第2章　贝叶斯学习基础

思考与计算

1. 对于2.1节给出的示例:“假设某个动物园里的雌性和雄性熊猫的比例是4:6，雌性熊猫中 90%的熊猫是干净整洁的，雄性熊猫中 20%是干净整洁的”。计算在该动物园中看到一只干净整洁的雄性熊猫的概率是多少？如果看到一只熊猫是干净整洁的，它是雄性的概率是多少？

答：已知 $p(F)=0.4$，$p(M)=0.6$，$p(C|F)=0.9$，$p(C|M)=0.2$

$$p(C,M)=p(C|M)\times p(M)=0.2\times 0.6=0.12 \tag{2-1}$$

$$p(M|C)=\frac{p(C|M)\times p(M)}{p(C|M)\times p(M)+p(C|F)\times p(F)}=\frac{0.12}{0.12+0.36}=0.25 \tag{2-2}$$

2. 举例说明最小风险贝叶斯决策与最小错误率贝叶斯决策的不同。

答：以基于白细胞浓度的血液疾病诊断为例，分别说明使用最小错误率贝叶斯决策和最小风险贝叶斯决策的决策过程。

假设患病白细胞浓度服从均值为 2000，标准差为 1000 的正态分布，未患病白细胞浓度服从均值为 7000，标准差为 3000 的正态分布，患病的人数比例为 0.5%，问当白细胞浓度为 3000 时，应该做出什么决策？

设 w 表示是否患病，x 表示白细胞浓度，根据题意可以得到

$$p(w=1)=0.5\% \tag{2-3}$$

$$p(w=2)=99.5\% \tag{2-4}$$

$$p(x|w=1)=\mathcal{N}(2000,1000^2) \tag{2-5}$$

$$p(x|w=2)=\mathcal{N}(7000,3000^2) \tag{2-6}$$

（1）若进行最小错误率贝叶斯决策，则需要根据贝叶斯公式计算随机变量 w 的后验分布，计算结果如下：

$$p(w=1|x)=\frac{p(x|w=1)p(w=1)}{p(x)}=1.9\% \tag{2-7}$$

$$p(w=2|x)=\frac{p(x|w=2)p(w=2)}{p(x)}=98.1\% \tag{2-8}$$

其中

$$p(x)=p(x|w=1)p(w=1)+p(x|w=2)p(w=2) \tag{2-9}$$

贝叶斯最小错误率决策会选择后验概率最大的类别，即 $h(x)=2$。

（2）若进行最小风险贝叶斯决策，需考虑不同决策的损失，假设决策损失矩阵为（只是假设，合理的数值应该视真实情况而定）

$$\Lambda=\begin{bmatrix}0 & 100\\ 1 & 0\end{bmatrix} \tag{2-10}$$

其中 λ_{ij} 表示将第 i 类数据判别为第 j 类的损失，也可以用 $\lambda(h(x)=j|w=i)$ 表示。在该例子中，λ_{12} 表示将患病判别为正常的损失，λ_{21} 表示将正常判别为患病的损失。最小风险决策综合考虑各种决策的损失，选择条件风险最小的类别，其中条件风险的计算如下：

$$R(h(x)|x)=\sum_i \lambda(h(x)|w=i)p(w=i|x) \tag{2-11}$$

可以得到将 x 判别为不同类别的条件风险为

$$\begin{aligned}&R(h(x)=1|x)\\&=\lambda(h(x)=1|w=1)p(w=1|x)+\lambda(h(x)=1|w=2)p(w=2|x)\\&=98.1\%\end{aligned} \tag{2-12}$$

$$\begin{aligned}&R(h(x)=2|x)\\&=\lambda(h(x)=2|w=1)p(w=1|x)+\lambda(h(x)=2|w=2)p(w=2|x)\\&=190\%\end{aligned} \tag{2-13}$$

最小风险贝叶斯决策会选择条件风险最小的类别，即 $h(x)=1$。

3. **给出在两类类别先验概率相等情况下，类条件概率分布是相等对角协方差矩阵的高斯分布的贝叶斯决策规则，并进行错误率分析。**

答：（1）首先给出决策面的表达式。根据类条件概率分布的高斯假设，可以

得到

$$p(\mathbf{x}\,|\,w=i)=\mathcal{N}(\boldsymbol{\mu}_i,\Sigma_i),\ i=1,...,C. \tag{2-14}$$

贝叶斯决策得到的判别函数为

$$\begin{aligned} g_i(\mathbf{x}) &= \ln p(\mathbf{x}\,|\,w=i)+\ln p(w=i) \\ &= -\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\Sigma_i|+\ln p(w=i)-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i). \end{aligned} \tag{2-15}$$

通过判别函数可以得到决策面 $g_i(x)=g_j(x)$，具体形式为

$$-\frac{1}{2}\left[(x-\boldsymbol{\mu}_i)^\top\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)-(\mathbf{x}-\boldsymbol{\mu}_j)^\top\Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right]+\ln\frac{p(w=i)}{p(w=j)}-\frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|}=0. \tag{2-16}$$

假设两类类别先验概率相等，即 $p(w=1)=p(w=2)$，那么 $\ln p(w=i)/p(w=j)=0$，则判别函数 $g_i(\mathbf{x})$ 中的 $\ln p(w=i)$ 可以忽略不计。当所有类别的协方差矩阵都相等且为对角阵时，假设 $\Sigma_1=\Sigma_2=...=\Sigma_C=\Lambda$，判别函数(2-15)可简化为

$$g_i(x)=-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top\Lambda^{-1}(\mathbf{x}-\boldsymbol{\mu}_i). \tag{2-17}$$

将上式展开，忽略与 i 无关的项 $\mathbf{x}^\top\Lambda^{-1}\mathbf{x}$，判别函数进一步简化为

$$g_i(\mathbf{x})=(\Lambda^{-1}\boldsymbol{\mu}_i)^\top\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_i^\top\Lambda^{-1}\boldsymbol{\mu}_i. \tag{2-18}$$

此时判别函数是 $\mathbf{x}$ 的线性函数，决策面是一个超平面。当决策区域 R_i 与 R_j 相邻时，决策面满足方程

$$g_i(\mathbf{x})-g_j(\mathbf{x})=0, \tag{2-19}$$

即

$$\left[\Lambda^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\right]^\top(\mathbf{x}-\mathbf{x}_0)=0, \tag{2-20}$$

其中

$$\mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j). \tag{2-21}$$

即 $\mathbf{x}_0$ 为 $\boldsymbol{\mu}_i$ 与 $\boldsymbol{\mu}_j$ 连线的中点。在两分类问题下，决策面方程为

$$\left[\Lambda^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]^{\top}\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^{\top}\Lambda^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=0. \tag{2-22}$$

（2）为了计算错误率，这里引入最小错误率决策的负对数似然比，

$$r(\mathbf{x})=-\ln p(\mathbf{x}\mid w=1)+\ln p(\mathbf{x}\mid w=2). \tag{2-23}$$

最小错误率贝叶斯决策可以表示为：

$$\begin{cases} x\text{被判定为第一类} & \text{如果} r(\mathbf{x})=-\ln p(\mathbf{x}\mid w=1)+\ln p(\mathbf{x}\mid w=2)<0, \\ \mathrm{x}\text{被判定为第二类} & \text{如果} r(\mathbf{x})=-\ln p(\mathbf{x}\mid w=1)+\ln p(\mathbf{x}\mid w=2)>0. \end{cases} \tag{2-24}$$

由于 $r(\mathbf{x})$ 是随机变量 $\mathbf{x}$ 的函数，因此 $r(\mathbf{x})$ 也是随机变量。记其条件概率密度函数为 $p(r\mid w)$，贝叶斯平均错误率的计算可以转变为关于 $r(\mathbf{x})$ 的积分。

令 $p_1(\text{error})$ 表示将第一类样本判定为第二类的错误率，$p_2(\text{error})$ 表示将第二类样本判定为第一类的错误率，则通过先验概率加权可得（平均）错误率，即

$$p(\text{error})=0.5p_1(\text{error})+0.5p_2(\text{error}), \tag{2-25}$$

其中每一类错误率可以表示为

$$\begin{aligned} p_1(\text{error})&=\int_{R_2}p(\mathbf{x}\mid w=1)dx=\int_{r_B}^{\infty}p(h\mid w=1)dh, \\ p_2(\text{error})&=\int_{R_1}p(\mathbf{x}\mid w=2)dx=\int_{-\infty}^{r_B}p(h\mid w=2)dh. \end{aligned} \tag{2-26}$$

其中 $r(\mathbf{x})$ 的决策边界为

$$r_B=0. \tag{2-27}$$

因此，如果知道 $r(\mathbf{x})$ 的条件概率密度函数，即可算出错误率 $p_1(\text{error})$ 和 $p_2(\text{error})$。根据 $p(\mathbf{x}\mid w=i)\sim\mathcal{N}(\boldsymbol{\mu}_i,\Sigma_i),\ \ i=1,2$，可得

$$\begin{aligned} r(\mathbf{x})&=-\ln p(\mathbf{x}\mid w=1)+\ln p(\mathbf{x}\mid w=2) \\ &=-\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\top}\Lambda^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\Lambda|\right] \\ &\quad+\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\top}\Lambda^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\Lambda|\right] \\ &=\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\top}\Lambda^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\top}\Lambda^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)+\frac{1}{2}\ln\frac{|\Lambda|}{|\Lambda|} \\ &=(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\top}\Lambda^{-1}\mathbf{x}+\frac{1}{2}(\boldsymbol{\mu}_1^{\top}\Lambda^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^{\top}\Lambda^{-1}\boldsymbol{\mu}_2). \end{aligned} \tag{2-28}$$

$\mathbf{x}$ 虽是 d 维高斯分布的随机变量，$r(\mathbf{x})$ 却是一维的随机变量，并且是关于 $\mathbf{x}$ 的线

性函数。上式可以看作对 $\mathbf{x}$ 的各分量进行线性组合，然后平移，所以 $r(\mathbf{x})$ 服从一维高斯分布。下面计算一维高斯分布 $p(r(\mathbf{x})\,|\,w=1)$ 的期望 m_1 和方差 σ_1：

$$\begin{aligned} m_1 &= \mathbb{E}[r(\mathbf{x})\,|\,w=1] \\ &= (\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\top}\Lambda^{-1}\boldsymbol{\mu}_1+\frac{1}{2}(\boldsymbol{\mu}_1^{\top}\Lambda^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^{\top}\Lambda^{-1}\boldsymbol{\mu}_2) \\ &= -\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\top}\Lambda^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2). \end{aligned} \tag{2-29}$$

令 $m=\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\top}\Lambda^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$，则 $m_1=-m$，并且

$$\sigma_1^2=\mathbb{E}\left[(r(\mathbf{x})-m_1)^2\,|\,w=1\right]=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\top}\Lambda^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=2m, \tag{2-30}$$

同理可得 $p(r\,|\,w=2)$ 的期望 m_2 和方差 σ_2 为

$$\begin{aligned} m_2 &= \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\top}\Lambda^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=m, \\ \sigma_2^2 &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\top}\Lambda^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=2m. \end{aligned} \tag{2-31}$$

现根据式(2-26)计算 $p_1(\text{error})$ 和 $p_2(\text{error})$，得到

$$\begin{aligned} p_1(\text{error}) &= \int_{r_B}^{\infty} p(r\,|\,w=1)dh \\ &= \int_{r_B}^{\infty}\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\exp\{-\frac{1}{2}(\frac{r+m}{\sigma})^2\}dh \\ &= \int_{r_B}^{\infty}(2\pi)^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\frac{r+m}{\sigma})^2\}d(\frac{r+m}{\sigma}) \\ &= \int_{\frac{r_B+m}{\sigma}}^{\infty}(2\pi)^{-\frac{1}{2}}\exp(-\frac{1}{2}\varphi^2)d\varphi, \end{aligned} \tag{2-32}$$

$$\begin{aligned} p_2(error) &= \int_{-\infty}^{r_B} p(r\,|\,w=2)dh \\ &= \int_{-\infty}^{r_B}(2\pi)^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\frac{r-m}{\sigma})^2\}d(\frac{r-m}{\sigma}) \\ &= \int_{-\infty}^{\frac{r_B-m}{\sigma}}(2\pi)^{-\frac{1}{2}}\exp(-\frac{1}{2}\varphi^2)d\varphi. \end{aligned} \tag{2-33}$$

其中，

$$r_B=0,\quad \sigma=\sqrt{2m}. \tag{2-34}$$

因此，$p_1(\text{error})$ 和 $p_2(\text{error})$ 表示为标准高斯分布 $\mathcal{N}(0,1)$ 在对应区域上的概率值。

4. 推导高斯分布的均值与协方差的最大似然估计。

答：最大似然估计的求解目标为

$$\arg\max_{\boldsymbol{\mu},\Sigma}\sum_{i=1}^{N}\ln N(\mathbf{x}_i\mid\boldsymbol{\mu},\Sigma)$$

下面分别求解均值和协方差。

（1）对对数似然函数关于均值求导并设置为零可以得到如下方程：

$$\begin{aligned}
&\frac{d\sum_{i=1}^{N}\ln N(\mathbf{x}_i\mid\boldsymbol{\mu},\Sigma)}{d\boldsymbol{\mu}}=\frac{d\sum_{i=1}^{N}\ln\frac{1}{2\pi\mid\Sigma\mid^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\}}{d\boldsymbol{\mu}}\\
&=\frac{-\frac{1}{2}\sum_{i=1}^{N}d(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}{d\boldsymbol{\mu}}\\
&=\frac{-\frac{1}{2}\sum_{i=1}^{N}(d(\mathbf{x}_i-\boldsymbol{\mu})^{\top})\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})+(\mathbf{x}_i-\boldsymbol{\mu})^{\top}d(\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}))}{d\boldsymbol{\mu}}\\
&=\frac{-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})\Sigma^{-1}(d(\mathbf{x}_i-\boldsymbol{\mu}))+(\mathbf{x}_i-\boldsymbol{\mu})^{\top}d(\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}))}{d\boldsymbol{\mu}}\\
&=\frac{\frac{1}{2}\sum_{i=1}^{N}2(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}d\boldsymbol{\mu}}{d\boldsymbol{\mu}}=\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}=\mathbf{0}^{\top}
\end{aligned}\tag{2-35}$$

解方程(2-35)得到

$$\boldsymbol{\mu}_{ml}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

（2）对对数似然函数关于协方差求导并设置为零可以得到如下方程：

$$
\begin{aligned}
&\arg\max_{\boldsymbol{\mu},\Sigma}\sum_{i=1}^{N}\ln N(\mathbf{x}_i\,|\,\boldsymbol{\mu},\Sigma)\\
&\frac{d\sum_{i=1}^{N}\ln N(\mathbf{x}_i\,|\,\boldsymbol{\mu},\Sigma)}{d\Sigma}=\frac{d\sum_{i=1}^{N}\ln\frac{1}{2\pi\,|\,\Sigma\,|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\}}{d\Sigma}\\
&=\frac{-\frac{N}{2}d\ln|\,\Sigma\,|-\frac{1}{2}\sum_{i=1}^{N}d((\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}))}{d\Sigma}\\
&=\frac{-\frac{N}{2}\mathrm{Tr}[\Sigma^{-1}d\Sigma]+\frac{1}{2}\sum_{i=1}^{N}d\mathrm{Tr}[(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(d\Sigma)\Sigma^{-1}]}{d\Sigma}\\
&=\frac{-\frac{N}{2}\mathrm{Tr}[\Sigma^{-1}d\Sigma]+\frac{1}{2}\sum_{i=1}^{N}d\mathrm{Tr}[\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(d\Sigma)]}{d\Sigma}\\
&=-N\Sigma^{-1}+\sum_{i=1}^{N}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}=0
\end{aligned}
\tag{2-36}
$$

解方程(2-36)得到

$$
\Sigma_{m\ell}=\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu}_{m\ell})(\mathbf{x}_i-\boldsymbol{\mu}_{m\ell})^{\top}
$$

5. 推导高斯分布的均值的最大后验估计。

答：均值$\boldsymbol{\mu}$的对数后验分布可以表示为

$$
\begin{aligned}
\ln p(\boldsymbol{\mu}\,|\,\mathcal{D})&=\sum_{i=1}^{N}\ln p(\mathbf{x}_i\,|\,\boldsymbol{\mu})+\ln p(\boldsymbol{\mu})+\text{const}\\
&=\sum_{i=1}^{N}\ln\mathcal{N}(\mathbf{x}_i\,|\,\boldsymbol{\mu},\Sigma)+\ln\mathcal{N}(\boldsymbol{\mu}\,|\,0,\Sigma_{\boldsymbol{\mu}})+\text{const},
\end{aligned}
\tag{2-37}
$$

其中 const 表示与均值$\boldsymbol{\mu}$无关的项。对对数后验关于均值求导并设置为零，可以得到如下方程：

$$
\begin{aligned}
&\frac{d\sum_{i=1}^{N}\ln\mathcal{N}(\mathbf{x}_i\,|\,\boldsymbol{\mu},\Sigma)+\ln\mathcal{N}(\boldsymbol{\mu}\,|\,0,\Sigma_{\boldsymbol{\mu}})}{d\boldsymbol{\mu}}\\
&=\frac{d\sum_{i=1}^{N}\ln\frac{1}{2\pi\,|\,\Sigma\,|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\}+d\ln\frac{1}{2\pi\,|\,\Sigma_{\boldsymbol{\mu}}\,|^{1/2}}\exp\{-\frac{1}{2}\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu}\}}{d\boldsymbol{\mu}}\\
&=\frac{-\frac{1}{2}\sum_{i=1}^{N}d(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})-\frac{1}{2}d\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu}}{d\boldsymbol{\mu}}\\
&=\frac{-\frac{1}{2}\sum_{i=1}^{N}(d(\mathbf{x}_i-\boldsymbol{\mu})^{\top})\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})+(\mathbf{x}_i-\boldsymbol{\mu})^{\top}d(\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}))-\frac{1}{2}(d\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^{\top}d\Sigma_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu})}{d\boldsymbol{\mu}}\\
&=\frac{-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(d(\mathbf{x}_i-\boldsymbol{\mu}))+(\mathbf{x}_i-\boldsymbol{\mu})^{\top}d(\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}))-\frac{1}{2}(\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}d\boldsymbol{\mu}+\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}d\boldsymbol{\mu})}{d\boldsymbol{\mu}}\\
&=\frac{\frac{1}{2}(\sum_{i=1}^{N}2(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}-2\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1})d\boldsymbol{\mu}}{d\boldsymbol{\mu}}=\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}+\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}=\mathbf{0}^{\top}
\end{aligned}
\tag{2-38}
$$

解方程(2-38)得到

$$
\begin{aligned}
\boldsymbol{\mu}_{\text{map}}&=N(N\Sigma^{-1}+\Sigma_{\boldsymbol{\mu}}^{-1})^{-1}\Sigma^{-1}\boldsymbol{\mu}_{\text{m}\ell}\\
&=N\Sigma_{\boldsymbol{\mu}}(N\Sigma_{\boldsymbol{\mu}}+\Sigma)^{-1}\boldsymbol{\mu}_{\text{m}\ell}.
\end{aligned}
$$

6. 推导高斯分布的均值的贝叶斯参数估计。

答：假设均值服从高斯分布$\boldsymbol{\mu}\sim\mathcal{N}(\mathbf{0},\Sigma_{\boldsymbol{\mu}})$，根据如下贝叶斯公式，

$$
p(\theta\,|\,\mathcal{D})=\frac{p(\mathcal{D}\,|\,\theta)p(\theta)}{p(\mathcal{D})}. \tag{2-39}
$$

可以得到均值的后验分布表达式为

$$
\begin{aligned}
p(\boldsymbol{\mu} \mid \mathcal{D}) &= p(\boldsymbol{\mu}) \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\mu}) \,/\, p(\mathcal{D}) \\
&= \mathcal{N}(\boldsymbol{\mu} \mid 0, \Sigma_{\boldsymbol{\mu}}) \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}, \Sigma) \,/\, p(\mathcal{D}) \\
&= \frac{1}{2\pi \mid \Sigma_{\boldsymbol{\mu}} \mid^{1/2}} \exp\{-\frac{1}{2} \boldsymbol{\mu}^{\top} \Sigma_{\boldsymbol{\mu}}{}^{-1} \boldsymbol{\mu}\} \prod_{i=1}^{N} \frac{1}{2\pi \mid \Sigma \mid^{1/2}} \exp\{-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\} \,/\, p(\mathcal{D}) \\
&= \frac{1}{2\pi \mid \Sigma_{\boldsymbol{\mu}} \mid^{1/2}} \left(\frac{1}{2\pi \mid \Sigma \mid^{1/2}} \right)^{N} \exp\{-\frac{1}{2} \boldsymbol{\mu}^{\top} \Sigma_{\boldsymbol{\mu}}{}^{-1} \boldsymbol{\mu} - \sum_{i=1}^{N} \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \,/\, p(\mathcal{D}) \\
&= \frac{1}{2\pi \mid \Sigma_{\boldsymbol{\mu}} \mid^{1/2}} \left(\frac{1}{2\pi \mid \Sigma \mid^{1/2}} \right)^{N} \exp\left\{ -\frac{1}{2} \boldsymbol{\mu}^{\top} \Sigma_{\boldsymbol{\mu}}{}^{-1} \boldsymbol{\mu} - \frac{1}{2} \sum_{i=1}^{N} \mathbf{x}_i{}^{\top} \Sigma^{-1} \mathbf{x}_i - 2\boldsymbol{\mu}^{\top} \Sigma^{-1} \mathbf{x}_i + \boldsymbol{\mu}^{\top} \Sigma^{-1} \boldsymbol{\mu} \right\} / \, p(\mathcal{D}) \\
&= \frac{1}{2\pi \mid \Sigma_{\boldsymbol{\mu}} \mid^{1/2}} \left(\frac{1}{2\pi \mid \Sigma \mid^{1/2}} \right)^{N} \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\mu}^{\top} (\Sigma_{\boldsymbol{\mu}}{}^{-1} + N\Sigma^{-1}) \boldsymbol{\mu} - 2N\boldsymbol{\mu}^{\top} \Sigma^{-1} \boldsymbol{\mu}_{\mathrm{m}\ell} + \sum_{i=1}^{N} \mathbf{x}_i{}^{\top} \Sigma^{-1} \mathbf{x}_i \right] \right\} / \, p(\mathcal{D}) \\
&= \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\mu}^{\top} (\Sigma_{\boldsymbol{\mu}}{}^{-1} + N\Sigma^{-1}) \boldsymbol{\mu} - 2N\boldsymbol{\mu}^{\top} \Sigma^{-1} \boldsymbol{\mu}_{\mathrm{m}\ell} + \sum_{i=1}^{N} \mathbf{x}_i{}^{\top} \Sigma^{-1} \mathbf{x}_i \right] \right\} \times \text{const} \\
&= \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\mu} - N(\Sigma_{\boldsymbol{\mu}}^{-1} + N\Sigma^{-1}) \Sigma^{-1} \boldsymbol{\mu}_{\mathrm{m}\ell} \right)^{\top} (\Sigma_{\boldsymbol{\mu}}^{-1} + N\Sigma^{-1})^{-1} \left(\boldsymbol{\mu} - N(\Sigma_{\boldsymbol{\mu}}^{-1} + N\Sigma^{-1}) \Sigma^{-1} \boldsymbol{\mu}_{\mathrm{m}\ell} \right) \right\} \times \text{const} \\
&= \mathcal{N}\left(N\Sigma_{\boldsymbol{\mu}} (N\Sigma_{\boldsymbol{\mu}} + \Sigma)^{-1} \boldsymbol{\mu}_{\mathrm{m}\ell}, (\Sigma_{\boldsymbol{\mu}}^{-1} + N\Sigma^{-1})^{-1} \right).
\end{aligned}
$$

第3章　逻辑回归

思考与计算

1.　选择一个 UCI 数据集，比较线性回归和岭回归的错误率。

答：（略）

2.　请编程实现二类分类的逻辑回归，要求采用牛顿法进行优化求解。

答：（略）

3.　证明一元高斯似然关于方差的共轭先验是逆伽马分布。

答：要证明一元高斯似然关于方差的共轭先验是逆伽马分布，需证明在该假设前提下，方差的后验分布是逆伽马分布。假设似然函数为如下高斯分布：

$$p(\mathbf{y}\mid X,\boldsymbol{\beta},\sigma^2)\propto(\sigma^2)^{-\frac{N}{2}}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-X\boldsymbol{\beta})^{\top}(\mathbf{y}-X\boldsymbol{\beta})\right), \tag{3-1}$$

假设方差的先验分布为逆伽马分布 Inv-Gamma(a_0,b_0)，即

$$p(\sigma^2)\propto(\sigma^2)^{-a_0-1}\exp(-\frac{b_0}{\sigma^2}), \tag{3-2}$$

根据贝叶斯公式可以得到方差的后验分布表达式如下

$$\begin{aligned}
p(\sigma^2\mid\mathbf{y},X,\boldsymbol{\beta}) &= p(\sigma^2)p(\mathbf{y}\mid X,\boldsymbol{\beta},\sigma^2)/\,p(\mathbf{y}\mid\boldsymbol{\beta})\\
&=(\sigma^2)^{-a_0-1}\exp\left\{-\frac{b_0}{\sigma^2}\right\}\times(\sigma^2)^{-\frac{N}{2}}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-X\boldsymbol{\beta})^{\top}(\mathbf{y}-X\boldsymbol{\beta})\right\}\times\text{const}\\
&=(\sigma^2)^{-a_0-1-\frac{N}{2}}\exp\left\{-\frac{b_0+\frac{1}{2}(\mathbf{y}-X\boldsymbol{\beta})^{\top}(\mathbf{y}-X\boldsymbol{\beta})}{\sigma^2}\right\}\times\text{const}\\
&=\text{Inv-Gamma}\left(a_0+\frac{N}{2},\ b_0+\frac{1}{2}(\mathbf{y}-X\boldsymbol{\beta})^{\top}(\mathbf{y}-X\boldsymbol{\beta})\right)
\end{aligned} \tag{3-3}$$

4.　思考如何优化使用 L_1 范数进行正则化的最小二乘。

答：对于目标函数不是连续可微的情况，可以用次梯度来进行优化，L_1 范数的次梯度表示为

$$\frac{\partial\|\boldsymbol{\beta}\|_{L_1}}{\partial\boldsymbol{\beta}}=\text{sign}(\boldsymbol{\beta})=\begin{cases}+1 & \beta_d>0\\ -1 & \beta_d<0\\ [-1,+1] & \beta_d=0\end{cases}$$

但次梯度存在两个问题：求解慢，通常不会产生稀疏解。此时可以用 Proximal Gradient Descent 对 L_1 范数正则化问题进行求解。求解过程如下。

使用 L_1 范数进行正则化的最小二乘的优化目标表达式如下：

$$S'=\sum_{i=1}^{N}\left(y_i-f(\mathbf{x}_i,\boldsymbol{\beta})\right)^2+\lambda\|\boldsymbol{\beta}\|_{L_1} \tag{3-4}$$

其中 $\sum_{i=1}^{N}\left(y_i-f(\mathbf{x}_i,\boldsymbol{\beta})\right)^2$ 是可导函数，记为 $g(\boldsymbol{\beta})$，且 $\lambda\|\boldsymbol{\beta}\|_{L_1}$ 不可导。

根据泰勒展开式 $g(\boldsymbol{\beta})$ 可以表示为

$$g(\boldsymbol{\beta})=g(\boldsymbol{\beta}_k)+(\nabla g(\boldsymbol{\beta}_k))^{\top}(\boldsymbol{\beta}-\boldsymbol{\beta}_k)+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_k)^{\top}H(g(\boldsymbol{\beta}_k))(\boldsymbol{\beta}-\boldsymbol{\beta}_k)+o(\|\boldsymbol{\beta}_k\|_2^2) \tag{3-5}$$

假设 $g(\boldsymbol{\beta})$ 满足 L-Lipschitz 条件，即存在常数 $a>0$ 使得

$$\|\nabla g(\boldsymbol{\beta}')-\nabla g(\boldsymbol{\beta})\|_2\le a\|\boldsymbol{\beta}'-\boldsymbol{\beta}\|_2\quad(\forall\boldsymbol{\beta},\boldsymbol{\beta}') \tag{3-6}$$

令 $H(g(\boldsymbol{\beta}_k))=a\times I$，公式(3-5)可以使用如下表达式近似

$$\begin{aligned}g(\boldsymbol{\beta})&\approx g(\boldsymbol{\beta}_k)+(\nabla g(\boldsymbol{\beta}_k))^{\top}(\boldsymbol{\beta}-\boldsymbol{\beta}_k)+\frac{a}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_k)^{\top}(\boldsymbol{\beta}-\boldsymbol{\beta}_k)\\&=\frac{a}{2}\left\|\boldsymbol{\beta}-\left(\boldsymbol{\beta}_k-\frac{1}{a}\nabla g(\boldsymbol{\beta}_k)\right)\right\|_2^2+\text{const}\end{aligned} \tag{3-7}$$

其中 const 是与变量 $\boldsymbol{\beta}$ 无关的常数。公式(3-7)中的 $g(\boldsymbol{\beta})$ 在 $\boldsymbol{\beta}=\boldsymbol{\beta}_k-\frac{1}{a}\nabla g(\boldsymbol{\beta}_k)$ 时取得最小值。

如果使用梯度下降来求得 $g(\boldsymbol{\beta})$ 的最小值，在第 $k+1$ 步，更新公式为

$$\boldsymbol{\beta}_{k+1}=\boldsymbol{\beta}_k-\frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \tag{3-8}$$

使用类似的思想来求解公式(3-4)的最小值，在第 $k+1$ 步，更新公式可以表示为

$$\boldsymbol{\beta}_{k+1}=\arg\max_{\boldsymbol{\beta}}\frac{a}{2}\left\|\boldsymbol{\beta}-\left(\boldsymbol{\beta}_k-\frac{1}{a}\nabla g(\boldsymbol{\beta}_k)\right)\right\|_2^2+\lambda\|\boldsymbol{\beta}\|_{L_1} \tag{3-9}$$

公式中的优化目标中可以分别计算 $\boldsymbol{\beta}$ 的每一维 $\boldsymbol{\beta}^d$，然后进行整合。优化问题可以表示为

$$\beta_{k+1}^{d} = \arg\max_{\beta^d} \frac{a}{2}\left(\beta^d - \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d \right)^2 + \lambda|\beta^d|$$

$$= \arg\max_{\beta^d} \frac{a}{2}\left(\beta^d - \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d \pm \frac{\lambda}{a} \right)^2 + \text{const}$$

$$= \begin{cases} \beta^d = \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d + \frac{\lambda}{a}, & \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d + \frac{\lambda}{a} < 0 \\ \beta^d = \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d - \frac{\lambda}{a}, & \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d - \frac{\lambda}{a} > 0 \\ \beta^d = 0, & -\frac{\lambda}{a} \le \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d \le \frac{\lambda}{a} \end{cases}$$

第4章 概率图模型基础

思考与计算

1. 设计一个贝叶斯网络的图模型，并写出所有变量的联合分布。

答：例如，图 4-1 所示的一种生成式混合高斯过程图模型，其中最左侧的节点以及中间方框中的 $\boldsymbol{\theta}_k$ 和 $\mathbf{I}_k$ 表示模型的超参数，其余中间与右侧方框中的节点表示模型的随机变量。

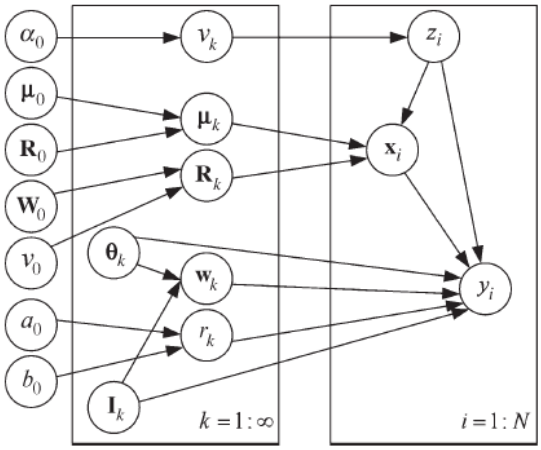
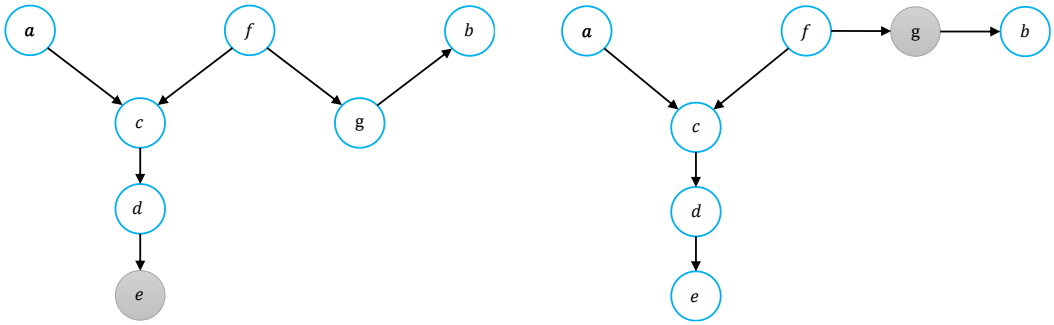


图 4-1 混合高斯过程的图模型

观察图 4-1 所示的贝叶斯网络的图模型，为了方便表示，使用 $\bar{\nu}$ 表示 $\nu_1,\nu_2,...,\nu_\infty$，中间方框中的其他随机变量使用类似表示，使用 $\mathbf{z}$ 表示 $z_1,z_2,...,z_N$，使用 $\mathbf{X}$ 表示 $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N$，使用 $\mathbf{y}$ 表示 $y_1,y_2,...,y_N$，可以得到如下联合分布：

$$
\begin{aligned}
&p(\bar{\nu},\bar{\boldsymbol{\mu}},\bar{\mathbf{R}},\bar{\mathbf{W}},\bar{r},\mathbf{z},\mathbf{X},\mathbf{y}) \\
&=\prod_{k=1}^{\infty} p(\nu_k\mid\alpha_0)p(\boldsymbol{\mu}_k\mid\boldsymbol{\mu}_0,\mathbf{R}_0)p(\mathbf{R}_k\mid\nu_0,\mathbf{W}_0)p(\mathbf{W}_k\mid\boldsymbol{\theta}_k,\mathbf{I}_k)p(r_k\mid a_0,b_0) \\
&\times\prod_{i=1}^{N} p(z_i\mid\bar{\nu})p(\mathbf{x}_i\mid z_i,\bar{\boldsymbol{\mu}},\bar{\mathbf{R}})p(y_i\mid z_i,\mathbf{x}_i,\bar{\mathbf{W}},\bar{r})
\end{aligned}
\tag{4-1}
$$

2. 分析图 4-7 所示的贝叶斯网络中的其他变量之间的条件独立性。



（a）示例一　　（b）示例二

图 4-7 使用 d-分隔判断条件独立性示例

答：下面分别分析图 4-7（a）与图 4-7（b）图中，节点 a 与其他所有变量的关系。

在图 4-7（a）中，节点 a 到节点 c,d,f,g,b 均只有一条路径。路径中，a 到 c 不被阻隔，a,c,d 是顺序结构且 c 未观测，因此 a 到 d 不被阻隔。由于节点 a,c,f 是汇总结构且 c 的后代节点 e 被观测，因此 a 到 f 不被阻隔；节点 c, f ,g 是发散结构且 f 未观测，因此 c 到 g 不被阻隔；节点 f,g,b 是顺序结构且 g 是未观测，因此 f 到 b 不被阻隔。总的来说，节点 a 到节点 c,d,f,g,b 均不是 d-分隔的，因此在给定节点 e 后，节点 a 和节点 c,d,f,g,b 均不是条件独立。

在图 4-7（b）中，节点 a 到节点 c,d,e,f,b 均只有一条路径。路径中，a 到 c 不被阻隔，a,c,d 是顺序结构且 c 未观测，因此 a 到 d 不被阻隔。c,d,e 是顺序结构且 d 未观测，因此 c 到 e 不被阻隔。由于节点 a,c,f 是汇总结构且 c 的后代节点不被观测，因此 a 到 f 被阻隔；节点 f,g,b 是顺序结构且 g 被观测，因此 f 到 b 是被阻隔的，所以 a 到 b 是 d-分隔的，因此在给定节点 g 后，节点 a 和 b 条件独立，即有 $a \perp f \mid g$ 和 $a \perp b \mid g$ 。

3. **写出图 4-13 所示的无向图的联合概率表示。**

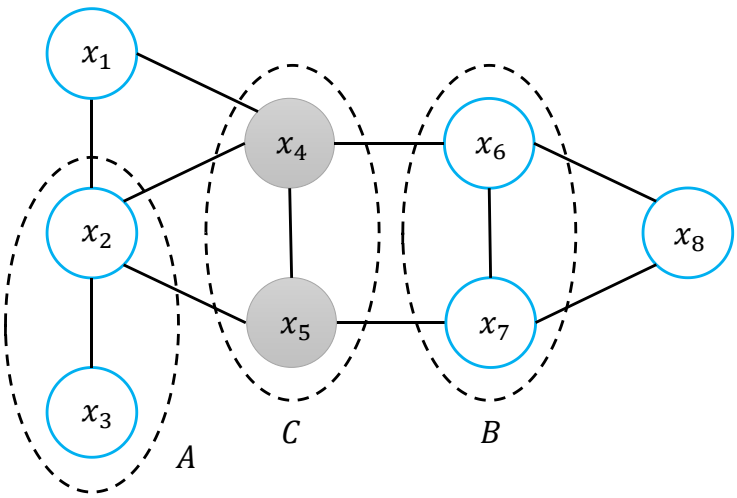


图 4-13 满足条件独立性 $A \perp B|C$ 的无向图模型示例

答：根据无向图模型的条件独立性，可以得到图 4-13 中的所有变量的联合分布表示为：

$$\begin{aligned} p(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8) &= p(x_4,x_5)p(x_1,x_2,x_3,x_6,x_7,x_8 \mid x_4,x_5) \\ &= p(x_4,x_5)p(x_1,x_2,x_3 \mid x_4,x_5)p(x_6,x_7,x_8 \mid x_4,x_5) \end{aligned} \tag{4-2}$$

若将概率分布根据最大团进行分解，则可以得到如下表示：

$$\begin{aligned} & p(x_4,x_5)p(x_1,x_2,x_3 \mid x_4,x_5)p(x_6,x_7,x_8 \mid x_4,x_5) \\ &= \frac{1}{Z_1}\phi_1(x_4,x_5)\frac{1}{Z_2}\prod\phi_2(x_1,x_2 \mid x_4,x_5)\phi_3(x_2,x_3 \mid x_4,x_5)\frac{1}{Z_3}\phi_4(x_6,x_7,x_8 \mid x_4,x_5) \end{aligned} \tag{4-3}$$

4. 使用和积算法推导出图 4-23 所示的因子图中其他变量的边缘分布。

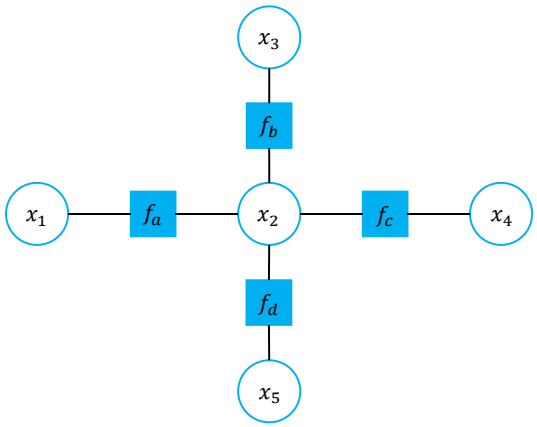
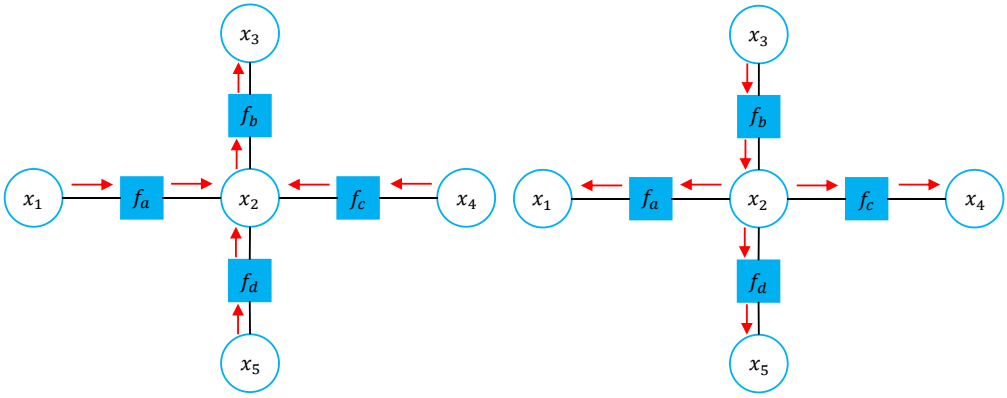


图 4-23 因子图示例

答：如图 4-23 所示，若只需计算节点 x_1,x_2,x_3,x_4,x_5 中某一个变量的边缘分布，只需将该变量作为根节点进行一次朝根节点方向的消息传递。若需要计算多个随机变量的边缘分布，则可以任意选取一个节点为根节点，然后进行两次消息传递既可以得出所有随机变量的边缘分布。这里将图 4-23 中的 x_3 作为根节点进行两次消息传递，两次消息传递示意图如图 4-24（a）和图 4-24（b）所示



（a）从叶子节点 x_1、x_4 和 x_5 向根节点 x_3 传递　（b）从根节点 x_3 向叶子节点 x_1、x_4 和 x_5 传递

图 4-24 对应图 4-23 的和积算法的消息流示意图

从叶子节点向根节点的消息传递如下：

$$\begin{aligned}
&\mu_{x_1\to f_a}(x_1)=1\\
&\mu_{f_a\to x_2}(x_2)=\sum_{x_1} f_a(x_1,x_2)\\
&\mu_{x_4\to f_c}(x_4)=1\\
&\mu_{f_c\to x_2}(x_2)=\sum_{x_4} f_c(x_2,x_4)\\
&\mu_{x_5\to f_d}(x_5)=1\\
&\mu_{f_d\to x_2}(x_2)=\sum_{x_5} f_d(x_2,x_5)\\
&\mu_{x_2\to f_b}(x_2)=\mu_{f_a\to x_2}(x_2)\mu_{f_c\to x_2}(x_2)\mu_{f_d\to x_2}(x_2)\\
&\mu_{f_b\to x_3}(x_3)=\sum_{x_2} f_b(x_2,x_3)\mu_{x_2\to f_b}(x_2).
\end{aligned} \tag{4-4}$$

从根节点向叶子节点的消息传递如下：

$$\begin{aligned}
&\mu_{x_3\to f_b}(x_3)=1\\
&\mu_{f_b\to x_2}(x_2)=\sum_{x_3} f_b(x_2,x_3)\\
&\mu_{x_2\to f_a}(x_2)=\mu_{f_b\to x_2}(x_2)\mu_{f_c\to x_2}(x_2)\mu_{f_d\to x_2}(x_2)\\
&\mu_{f_a\to x_1}(x_1)=\sum_{x_2} f_a(x_1,x_2)\mu_{x_2\to f_a}(x_2)\\
&\mu_{x_2\to f_c}(x_2)=\mu_{f_a\to x_2}(x_2)\mu_{f_b\to x_2}(x_2)\mu_{f_d\to x_2}(x_2)\\
&\mu_{f_c\to x_4}(x_4)=\sum_{x_2} f_c(x_2,x_4)\mu_{x_2\to f_c}(x_2)\\
&\mu_{x_2\to f_d}(x_2)=\mu_{f_a\to x_2}(x_2)\mu_{f_b\to x_2}(x_2)\mu_{f_c\to x_2}(x_2)\\
&\mu_{f_d\to x_5}(x_4)=\sum_{x_2} f_d(x_2,x_5)\mu_{x_2\to f_d}(x_2).
\end{aligned} \tag{4-5}$$

然后，根据所得到的消息的值，可以计算任意节点的边缘分布，x_1,x_2,x_3,x_4,x_5 的边缘分布如下：

$$
\begin{aligned}
p(x_1) &= \mu_{f_a \to x_1}(x_2) = \sum_{x_2} f_a(x_1, x_2)\, \mu_{x_2 \to f_a}(x_2) \\
&= \sum_{x_2} f_a(x_1, x_2)\, \mu_{f_b \to x_2}(x_2)\, \mu_{f_c \to x_2}(x_2)\, \mu_{f_d \to x_2}(x_2) \\
&= \sum_{x_2} f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_c(x_2, x_4) \sum_{x_5} f_d(x_2, x_5) \\
p(x_2) &= \mu_{f_a \to x_2}(x_2)\, \mu_{f_b \to x_2}(x_2)\, \mu_{f_c \to x_2}(x_2)\, \mu_{f_d \to x_2}(x_2) \\
&= \sum_{x_1} f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_c(x_2, x_4) \sum_{x_5} f_d(x_2, x_5) \\
p(x_3) &= \mu_{f_b \to x_3}(x_3) = \sum_{x_2} f_b(x_2, x_3)\, \mu_{x_2 \to f_b}(x_2) \\
&= \sum_{x_2} f_b(x_2, x_3)\, \mu_{f_a \to x_2}(x_2)\, \mu_{f_c \to x_2}(x_2)\, \mu_{f_d \to x_2}(x_2) \\
&= \sum_{x_2} f_b(x_2, x_3) \sum_{x_1} f_a(x_1, x_2) \sum_{x_4} f_c(x_2, x_4) \sum_{x_5} f_d(x_2, x_5). \qquad (4\text{-}6) \\
p(x_4) &= \mu_{f_c \to x_4}(x_4) = \sum_{x_2} f_c(x_2, x_4)\, \mu_{x_2 \to f_c}(x_2) \\
&= \sum_{x_2} f_c(x_2, x_4)\, \mu_{f_a \to x_2}(x_2)\, \mu_{f_b \to x_2}(x_2)\, \mu_{f_d \to x_2}(x_2) \\
&= \sum_{x_2} f_c(x_2, x_4) \sum_{x_1} f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_5} f_d(x_2, x_5) \\
p(x_5) &= \mu_{f_d \to x_5}(x_4) = \sum_{x_2} f_d(x_2, x_5)\, \mu_{x_2 \to f_d}(x_2) \\
&= \sum_{x_2} f_d(x_2, x_5)\, \mu_{f_a \to x_2}(x_2)\, \mu_{f_b \to x_2}(x_2)\, \mu_{f_c \to x_2}(x_2) \\
&= \sum_{x_2} f_d(x_2, x_5) \sum_{x_1} f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_c(x_2, x_4)
\end{aligned}
$$

第7章 支持向量机

思考与计算

1. 查找资料理解并推导互补松弛条件。

答：互补松弛条件是在拉格朗日对偶优化中，保证对偶问题的解与原问题的解相同的必要条件之一，所有的必要条件也称为 KKT 条件。下面介绍拉格朗日对偶优化的原理，以及 KKT 条件的推导过程。

假设带有约束的优化问题表示为

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \end{array} \tag{7-1}$$

定义原问题的最优解为 x^*，目标的最优值为 p^*。

拉格朗日对偶优化的基本思想是将约束条件以一定的权重加入到原优化目标中，从而得到带有简单约束条件的优化问题，新的优化问题则叫做原问题的对偶问题。对偶变换会改变优化目标以及优化变量，并且在一定的条件下可以得到等价的解。

首先，分别对约束条件引入拉格朗日乘子（也叫做对偶变量）λ_i，ν_i，定义拉格朗日函数 L，具体表示如下：

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x), \tag{7-2}$$

其次，计算拉格朗日函数关于 x 的下界，得到拉格朗日对偶函数 $g(\lambda,\nu)$

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right). \tag{7-3}$$

进一步分析(7-3)与原问题最优值的关系。假设 $\tilde{x}$ 表示满足约束条件的任意变量，那么可以得到如下不等式：

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \le 0, \tag{7-4}$$

因此拉格朗日函数满足如下关系

$$L(\tilde{x},\lambda,\nu)=f_0(\tilde{x})+\sum_{i=1}^{m}\lambda_i f_i(\tilde{x})+\sum_{i=1}^{p}\nu_i h_i(\tilde{x})\le f_0(\tilde{x}). \tag{7-5}$$

拉格朗日对偶函数满足如下关系

$$g(\lambda,\nu)=\inf_{x\in\mathcal{D}}L(x,\lambda,\nu)\le L(\tilde{x},\lambda,\nu)\le f_0(\tilde{x}). \tag{7-6}$$

由于原问题的最优解为 x^* 也包含在 $\tilde{x}$ 中，因此可以得到

$$g(\lambda,\nu)\le p^\star. \tag{7-7}$$

至此我们可以得到，对偶函数一定小于或等于原优化问题的最优值。因此，在乘子变量非负的约束下，最大化对偶函数的值（即求解对偶优化问题）就很有意义，它可能会达到原优化问题的最优值。定义对偶函数的解为 λ^*,ν^*，最优值为 d^*，可以得到 $d^*\le p^*$

下面分析满足什么条件时，对偶函数的解与原问题的解等价。

首先，根据原问题和对偶问题的定义，可以得到如下等式和不等式关系

$$\begin{aligned} f_0(x^\star) &= g(\lambda^\star,\nu^\star)\\ &= \inf_x\left(f_0(x)+\sum_{i=1}^{m}\lambda_i^\star f_i(x)+\sum_{i=1}^{p}\nu_i^\star h_i(x)\right)\\ &\le f_0(x^\star)+\sum_{i=1}^{m}\lambda_i^\star f_i(x^\star)+\sum_{i=1}^{p}\nu_i^\star h_i(x^\star)\\ &\le f_0(x^\star). \end{aligned} \tag{7-8}$$

其中第一行等式表示原问题的解对偶解相同，第二行等式是将最优对偶解代入对偶函数的定义式(7-3)中的结果，第三行不等式利用了下界的含义，第四行利用 1 不等式约束条件。

根据公式(7-8)中的最后一行不等式可以得到所有的不等号都应该是等号。因此，可以得到

$$\sum_{i=1}^{m}\lambda_i^\star f_i(x^\star)=0. \tag{7-9}$$

由于每一个不等式约束都小于等于 0，且每一个对偶变量都是大于等于 0，因此，

$$\lambda_i^\star f_i(x^\star)=0,\quad i=1,\dots,m. \tag{7-10}$$

该等式被称为互补松弛条件，表示 $f_i(x^*)$ λ_i^* 中至少有一个为 0。此外，根据(7-8)

还可以得到 $f_0(x^*)$ 是 $f_0(x)+\sum_{i=1}^{m}\lambda_i^* f_i(x)+\sum_{i=1}^{p}\nu_i^* h_i(x)$ 的最小值， x^* 为其中的一个最优解。综上分析，可以得到对偶解与原问题的解等价的充分必要条件（KKT）如下：

$$\begin{array}{rcll} f_i(x^\star) & \leq & 0, & i=1,\ldots,m \\ h_i(x^\star) & = & 0, & i=1,\ldots,p \\ \lambda_i^\star & \geq & 0, & i=1,\ldots,m \\ \lambda_i^\star f_i(x^\star) & = & 0, & i=1,\ldots,m \\ \nabla f_0(x^\star)+\sum_{i=1}^m \lambda_i^\star \nabla f_i(x^\star)+\sum_{i=1}^p \nu_i^\star \nabla h_i(x^\star) & = & 0, & \end{array} \tag{7-11}$$

2. 由支持向量机的原始优化目标推导得出拉格朗日对偶优化，即给出书中公式（7.7）到公式（7.9）的推导过程。

答：已知对偶函数一定小于或等于原优化问题的最优值，因此，在乘子变量非负的约束下，最大化对偶函数的值（即求解对偶优化问题）就很有意义，它可能会达到原优化问题的最优值，特别是当满足 KKT 条件时，它等于原优化问题的最优值。此时，原问题可以转换为对偶问题进行求解，此时拉格朗日函数与原问题的解以及对偶问题的解满足强最大-最小性质，或者叫做鞍点性质。下面描述强最大-最小性质，并给出支持向量机的拉格朗日对偶优化表示。

首先，使用拉格朗日函数 L 来描述原问题的最优解 p^*。这里考虑只包含不等式约束的情形，包含等式约束时类似。拉格朗日函数关于对偶变量的上界可以表示为满足条件的原问题或者是无穷大，形式化表示如下：

$$\begin{array}{rcl} \sup_{\lambda\succeq 0} L(x,\lambda) & = & \sup_{\lambda\succeq 0}\left(f_0(x)+\sum_{i=1}^{m}\lambda_i f_i(x)\right) \\ & = & \begin{cases} f_0(x) & f_i(x)\leq 0, \quad i=1,\ldots,m \\ \infty & \text{otherwise.} \end{cases} \end{array} \tag{7-12}$$

这意味着拉格朗日函数的上界的最小值为原问题的最优值，即

$$p^\star = \inf_x \sup_{\lambda\succeq 0} L(x,\lambda). \tag{7-13}$$

其次，根据前文中对偶问题的最优解的定义，可以得到

$$d^\star = \sup_{\lambda\succeq 0} \inf_x L(x,\lambda). \tag{7-14}$$

并且两个问题的最优解的满足不等式关系 $d^* \leq p^*$，也被称为弱对偶性，具体表示如下：

$$\sup_{\lambda\succeq 0} \inf_x L(x,\lambda) \leq \inf_x \sup_{\lambda\succeq 0} L(x,\lambda), \tag{7-15}$$

最后，若想保证对偶问题的解与原问题的解相同，需满足等式关系 $d^* = p^*$，

也被称为强对偶性，具体表示为

$$\sup_{\lambda\succeq 0}\ \inf_{x} L(x,\lambda) = \inf_{x}\ \sup_{\lambda\succeq 0} L(x,\lambda). \tag{7-16}$$

支持向量机的原优化问题表示为，

$$\begin{aligned} &\min_{w,b} && \frac{1}{2}\|\mathbf{w}\|^2 \\ &s.t. && y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \,(i = 1, 2, \ldots, N). \end{aligned} \tag{7-17}$$

引入拉格朗日乘子向量 $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_N]$，我们便可以通过拉格朗日函数将约束条件融入到目标函数中，得到优化问题(7.7)对应的拉格朗日函数为

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i (y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1), \tag{7-18}$$

其中各乘子变量 $\alpha_i (i = 1, 2, ..., N)$ 均为非负值。根据公式(7-13)，原问题表示为

$$\min_{\mathbf{w},b} \max_{\alpha_i \geq 0} L(\mathbf{w}, b, \boldsymbol{\alpha}). \tag{7-19}$$

根据公式(7-16)，对应的拉格朗日对偶优化问题表示为：

$$\max_{\alpha_i \geq 0} \min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}). \tag{7-20}$$

3. 支持向量机是否可以处理多类分类，如何实现？

答：支持向量机可以处理多类分类，进行多类（例如 K 类）分类的思想是学习多个二类分类器，常用的方法有三种：

（1）分别学习 K 个二类分类器，第 k 个分类器用于判别样本是否属于第 k 类。该类方法也被称为一对多策略。训练时需构建对应的训练数据，例如，在训练第 k 个分类器时，使用属于第 k 类的数据作为正样本，不属于第 k 类的所有数据作为负样本。训练完成后，根据 K 个分类器的预测结果来判断新的样本应该属于哪一个类别，假设第 k 个分类器的预测函数为 $f_k(\mathbf{x}) = \sum_{i=1}^{N} y_i a_i \mathbf{x}^\top \mathbf{x}_i + b$, 或者 $f_k(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x}) + b$，那么最终的决策类别为，$h(\mathbf{x}) = \arg\max_k f_k(\mathbf{x})$。一对多策略的优点是最终的决策比较明确，相比于一对一策略所需训练的分类器较少；缺点是训练数据不平衡，且决策时使用的预测函数值的可能尺度不同，可能导致决策不准确。

（2）分别学习 $K(K-1)$ 个二类分类器，每个分类器用于判别该样本是属于成对类别（第 i 类和第 j 类，$i = 1, 2, ..., K, j = 1, 2, ..., K, i \neq j$ ）中的哪一类。该类方

法也被称为一对一策略。训练时需构建对应的训练数据，例如，在训练$i-j$分类器时，使用第i类的数据作为正样本，第j类的数据作为负样本。训练完成后，根据所有分类器对测试数据进行投票决策，例如，如果$i-j$分类器将测试数据判别为i，那么第 i 类投票数加 1 而第j类投票数减 1，最终决策为所有分类器投票数目总和最多的类别。这里需要注意的是使用投票决策时会产生平票的情况，此时可以随机选取其中某一类别，或者选择预测函数值最大的类别。一对一策略的优点是训练数据平衡；缺点是需要训练较多分类器，且容易出现票数相等的情况。

（3）设计新的优化目标，使用所有训练数据同时训练K个分类器，优化目标为最大化每一个类别到其他类别的间隔，带有松弛变量的多类 SVM 优化问题表示如下：

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|^2 + C\sum_{i=1}^{N}\sum_{k\neq y_i}\xi_i^k \\ s.t. \quad & \mathbf{w}_{y_i}^{\top}\mathbf{x}_i + b_{y_i} - (\mathbf{w}_k^{\top}\mathbf{x}_i + b_k) \geq 2 - \xi_i^k, \\ & \xi_i \geq 0, \ (i=1,2,\ldots,N), \ k \in \{1,2,\ldots,K\} \setminus y_i. \end{aligned} \tag{7-21}$$

该多类 SVM 的优点是不需要划分训练数据，缺点是需要在 N 个训练样本上解决一个复杂度$O(K^2N^2)$（较高）的优化问题，因此训练速度较慢。

4. 编程实现基于核方法的支持向量机，并在 UCI 数据集上测试性能。

答：略

第8章 人工神经网络与深度学习

思考与计算

1. 思考如果将线性函数 $f(\mathbf{x})=\mathbf{w}^{\top}\mathbf{x}$ 用作神经网络激活函数会有什么问题。

答：激活函数的主要目的一是实现数据的非线性变换，解决线性模型表达、分类能力不足的缺陷；二是实现数据的归一化，将当前数据映射到某个范围内，以限制数据在尺度上的扩张，防止溢出风险。如果将线性函数 $f(\mathbf{x})=\mathbf{w}^{\top}\mathbf{x}$ 用作神经网络激活函数，则多层网络可以通过矩阵变换，直接转换成一层神经网络，失去了多层的意义。所以在神经网络的隐含层通常都使用非线性激活函数，使网络具有非线性表达的能力，只有在输出层进行回归任务时可能使用线性激活函数。

2. 感知机神经网络的优缺点是什么？

感知机一种线性二分类算法，优点是能够实现在线学习，训练过程简单高效。缺点是它要求数据线性可分，难以实现非线性分类问题，比如异或、同或运算。

3. 如何保证神经网络具有较好的泛化能力？

答：神经网络的泛化能力通常指在训练数据以外的测试数据上的性能。这里讲的泛化能力是指保证了训练性能的基础上的测试性能，通常是为了解决过拟合问题。提升测试性能的常用方法包括如下几类：1、设计更合适的网络结构，如卷积神经网络和循环神经网络；2、使用与训练数据规模匹配的神经网络架构，简化或复杂化网络结构；3、早停止（early stopping）；4、权重衰减（weight decay）；5、丢弃法（dropout）；6、大多数模型都适用的数据增强方法，比如数据增加噪声、调整数据分布。

4. 编程实现神经网络的误差反向传播算法，尝试动态调整学习率以提升收敛速度，并在两个公开数据集上进行实验。

答：略

5. 编程实现卷积神经网络，并在手写字符识别数据集 MNIST 上进行实验。

答：略

6. 编程实现循环神经网络，运用不同的结构（如 LSTM 和 GRU），在公开数据集上进行机器翻译实验。

答：略

第9章 高斯过程

思考与计算

1. 思考高斯分布与高斯过程的区别。

答：高斯分布用于刻画观测数据 x 的分布，考虑更一般的情况，当 x 是多维随机变量时，即 $\mathbf{x}=x_1,x_2,...,x_D$，其分布的均值和协方差矩阵是固定的参数。使用多个样本点可以估计分布的参数，通常假设 $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N$ 是服从 D 维高斯分布的独立样本，其均值和协方差矩阵可以通过最大似然估计获得。高斯过程用于刻画成对观测数据 $\{\mathbf{x},y\}$ 中 $\mathbf{x}$ 与 y 之间映射的分布，其分布的参数是均值函数和协方差函数（也叫核函数），均值函数和协方差函数的参数称为超参数。使用多个样本点可以估计高斯过程的超参数。假设 $\{\mathbf{x}_1,y_1\},\{\mathbf{x}_2,y_2\},...,\{\mathbf{x}_N,y_N\}$ 均是服从高斯过程的样本，高斯过程假设样本之间不独立，而是假设多个输出 $y_1,y_2,...,y_N$ 的联合分布是一个多元高斯分布，维度是样本大小 N，多元分布的均值向量与协方差矩阵由输入、均值函数与核函数确定。

2. 参考图 9-1 自己构建一个高斯过程模型，从模型中产生样本，并绘制样本图像。

答：图 9-1 展示了高斯过程 $\mathcal{GP}(0,k(x,x'))$ 的三个样本，其中 $k(x,x')=\exp\{-(x-x')^2/4\}$。

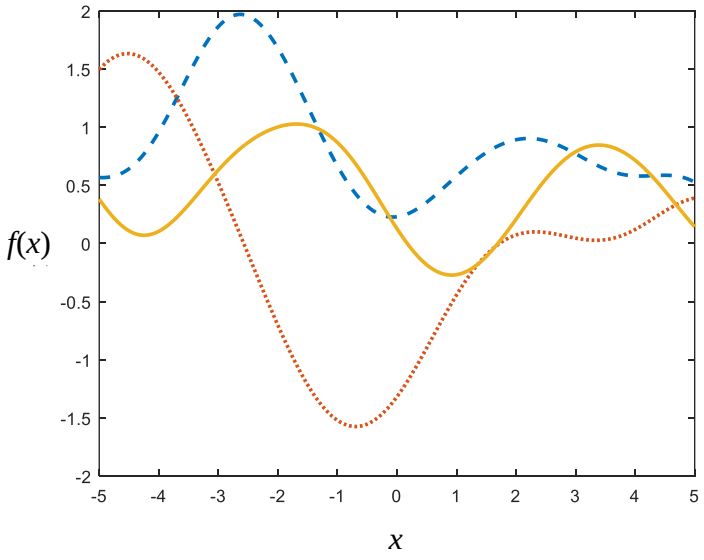


图 9-1 高斯过程的三个样本示例
（三条不同的曲线对应高斯过程的三个样本）

构建过程如下：

首先，在输入空间上均匀采样以获得输入$\{x_1, x_2, ..., x_N\}$。在该例子中输入为[-5,5]上的实数，假设采样 100 个数据输入$\{x_1, x_2, ..., x_N\}$。

其次，采样获得对应的 100 个数据输出$\{y_1, y_2, ..., y_N\}$。由于输出服从多元高斯分布，其均值向量长度为 100，协方差矩阵大小为100×100。假设均值向量为零向量$\mathbf{0}$，利用输入和核函数$k(x, x') = \exp\{-(x-x')^2/4\}$，依次计算协方差矩阵$K$中的每一个元素。那么$\{y_1, y_2, ..., y_N\}$服从多元高斯分布$\mathcal{N}(\mathbf{0}, K)$，从该分布中采样一个样本即可以得到一组$\{y_1, y_2, ..., y_N\}$，与输入$\{x_1, x_2, ..., x_N\}$对应则可以绘制出图 1 中的一条曲线。

3. 思考高斯过程与用于回归的其他概率模型相比有哪些特殊性质。

答：

1) 高斯过程是概率模型，建模了数据的不确定性，假设输出是带噪声的高斯分布，在输出有噪声的数据上性能鲁棒。
2) 高斯过程通过贝叶斯模型选择的方法来对模型中的超参数进行估计，这与支持向量机使用交叉验证选参数是不同的。高斯过程中的贝叶斯学习方法可以自动平衡模型拟合程度与模型复杂度之间的关系，从而保证一定的泛化能力。
3) 高斯过程预测结果具有概率意义，提供预测均值、方差和概率密度，其中均值表示最可能的预测值，方差和概率密度用于衡量预测的信心程度。这与神经网络回归模型只输出预测值是不同的。
4) 高斯过程是一种核方法，不需要显示地构建非线性映射，实现非线性回归。这与贝叶斯线性回归模型不同，其核函数的思想与核 SVM 类似。
5) 高斯过程不假设数据独立，假设所有数据相互依赖，在进行预测时会充分运用训练数据。这与神经网络假设样本之间相互独立是不同的。
6) 高斯过程的训练时间复杂度是$O(N^3)$，预测时间复杂度是$O(N^2)$。这比很多非核方法模型的复杂度都要高，目前有许多研究能够降低复杂度。

4. 编程实现高斯过程回归模型，并用回归模型实现二类分类问题。

答：略

第10章　聚类

思考与计算

1.　尝试其他方法证明随着聚类数目的增加，K-均值聚类的总误差逐渐减小。

答：略。

可补充 1 题：“尝试证明批处理K-均值聚类算法的收敛性”。答：K-均值聚类算法可以看作一种 EM 算法，利用 EM 算法的收敛性可以证明K-均值聚类算法的收敛性。

2.　分析K-均值聚类算法与基于 EM 算法的高斯混合模型聚类的关系。

答：这里分析批处理的K-均值算法与高斯混合模型的关系。批处理的K-均值聚类算法可以看作高斯混合模型的一个特殊版本。高斯混合模型和K-均值聚类的结果受初始值影响，均无法保证得到全局最优解；两者的聚类数目都要事先设置。具体的关系分析如下。

K-均值聚类算法属于硬聚类，迭代过程中，样本确定性地属于某一个簇，计算依据如下：

$$z_n \leftarrow \arg\min_k \| \mathbf{x}_n - \boldsymbol{\mu}_k \|^2 . \tag{10-1}$$

在得到当前的划分之后，每个聚类簇仅利用簇内的样本计算每个簇的中心，表达式如下：

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} I(z_n = k)\mathbf{x}_n, \tag{10-2}$$

其中，$N_k = \sum_{n=1}^{N} I(z_n = k).$

高斯混合模型属于软聚类，一个样本以一定的概率属于某一个簇，所有簇都会计算，计算公式如下：

$$p(z_{nk} = 1|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \Sigma_j)}. \tag{10-3}$$

其中 $\pi_k = \frac{1}{N} \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{x}_n)$。在得到每个样本对所有簇的所属概率后，利用所有

样本点计算每个簇的分布的均值和协方差。

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{x}_n)}, \tag{10-4}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{x}_n)(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\top}}{\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{x}_n)}. \tag{10-5}$$

当概率 $p(z_{nk}=1|\mathbf{x}_n)$ 取值非 0 即 1 时，即对 $p(z_{nk}=1|\mathbf{x}_n)$ 进行如下近似处理

$$p(z_{nk}=1|\mathbf{x}_n) = \begin{cases} 1, & k = \arg\max\limits_{k} p(z_{nk}=1|\mathbf{x}_n), \\ 0, & \text{otherwise}, \end{cases} \tag{10-6}$$

并且固定 $\pi_k = 1/K$ 时，高斯混合模型近似于基于马氏距离的 K-均值聚类，其中马氏距离考虑数据的协方差，马氏距离的平方的表示为

$$d = (\mathbf{x}_n-\boldsymbol{\mu}_k)^{\top}\Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k) \tag{10-7}$$

这里之所以是近似关系，原因是公式(10-3)中包含的高斯密度函数不仅仅依赖于马氏距离。如果再进一步约束所有类别的协方差矩阵为单位阵，那么此时的高斯混合模型等价于 K-均值聚类算法。

3. 推导谱聚类中与归一化的拉普拉斯矩阵相关的优化目标，即给出公式(10.23)到公式(10.31)的推导过程。

答：归一化切割（normalized cut）的表示是

$$\arg\min_{H} \frac{1}{2}\sum_{k=1}^{K}\frac{W(A_k, \mathcal{G}\setminus A_k)}{vol(A_k)},$$

$$h_{i,k} = \begin{cases} \dfrac{1}{\sqrt{vol(A_k)}} & \text{if } v_i \in A_k, \\ 0 & \text{otherwise}, \end{cases} \tag{10-8}$$

其中 $vol(A_i)$ 表示集合 A_i 中所有边的权重的和，即 $vol(A_i)=\sum_{j\in A_i} d_{jj}$ 。

下面推导出最优切割目标(10-8)关于拉普拉斯矩阵的表示以及优化问题的解。由于公式(10-8)中优化目标的每一项可以写为

$$
\begin{aligned}
\frac{W(A_k,\mathcal{G}\setminus A_k)}{vol(A_k)} &= \frac{1}{2}\Big(\sum_{i\in A_k, j\notin A_k} w_{ij}\frac{1}{vol(A_k)} + \sum_{i\notin A_k, j\in A_k} w_{ij}\frac{1}{vol(A_k)}\Big) \\
&= \frac{1}{2}\left(\sum_{i\in A_k, j\notin A_k} w_{ij}\Big(\frac{1}{\sqrt{vol(A_k)}}-0\Big)^2 + \sum_{i\notin A_k, j\in A_k} w_{ij}\Big(0-\frac{1}{\sqrt{vol(A_k)}}\Big)^2\right) \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(h_{ik}-h_{jk})^2 \\
&= \mathbf{h}_k^{\top} L \mathbf{h}_k .
\end{aligned}
\tag{10-9}
$$

因此，优化问题(10-8)可以表示为

$$
\begin{aligned}
&\arg\min_{H} \mathrm{Tr}(H^{\top}LH), \\
&h_{i,k} = \begin{cases} \dfrac{1}{\sqrt{vol(A_k)}} & \text{if } v_i \in A_k, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
\tag{10-10}
$$

如果将 H 约束为指示矩阵进行优化，那么可能的解有 $2^N K$ 种，难以遍历求解。为了解决这个问题，可以对原优化目标进行近似求解，找到满足优化目标 $\arg\min_H \mathrm{Tr}(H^{\top}LH)$ 的矩阵 H，但不约束 H 为指示矩阵。根据(10-10)中 $h_{i,k}$ 的定义，可得到 $H^{\top}DH=\mathbf{I}$。不约束 H 为指示矩阵，但仍约束 $H^{\top}DH=\mathbf{I}$，归一化切割的优化问题(10-10)可以近似表示为

$$
\begin{aligned}
&\arg\min_{F} \mathrm{Tr}(F^{\top}D^{-1/2}LD^{-1/2}F), \\
&s.t.\ F^{\top}F=\mathbf{I},
\end{aligned}
\tag{10-11}
$$

其中 $F=D^{1/2}H$。对优化问题(10-11)引入 K 个拉格朗日乘子 $\{\lambda_1,\lambda_2,...,\lambda_K\}$，记为向量 $\boldsymbol{\lambda}$，可以得到与公式(10-11)等价的无约束优化问题，表示如下：

$$
\arg\min_{H} \mathrm{Tr}(F^{\top}D^{-1/2}LD^{-1/2}F)+\mathrm{Tr}\Big[\mathrm{diag}(\boldsymbol{\lambda})(F^{\top}F-\mathbf{I})\Big].
\tag{10-12}
$$

对公式(10-12)中的优化目标关于 F 求导并设置为零，可以得到最优解 H 满足

$$
L_N F=\mathrm{diag}(\boldsymbol{\lambda})F,
\tag{10-13}
$$

其中 $L_N=D^{-1/2}LD^{-1/2}$ ，且对应的目标值为

$$
\mathrm{Tr}(F^{\top}L_N F)=\sum_{k=1}^{K}\lambda_k .
\tag{10-14}
$$

因此 F 的解为为归一化的拉普拉斯矩阵 $L_N=D^{-1/2}LD^{-1/2}$ 的前 K 个最小特征值对

应的特征向量构成的矩阵。在得到 F 之后，使用变换 $H = D^{-1/2}F$ 得到矩阵 H 并对 H 按照行归一化（此过程可以合并为一步，即直接对 F 按照行归一化得到 H ），可得到新的数据表示。为了得到图的最优切割，即数据的聚类结果，可以对新的数据表示 H 进行一次聚类，比如使用 K -均值聚类。

4. 使用一种谱聚类算法实现对人工合成的环状数据的聚类（如图 10-1）。

答：略

5. 编程实现 K -均值聚类算法和高斯混合模型聚类算法，并选择 3 个不同数据集进行实验分析。

答：略

6. 选择一种聚类算法实现图像分割。

答：略

第11章　主成分分析与相关的谱方法

思考与计算

1.　给出图 11-1 中的示例数据上 PCA 的求解过程。

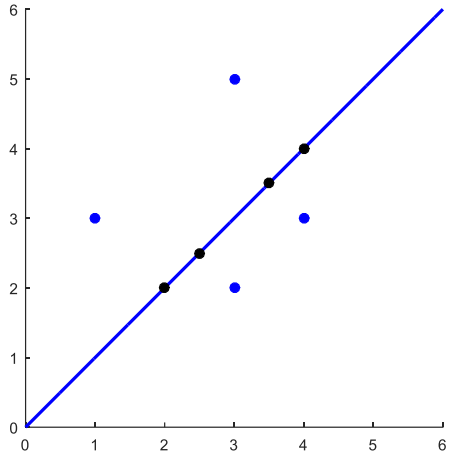


图 11-1　使用 PCA 将数据投影到一维空间的示例

(图中直线上的点是使用 PCA 将数据{(1,3),(3,2),(3,5),(4,3)}投影到一维的结果。)

首先，根据如下公式计算数据在原始空间的协方差矩阵 S：

$$\begin{aligned} S &= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i)(\mathbf{x}_i - \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i)^{\top} \\ &= \begin{bmatrix} 1.4286 & 1.4286 \\ 1.4286 & 1.4286 \end{bmatrix}. \end{aligned} \tag{11-1}$$

其次，对协方差矩阵进行特征值分解，即

$$S\mathbf{u}_1 = \lambda_1\mathbf{u}_1. \tag{11-2}$$

得到最大的特征值为 $\lambda_1 = 2.8571$，对应的特征向量为

$$\mathbf{u}_1 = [0.7071, 0.7071]^{\top}. \tag{11-3}$$

该向量为 PCA 意义下的最优投影向量，对原始数据进行投影可以得到

$$\mathbf{z} = X\mathbf{u}_1^{\top} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \\ 3 & 5 \\ 4 & 3 \end{bmatrix} \times [0.7071, 0.7071]^{\top} = [2.8284, 3.5355, 5.6569, 4.9497]^{\top} \tag{11-4}$$

在原空间中的位置为

$$\mathbf{z}' = \mathbf{z} \otimes \mathbf{u}_1 = [2.8284, 3.5355, 5.6569, 4.9497]^\top \otimes [0.7071, 0.7071] = \begin{bmatrix} 2.0 & 2.0 \\ 2.5 & 2.5 \\ 4.0 & 4.0 \\ 3.5 & 3.5 \end{bmatrix}^\top \tag{11-5}$$

2. 思考概率 PCA 的解与 PCA 的解的区别。

答： PCA 的最优投影 U 是协方差矩阵 S 的 M 个最大特征值 $\lambda_1, \lambda_2, ..., \lambda_M$ 对应的特征向量 $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M$ 构成的矩阵 U，投影后的数据表示为

$$\mathbf{z} = U^\top \mathbf{x} \text{ 或者 } \mathbf{z} = U^\top \left(\mathbf{x} - \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \right) \tag{11-6}$$

概率 PCA 的最优潜变量 $\mathbf{z}$ 的后验概率分布为

$$p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mathbf{z} \mid M^{-1} W^\top (\mathbf{x} - \boldsymbol{\mu}), \sigma^2 M^{-1}), \tag{11-7}$$

其中 $M = W^\top W + \sigma^2 \mathbf{I}$。模型参数的最大似然解为

$$W_{\mathrm{m}\ell} = U_M (\mathrm{diag}(\boldsymbol{\lambda}) - \sigma_{\mathrm{m}\ell}^2 \mathbf{I})^{1/2} R, \tag{11-8}$$

$$\sigma_{\mathrm{m}\ell}^2 = \frac{1}{D - M} \sum_{m=M+1}^{D} \lambda_m, \tag{11-9}$$

$$\boldsymbol{\mu}_{\mathrm{m}\ell} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n, \tag{11-10}$$

其中 R 通常设置为单位阵。通常使用后验概率分布的均值作为观测数据的主成分表示，即 $\mathbf{z} = M^{-1} W^\top (\mathbf{x} - \boldsymbol{\mu})$。当 $\sigma_{\mathrm{m}\ell}^2 \to 0$ 时，可以得到

$$\begin{aligned} \mathbf{z} &= M^{-1} W^\top (\mathbf{x} - \boldsymbol{\mu}) \\ &= (W^\top W)^{-1} W^\top (\mathbf{x} - \boldsymbol{\mu}) \\ &= (\mathrm{diag}(\boldsymbol{\lambda})^{1/2} U_M^\top U_M \mathrm{diag}(\boldsymbol{\lambda})^{1/2})^{-1} \mathrm{diag}(\boldsymbol{\lambda})^{1/2} U_M^\top (\mathbf{x} - \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n) \\ &= \mathrm{diag}(\boldsymbol{\lambda})^{-1/2} U_M^\top (\mathbf{x} - \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n) \end{aligned} \tag{11-11}$$

分析 PCA 与在 $\sigma_{\mathrm{m}\ell}^2 \to 0$ 时概率 PCA 的结果，如公式(11-6)和(11-11)，可以发现此时概率 PCA 得到的投影方向与 PCA 的投影方向一致，但是概率 PCA 在各

个主成分方向上的数值会比 PCA 放缩 $\boldsymbol{\lambda}_m^{-1/2}$ 倍。当 $\sigma_{\mathrm{m}\ell}^2>0$ 时，二者投影方向不同。

3. 思考在 LDA 中，为什么 S_B 的秩最大为 $C-1$。

答：S_B 的表示为

$$S_B=\sum_{c=1}^{C}N_c(\mathbf{m}_c-\mathbf{m})(\mathbf{m}_c-\mathbf{m})^{\top},$$

(11-12)

其中 $\mathbf{m}_c-\mathbf{m}$ 列向量，其秩为 1，即 $r(\mathbf{m}_c-\mathbf{m})=1$。根据秩的乘法不等式性质 $r(AB)\le\min(r(A),r(B))$，可以得到

$$r\left((\mathbf{m}_c-\mathbf{m})(\mathbf{m}_c-\mathbf{m})^{\top}\right)\le 1 \tag{11-13}$$

根据秩的加法不等式性质 $r(A+B)\le r(A)+r(B)$，可以得到

$$r\left(\sum_{c=1}^{C}N_c(\mathbf{m}_c-\mathbf{m})(\mathbf{m}_c-\mathbf{m})^{\top}\right)\le C \tag{11-14}$$

由于 $\mathbf{m}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i=\frac{1}{N}\sum_{c=1}^{C}N_c\mathbf{m}_c$，因此 $\sum_{c=1}^{C}N_c(\mathbf{m}_c-\mathbf{m})(\mathbf{m}_c-\mathbf{m})^{\top}$ 中的任意一项 $N_c(\mathbf{m}_c-\mathbf{m})(\mathbf{m}_c-\mathbf{m})^{\top}$ 可以表示为其他项的线性组合，继而可以得到

$$r\left(\sum_{c=1}^{C}N_c(\mathbf{m}_c-\mathbf{m})(\mathbf{m}_c-\mathbf{m})^{\top}\right)\le C-1 \tag{11-15}$$

4. 思考求解 CCA 中公式(11.71)所示的广义特征值问题的计算复杂度。

答：CCA 中公式(11.71)所示的广义特征值问题是求解 $\mathbf{w}_x$，使其满足如下等式

$$C_{xy}C_{yy}^{-1}C_{xy}^{\top}\mathbf{w}_x=4\lambda^2C_{xx}\mathbf{w}_x. \tag{11-16}$$

该优化问题等价于求解公式形如 $A\hat{\mathbf{x}}=\hat{\lambda}B\hat{\mathbf{x}}$ 的广义特征值问题。广义特征值问题可以通过一些变换转化为普通特征值问题。

下面分别分析两种常用方法的复杂度。假设 $\mathbf{x}$ 的维度为 D_x，$\mathbf{y}$ 的维度为 D_y，样本个数为 N。

（1）针对广义特征值问题 $A\hat{\mathbf{x}}=\hat{\lambda}B\hat{\mathbf{x}}$，如果 B 矩阵可逆，对等式两边左乘 B^{-1} 可以得到 $B^{-1}A\hat{\mathbf{x}}=\hat{\lambda}\hat{\mathbf{x}}$。该方法的复杂度包括计算协方差矩阵 C_{xx},C_{xy},C_{yy} 的复杂度分别为 $O(ND_x^2)$，$O(ND_xD_y)$，$O(ND_y^2)$，矩阵 C_{xx} 求逆的复杂度 $O(D_x^3)$，计算 C_{yy}^{-1} 的复杂度 $O(D_y^3)$，矩阵乘法 $C_{xx}^{-1}C_{xy}C_{yy}^{-1}C_{xy}^{\top}$ 的复杂度 $O(\max\{D_x^2D_y,D_xD_y^2\})$，特征值

分解的复杂度$O(D_x^3)$，因此总复杂度为$O(\max\{ND_x^2+D_x^3, ND_y^2+D_y^3\})$，这里分开表示计算协方差矩阵与特征值分解的复杂度。；

（2）如果B矩阵可进行 Cholesky 分解，即$B=LL^\top$，其中L是下三角阵，通过引入新的向量$\hat{\mathbf{y}}=L^\top\hat{\mathbf{x}}$，可以得到等价的特征值问题$L^{-1}A(L^{-1})^\top\hat{\mathbf{y}}=\lambda\hat{\mathbf{y}}$，其中$\hat{\mathbf{x}}=(L^{-1})^\top\hat{\mathbf{y}}$。该方法的复杂度包括计算协方差矩阵$C_{xx},C_{xy},C_{yy}$的复杂度分别为$O(ND_x^2)$，$O(ND_xD_y)$，$O(ND_y^2)$，对$C_{xx}$进行 Cholesky 分解的复杂度$O(D_x^3)$，计算$L^{-1}$的复杂度$O(D_x^3)$，矩阵乘法$L^{-1}C_{xy}C_{yy}^{-1}C_{xy}^\top(L^{-1})^\top$的复杂度$O(\max\{D_x^2D_y, D_xD_y^2, D_x^3\})$，特征值分解的复杂度 O($D_x^3$)，因此总复杂度为$O(\max\{ND_x^2+D_x^3, ND_y^2+D_y^3\})$，这里分开表示计算协方差矩阵与特征值分解的复杂度。

5. 试推导核 PCA 在$\{\phi(\mathbf{x}_n)\}_{n=1}^N$没有中心化情况下的投影表示。

答：$\{\phi(\mathbf{x}_n)\}_{n=1}^N$没有中心化，表示变换后的数据均值不为零，即$\sum_{n=1}^N\phi(\mathbf{x}_n)\neq 0$，令$\tilde{\phi}(\mathbf{x}_n)=\phi(\mathbf{x}_n)-\frac{1}{N}\sum_{n=1}^N\phi(\mathbf{x}_n)$表示数据从原始空间到高维空间的中心化后的非线性映射，如果投影到M维子空间，核 PCA 寻找投影向量$\{\mathbf{v}_m\}_{m=1}^M$，使得映射后的数据$\tilde{\phi}(X)=[\tilde{\phi}(\mathbf{x}_1),\tilde{\phi}(\mathbf{x}_2),...,\tilde{\phi}(\mathbf{x}_N)]$经投影后在子空间各个维度的表示$\{\mathbf{v}_m^\top\tilde{\phi}(X)\}_{m=1}^M$的方差总和最大。那么在高维空间上的协方差矩阵为

$$S'=\frac{1}{N}\sum_{n=1}^N\tilde{\phi}(\mathbf{x}_n)\tilde{\phi}(\mathbf{x}_n)^\top \tag{11-17}$$

可以得到其主成分子空间的基向量$\mathbf{v}_m$满足如下公式

$$S'\mathbf{v}_m=\lambda_m\mathbf{v}_m, \tag{11-18}$$

其中$\mathbf{v}_m$也表示投影向量，且$\mathbf{v}_m$满足

$$\frac{1}{N}\sum_{n=1}^N\tilde{\phi}(\mathbf{x}_n)\tilde{\phi}(\mathbf{x}_n)^\top\mathbf{v}_m=\lambda_m\mathbf{v}_m. \tag{11-19}$$

由于$\tilde{\phi}(\mathbf{x}_n)^\top\mathbf{v}_m$是一个标量，$\mathbf{v}_m$可以表示为$\{\tilde{\phi}(\mathbf{x}_n)\}$的线性组合，记为

$$\mathbf{v}_m=\sum_{n=1}^N\alpha_{mn}\tilde{\phi}(\mathbf{x}_n), \tag{11-20}$$

将式(11-20)带入式(11-19)中，可得

$$\frac{1}{N}\sum_{n=1}^N\tilde{\phi}(\mathbf{x}_n)\tilde{\phi}(\mathbf{x}_n)^\top\sum_{n'=1}^N\alpha_{mn'}\tilde{\phi}(\mathbf{x}_{n'})=\lambda_m\sum_{n=1}^N\alpha_{mn}\tilde{\phi}(\mathbf{x}_n). \tag{11-21}$$

为了使用核技巧，在公式(11-21)的等号两边左乘$\tilde{\boldsymbol{\phi}}(\mathbf{x}_\ell)^\top$，$\ell=1,2,...,N$，并且引入函数$\tilde{k}(\mathbf{x},\mathbf{x}')$，其计算可以用核函数$k(\mathbf{x},\mathbf{x}')=\boldsymbol{\phi}(\mathbf{x})^\top\boldsymbol{\phi}(\mathbf{x}')$表示，表达式如下：

$$\begin{aligned}\tilde{k}(\mathbf{x},\mathbf{x}') &= \tilde{\boldsymbol{\phi}}(\mathbf{x})^\top\tilde{\boldsymbol{\phi}}(\mathbf{x}') \\ &= \boldsymbol{\phi}(\mathbf{x})^\top\boldsymbol{\phi}(\mathbf{x}') - \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\phi}(\mathbf{x})^\top\boldsymbol{\phi}(\mathbf{x}_j) - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}(\mathbf{x}_i)^\top\boldsymbol{\phi}(\mathbf{x}') + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\phi}(\mathbf{x}_i)^\top\boldsymbol{\phi}(\mathbf{x}_j) \\ &= k(\mathbf{x},\mathbf{x}') - \frac{1}{N}\sum_{j=1}^{N}k(\mathbf{x},\mathbf{x}_j) - \frac{1}{N}\sum_{i=1}^{N}k(\mathbf{x}_i,\mathbf{x}') + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}k(\mathbf{x}_i,\mathbf{x}_j)\end{aligned} \tag{11-22}$$

公式(11-21)可以转换为

$$\frac{1}{N}\sum_{n=1}^{N}\tilde{k}(\mathbf{x}_\ell,\mathbf{x}_n)\sum_{n'=1}^{N}\alpha_{mn'}\tilde{k}(\mathbf{x}_n,\mathbf{x}_{n'}) = \lambda_m\sum_{n=1}^{N}\alpha_{mn}\tilde{k}(\mathbf{x}_\ell,\mathbf{x}_n),\ \ell=1,2,...,N. \tag{11-23}$$

公式(11-23)可以写为矩阵的形式：

$$\tilde{K}\tilde{K}\boldsymbol{\alpha}_m = \lambda_m N\tilde{K}\boldsymbol{\alpha}_m, \tag{11-24}$$

其非零特征值及其对应的特征向量满足$\tilde{K}\boldsymbol{\alpha}_m=\lambda_m N\boldsymbol{\alpha}_m$。因此，核 PCA 等价于求解如下特征值问题：

$$\tilde{K}\boldsymbol{\alpha}_m = \lambda_m N\boldsymbol{\alpha}_m, \tag{11-25}$$

其中$\tilde{K}$是函数$\tilde{k}(\mathbf{x},\mathbf{x}')$在所有训练数据上构建的矩阵，$\boldsymbol{\alpha}_m=[\alpha_{m1},\alpha_{m2},...,\alpha_{mN}]^\top$是需要求解的特征向量，根据所需投影的维度$M$，选择前$M$个最大特征值对应的特征向量。

在得到所需的特征向量之后，投影后的主成分表示可以使用如下计算获得：

$$z_m = \mathbf{v}_m^\top\tilde{\boldsymbol{\phi}}(\mathbf{x}) = \sum_{n=1}^{N}\alpha_{mn}\tilde{k}(\mathbf{x},\mathbf{x}_n),\ m=1,2,...,M. \tag{11-26}$$

第14章　强化学习

思考与计算

1. **列举三个可以使用强化学习框架的任务，并确定这些任务中的状态、动作、奖励函数。**

   答：略

2. **假设需要把汽车驾驶到某一目的地，尝试为这一任务设计奖励函数，并讨论这样设计的作用和意义。**

   答：强化学习目的可以归结为：最大化智能体接收到的累积收益的期望值。这意味着需要最大化的不是当前收益，而是长期的累积收益。将汽车驾驶到某一目的地可以使用强化学习来学习驾驶动作，其中状态定义为当前汽车的地理位置和所获取的各种环境参数，动作定义为可作用于汽车的操作，奖励函数可以包含多种奖励：例如，距离奖励，可定义为与目的地的距离的反函数，作用是为了获取最优路径；速度奖励，速度需保证在一定的范围内，超过范围越多扣除奖励分数越多，作用是为了防止停在原地或无限加速；违规奖励，越线的比例或开到人行道的比例越多，扣除奖励分数越多，作用是为防止越线或开到人行道；舒适度奖励，方向盘转角越大扣除奖励分数越多，作用是为了使车辆贴近人的开车行为；油耗奖励，油耗越大扣除奖励分数越多，作用是为了减少油耗。

3. **请自行设计状态空间和动作空间的大小，随机生成奖励和转移概率，用动态规划法编程实现策略迭代算法和值迭代算法。**

   答：略

4. **在未给定环境模型的情况下，可以通过随机策略采样估计出环境模型，再通过规划的方法强化学习，这种做法与无环境模型的控制方法有什么区别？**

   在未给定环境模型的规划方法中，智能体先通过随机策略采样与环境进行交互，得到对环境模型的估计。在得到估计的环境模型之后，可以直接使用基于规划的强化学习，规划过程不再需要智能体与外部环境作任何交互。而在无环境模型的控制方法中，不会建立环境模型，强化学习训练过程中需要使用最优或随机策略与环境进行大量的交互来估计值函数，最终求解最优策略实现对智能体的控制。

5. 蒙特卡洛控制采用的是动作值函数，可以采用状态值函数吗？为什么？

答：不可以。蒙特卡洛控制在无环境的情况下，通过智能体和环境进行交互收集一些经验样本，然后再根据这些样本来求解最优策略。交互过程中需要根据动作值函数来优化策略进而执行动作，而由于环境模型未知，动作值函数无法通过状态值函数的值进行计算，所以需要直接计算动作值函数。

6. 推导算法14-4中SARSA算法的更新公式。

答：SARSA 算法是一种基于时序差分学习方法的同策略学习算法，该学习方法仅使用当前时刻的状态、动作与收益以及下一时刻的状态和动作。SARSA 的主要思路是通过迭代的方式优化 Q 函数，并更新策略。

根据如下动作值函数的贝尔曼期望方程

$$Q_\pi(s,a)=\mathbb{E}_{(r,s',a')}\left[r+\gamma Q_\pi(s',a')\right], \tag{14-1}$$

可以得到，理想情况下使用迭代优化的更新过程为

$$Q_\pi(s,a)=Q_\pi(s,a)+\alpha\left(\mathbb{E}_{(r,s',a')}\left[r+\gamma Q_\pi(s',a')\right]-Q_\pi(s,a)\right) \tag{14-2}$$

其中 α 是学习率。

由于环境模型未知，无法计算期望，因此结合蒙特卡洛技术对期望进行近似。具体过程是：首先初始化 Q 函数，之后根据策略 $\pi^\varepsilon(a|s)$ 执行一个动作 a，在执行完该动作后到达新的状态 s'，得到即时收益为 r，Q 函数值为 $Q_\pi(s,a)$；再次根据策略 $\pi^\varepsilon(a|s)$ 执行动作 a'，得到的 Q 函数值为 $Q_\pi(s',a')$。蒙特卡洛技术利用 $r+\gamma Q_\pi(s',a')$ 与 $Q_\pi(s,a)$ 之间的差距来优化 Q 函数，最后更新策略。循环往复直至 Q 函数收敛，其中 $r+\gamma Q_\pi(s',a')$ 被称为时序差分目标，它与 $Q_\pi(s,a)$ 的差被称为时序差分误差，具体的更新过程如下：

$$Q_\pi(s,a)=Q_\pi(s,a)+\alpha\left(r+\gamma Q_\pi(s',a')-Q_\pi(s,a)\right) \tag{14-3}$$

7. 分析SARSA算法和 Q 学习算法的异同。

答：SARSA 算法和 Q 学习算法的都是基于单步时序差分的对动作值函数进行迭代优化的方法，迭代更新公式形式相同，过程如下：

$$Q_\pi(s,a)=Q_\pi(s,a)+\alpha\left(r+\gamma Q_\pi(s',a')-Q_\pi(s,a)\right) \tag{14-4}$$

然而其中执行动作 a 与 a' 依据的策略有所区别。SARSA 算法是同策略方法，下一步执行动作 a' 与当前步执行动作 a 依据的策略是相同的，且策略依据的动作

值函数是相同的。Q 学习算法是异策略方法，下一步执行动作 a' 与当前步执行动作 a 依据的策略是不同的，虽然不同策略可能依据相同的动作值函数。同策略依据自己的现有策略进行策略提升，异策略可以结合其他策略进行策略提升。

8. 推导强化学习的目标关于参数 θ 的梯度，并写出REINFORCE算法的详细步骤。

答：略（书中较详细）。

9. 查找资料，学习优势函数的定义、优势行动者-评论者算法（Advantage Actor-Critic，A2C）、异步优势行动者-评论者算法（Asynchronous Advantage Actor-Critic，A3C），并分析同步与异步算法的优缺点。

答：优势函数表达在某状态下，某动作相对于平均动作而言的优势，如果优势函数大于零，则说明该动作比平均动作好，如果优势函数小于零，则说明当前动作还不如平均动作好。优势函数可以表示为动作值函数相比于当前状态值函数的优势，表达式为

$$A_\pi(s,a)=Q_w(s,a)-V_\pi(s) \tag{14-5}$$

A2C 与行动者-评论者算法大致相同，区别在于 A2C 使用优势函数代替行动者的梯度更新公式 $\theta=\theta+\alpha[\nabla_\theta \log \pi_\theta(a\mid s)]Q_w(s,a)$ 中的 $Q_w(s,a)$，即更新公式为

$$\theta=\theta+\alpha[\nabla_\theta \log \pi_\theta(a\mid s)](Q_w(s,a)-V_\pi(s)) \tag{14-6}$$

有时也会使用状态值函数的时序差分误差来替代优势函数，那么策略参数的更新公式为

$$\theta=\theta+\alpha[\nabla_\theta \log \pi_\theta(a\mid s)](r+\gamma V_\pi(s')-V_\pi(s)) \tag{14-7}$$

A3C 算法是考虑深度强化学习中的神经网络在训练时要求数据独立而采取异步更新方法。A3C 中包含一个全局网络和多个局部网络，训练过程中开启多个线程。每个线程包含一个局部网络，每次训练先从全局网络获取全局参数，独立地与环境交互产生样本，独立完成训练并得到参数的梯度。每个线程的训练方法都是一个 A2C。每个线程在得到参数梯度之后，异步地上传给全局网络，完成全局参数的更新。