HW7

PB20111689 蓝俊玮

习题 7.4

> 实践中使用式 $h_{nb}(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c)\prod_{i=1}^{d} P(x_i|c)$ 决定分类类别时，若数据的维数非常高，则概率连乘 $\prod_{i=1}^{d} P(x_i|c)$ 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。

连乘操作容易造成下溢，通常使用对数似然：$\log P(c) + \log\prod_{i=1}^{d} P(x_i|c)$ 求解即可。在使用对数似然的时候，求概率的时候应该使用拉普拉斯修正，防止出现 $\log 0$ 的情况。

习题 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）

最小化分类错误率的贝叶斯最优分类器为：

$$h^*(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c|\boldsymbol{x})$$

基于贝叶斯定理，$P(c|\boldsymbol{x})$ 可以写成

$$p(c|\boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})}$$

所以最小化分类错误率的贝叶斯最优分类器可以转化为：$h^*(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(\boldsymbol{x}|c)P(c)$。那么在数据满足高斯分布的时候，则有：$h^*(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} f(\boldsymbol{x}|c)P(c)$。对其取对数转化，则可以得到：

$$\begin{aligned}
h^*(\boldsymbol{x}) =& \arg\max_{c\in\mathcal{Y}} \log\left(f(\boldsymbol{x}|c)P(c)\right) \\
=& \arg\max_{c\in\mathcal{Y}} \log\left(\frac{1}{\sqrt{2\pi}^n\sqrt{|\Sigma|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu_c)^T\Sigma^{-1}(\boldsymbol{x}-\mu_c)\right)\right) + \log P(c) \\
=& \arg\max_{c\in\mathcal{Y}} \log -\frac{1}{2}(\boldsymbol{x}-\mu_c)^T\Sigma^{-1}(\boldsymbol{x}-\mu_c) + \log P(c) \\
=& \arg\max_{c\in\mathcal{Y}} \boldsymbol{x}^T\Sigma^{-1}\mu_c - \frac{1}{2}\mu_c^T\Sigma^{-1}\mu_c + \log P(c)
\end{aligned}$$

在二分类任务中，贝叶斯决策边界可以表示为：

$$\begin{aligned}
g(x) =& \boldsymbol{x}^T\Sigma^{-1}\mu_1 - \boldsymbol{x}^T\Sigma^{-1}\mu_0 - \left(\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0\right) + \log\left(\frac{P(1)}{P(0)}\right) \\
=& \boldsymbol{x}^T\Sigma^{-1}(\mu_1-\mu_0) - \frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0) + \log\left(\frac{P(1)}{P(0)}\right)
\end{aligned}$$

再看线性判别分析：因为两个类别的方差相同时有 $w = \frac{1}{2}\Sigma^{-1}(\mu_1-\mu_0)$，则线性判别分析的决策边界可以表示为：$g(x) = x^T\Sigma^{-1}(\mu_1-\mu_0) - \frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)$。

因为题目中提到了假设同先验，所以可以无需考虑 $\log\left(\frac{P(1)}{P(0)}\right)$，因此贝叶斯最优分类器和线性判别分析的决策边界时相同的。

T3

> 证明 EM 算法的收敛性

设 $P(Y|\theta)$ 为观测数据的似然函数，$\theta^{(i)}$ 为 EM 算法得到的参数估计序列，$P(Y|\theta^{(i)})$ 为对应的似然函数序列。

由于 $P(Y|\theta)=\frac{P(Y,Z|\theta)}{P(Z|Y,\theta)}$，取对数有：

$$\log P(Y|\theta)=\log P(Y,Z|\theta)-\log P(Z|Y,\theta)$$

由 Q 函数 $Q(\theta,\theta^{(i)})=\sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^{(i)})$，设函数 $H(\theta,\theta^{(i)})=\sum_Z \log P(Z|Y,\theta)P(Z|Y,\theta^{(i)})$

于是对数似然函数可以写成

$$\log P(Y|\theta)=Q(\theta,\theta^{(i)})-H(\theta,\theta^{(i)})$$

则有：

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

由于在 M 步的时候，$\theta^{(i+1)}$ 让 $Q(\theta,\theta^{(i)})$ 达到极大值，所以

$$Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})\geq 0$$

而对于后面一项有：

$$\begin{aligned}H(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})&=\sum_Z(\log\frac{P(Z|Y,\theta^{(i+1)})}{P(Z|Y,\theta^{(i)})})P(Z|Y,\theta^{(i)})\\&\leq\log\left(\sum_Z\frac{P(Z|Y,\theta^{(i+1)})}{P(Z|Y,\theta^{(i)})}P(Z|Y,\theta^{(i)})\right)\\&=\log\left(\sum_Z P(Z|Y,\theta^{(i+1)})\right)=0\end{aligned}$$

因此得知：

$$P(Y|\theta^{(i+1)})\geq P(Y|\theta^{(i)})$$

所以单调递增且有上界为 1，说明 EM 算法是收敛的。

T4

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

记 $P_n=P(x_n|x_1,x_2,\ldots,x_{n-1})$，则有

$$\prod_{i=1}^{n+1}P_i=P(x_1,x_2,\ldots,x_{n+1})$$

那么就有：

$$
\begin{aligned}
P_{n+1} = P(x_{n+1}|x_1, x_2, \ldots, x_n) =& \frac{P(x_1, x_2, \ldots, x_{n+1})}{P(x_1, x_2, \ldots, x_n)} \\
=& \frac{P(y_1)P(x_1|y_1)\prod_{k=2}^{n+1} P(y_k|y_{k-1})P(x_k|y_k)}{P(y_1)P(x_1|y_1)\prod_{k=2}^{n} P(y_k|y_{k-1})P(x_k|y_k)} \\
=& P(y_{n+1}|y_n)P(x_{n+1}|y_{n+1})
\end{aligned}
$$