta,\theta^{(i)}) = \sum_Z \log P(X,Z|\theta)P(Z|X,\theta^{(i)})$$

令 $H(\theta,\theta^{(i)}) = \sum_Z \log P(Z|X,\theta)P(Z|X,\theta^{(i)})$

于是对数似然函数可以写成 $\log P(X|\theta) = Q(\theta,\theta^{(i)}) - H(\theta,\theta^{(i)})$

$$\log P(X|\theta^{(i+1)}) - \log P(X|\theta^{(i)}) = [Q(\theta^{(i+1)},\theta^{(i)}) - Q(\theta^{(i)},\theta^{(i)})] - [H(\theta^{(i+1)},\theta^{(i)}) - H(\theta^{(i)},\theta^{(i)})]$$

$\because$ $\theta^{(i+1)}$ 使 $Q(\theta,\theta^{(i)})$ 达极大 $\therefore$ $Q(\theta^{(i+1)},\theta^{(i)}) - Q(\theta^{(i)},\theta^{(i)}) \geq 0$

$$H(\theta^{(i+1)},\theta^{(i)}) - H(\theta^{(i)},\theta^{(i)}) = \sum_Z\left(\log\frac{P(Z|X,\theta^{(i+1)})}{P(Z|X,\theta^{(i)})}\right)P(Z|X,\theta^{(i)})$$

$$\overset{\text{Jenson不等式}}{\leq} \log\left(\sum_Z\frac{P(Z|X,\theta^{(i+1)})}{P(Z|X,\theta^{(i)})}P(Z|X,\theta^{(i)})\right)$$

$$= \log P(Z|X,\theta^{(i+1)}) = 0$$

$\therefore$ $\log P(X|\theta^{(i+1)}) - \log P(X|\theta^{(i)}) \geq 0$

$P(X|\theta)$ 有界，$L(\theta)$ 单调，则 $L(\theta^{(i)})$ 收敛到某一值 L^*. EM算法的收敛性得证.

• 在HMM中，求解概率$P(x_{n+1}|x_1, x_2, . . . , x_n)$

$$P(x_{n+1}|x_1, x_2, \cdots, x_n) = \frac{P(x_1, x_2, \cdots, x_n, x_{n+1})}{P(x_1, x_2, \cdots, x_n)}$$

$$= \frac{P(y_1)P(x_1|y_1)\prod_{i=2}^{n+1} P(y_k|y_{k-1})P(x_k|y_k)}{P(y_1)P(x_1|y_1)\prod_{i=2}^{n} P(y_k|y_{k-1})P(x_k|y_k)}$$

$$= P(y_{n+1}|y_n)P(x_{n+1}|y_{n+1})$$

2. 假设数据集$D = \{x_1, x_2, ..., x_m\}$, 任意$x_i$是从均值为μ、方差为$\lambda^{-1}$的正态分布$\mathcal{N}(\mu, \lambda^{-1})$中独立采样而得到。假设$\mu$和$\lambda$的先验分布为$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1})\mathrm{Gam}(\lambda|a_0, b_0)$ 其中 $\mathrm{Gam}(\lambda|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0}\lambda^{a_0-1}\exp(-b_0\lambda)$

(1) 请写出联合概率分布$p(D, \mu, \lambda)$

(2) 请写出证据下界（即变分推断的优化目标），并证明其为观测数据边际似然$\sum_{i=1}^{m}\log p(x_i)$的下界

(3) 请用变分推断法近似推断后验概率$p(\mu, \lambda|D)$

(1) $p(D, \mu, \lambda) = p(D|\mu, \lambda)\, p(\mu, \lambda)$

$$= p(\mu, \lambda) \prod_{i=1}^{m} p(x_i|\mu, \lambda)$$

$$= \frac{1}{\sqrt{2\pi(k_0\lambda)^{-1}}} \exp\left(-\frac{1}{2(k_0\lambda)^{-1}}(\mu-\mu_0)^2\right) \frac{1}{\Gamma(a_0)} b^{a_0}\lambda^{a_0-1}\exp(-b_0\lambda)\left(\frac{\lambda}{2\pi}\right)^{\frac{m}{2}}\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{m}(x_i-\mu)^2\right)$$

(2) 证据下界： $E[\log p(z|x)] - E[\log q(z)]$

$$= E_q[\log P(\lambda)] + E_q[\log P(\mu|\lambda)] + E_q[\log P(x|\mu, \lambda)] - E_q[\log q(\lambda)] - E_q[\log q(\mu)]$$

变分推断的目的是寻找一个参数的概率密度分布 $q(z)$, s.t. $q^*(z) = \arg\min_{q(z)} KL(q(z)\|p(z|x))$

$KL(q(z)\|p(z|x)) = E[\log q(z)] - E[\log p(z|x)] + \log p(x)$

$\because$ KL散度 ≥ 0 $\quad\therefore \log p(x) \geq E[\log p(z|x)] - E[\log q(z)]$

$\sum_{i=1}^{m}\log P(x_i) = \log P(X) \geq E[\log(z|x)] - E[\log q(z)]$

故证据下界也为 $\sum_{i=1}^{m}\log P(x_i)$ 的下界.

(3) $KL(q\|p) = -L + \ln p$ ，可通过最大化 L 使得 KL 散度最小

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

$$\log q^*(\mu) = -\frac{E\lambda k_0}{2}(\mu-\mu_0)^2 - \frac{E\lambda}{2}\sum_{i=1}^{m}(x_i-\mu)^2$$

$$= -\frac{E\lambda}{2}\left[(k_0+m)\mu^2 + \sum_{i=1}^{m}x_i^2 - 2\mu(k_0\mu_0 + m\bar{x})\right]$$

$$= -\frac{E\lambda}{2}\left[(k_0+m)\left(\mu - \frac{k_0\mu_0+m\bar{x}}{k_0+m}\right)^2 + \sum_{i=1}^{m}x_i^2 - \frac{(k_0\mu_0+m\bar{x})^2}{k_0+m}\right]$$

$\therefore q^*(\mu) \sim N(\mu \mid \mu_m, \lambda_m^{-1})$ ，其中 $\mu_m = \dfrac{k_0\mu_0 + m\bar{x}}{k_0+m}$ ， $\lambda_m = (k_0+m)E\lambda$

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

$$\log q^*(\lambda) = -\frac{\lambda}{2}E_\mu\left[k_0(\mu-\mu_0)^2 + \sum_{i=1}^{m}(x_i-\mu)^2\right] + (a_0-1)\log\lambda - b_0\lambda + \frac{m+1}{2}\log\lambda$$

$$= \left(a_0 + \frac{m-1}{2}\right)\log\lambda - \left(b_0 + \frac{1}{2}E_\mu\left[k_0(\mu-\mu_0)^2 + \sum_{i=1}^{m}(x_i-\mu)^2\right]\right)\lambda$$

$\therefore q^*(\lambda) \sim Gam(\lambda \mid a_m, b_m)$ ，其中 $a_m = a_0 + \frac{m+1}{2}$ ， $b_m = b_0 + \frac{1}{2}E_\mu\left[k_0(\mu-\mu_0)^2 + \sum_{i=1}^{m}(x_i-\mu)^2\right]$

$\because$ 独立 $\therefore q^*(\mu,\lambda) = q^*(\mu)q^*(\lambda) \sim N(\mu \mid \mu_m, \lambda_m^{-1})\,Gam(\lambda \mid a_m, b_m)$

无先验时，有 $\mu_0 = a_0 = b_0 = k_0 = 0$

$E\mu^2 = (\bar{x})^2 + \dfrac{1}{mE\lambda}$ $\quad E\mu = \mu_m = \bar{x}$ $\quad E\lambda = \dfrac{a_m}{b_m}$

联立得 $E\lambda = \dfrac{1}{\overline{x^2} - \bar{x}^2} = \dfrac{1}{Var\, x}$

带回 λ_m 、b_m 得 λ', b', 则 $P(\mu,\lambda|D) \sim N(\mu \mid \mu_m, \lambda')\,Gam(\lambda \mid a_m, b')$

3. PPT 46，给出CRF的预测问题的解法

在条件随机场预测问题中，CRF预测问题就成为了求非规范化概率最大的最优路径问题.

使用 Viterbi 算法：

输入：模型特征向量 $F(y,x)$，权值向量 w，观测序列 $x=(x_1,x_2,\cdots,x_n)$

输出：最优路径 $y^*=(y_1^*,y_2^*,\cdots,y_n^*)$.

1. 初始化位置1的各个标记 $j=1,2,\cdots,m$ 的非规范化概率.

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

2. 对 $i=2,3,\cdots,n$ 递推计算.

$$\delta_i(l)=\max_{1\le j\le m}\{\delta_{i-1}(j)+w\cdot F_i(y_{i-1}=j,y_i=l,x)\},\quad l=1,2,\cdots,m$$

$$\Psi_i(l)=\arg\max_{1\le j\le m}\{\delta_{i-1}(j)+w\cdot F_i(y_{i-1}=j,y_i=l,x)\},\quad l=1,2,\cdots,m$$

3. 终止

$$\max_y(w\cdot F(y,x))=\max_{1\le j\le m}\delta_n(j)$$

$$y_n^*=\arg\max_{1\le j\le m}\delta_n(j)$$

4. 返回路径

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

Ψ 保存每个节点的最大非规范化概率的前导节点，即当前节点的最优前节点，最优路径先已知最优终结点，然后根据最优终结点从 Ψ 里面求出他们最优前前节点，依次求出 $y_n^*,y_{n-1}^*,\cdots,y_1^*$.

则 $y^*=(y_1^*,y_2^*,\cdots,y_n^*)^T$ 就是条件随机场预测问题的最优路径.