模式识别作业四

数据科学与计算机学院　17大数据与人工智能

17341015　陈鸿峥

问题 1 (§3 Q4)**.** 设$\mathbf{x}$为一个d维的二值向量，服从多维伯努利分布

$$P(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta)^{1-x_i}$$

其中$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^{\mathrm{T}}$为一个未知的参数向量，而$\theta_i$为$x_i = 1$的概率。证明：对于$\boldsymbol{\theta}$的最大似然估计为

$$\hat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

解答. 考虑n个样本的采样$\{\mathbf{x}_k\}_{k=1}^{n}$，似然函数为

$$L(\boldsymbol{\theta}) = \prod_{k=1}^{n}\prod_{i=1}^{d} \theta_i^{x_{ki}}(1-\theta_i)^{1-x_{ki}}$$

对数似然函数为

$$\ell(\boldsymbol{\theta}) = \sum_{k=1}^{n}\sum_{i=1}^{d} x_{ki}\left(\ln\theta_i + (1-x_{ki})\ln(1-\theta_i)\right)$$

对上式求梯度有

$$[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})]_i = \nabla_{\theta_i}\ell(\theta) = \frac{1}{\theta_i}\sum_{k=1}^{n} x_{ki} - \frac{1}{1-\theta_i}\sum_{k=1}^{n}(1-x_{ki}) = 0$$

对上式整理可得

$$\hat{\theta}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$$

表为向量形式即得结果。

问题 2 (§3 Q19)**.** 假设我们有一组训练样本，都服从高斯分布，其协方差矩阵Σ已知，而均值$\boldsymbol{\mu}$未知。进一步假设这个均值$\boldsymbol{\mu}$本身是随机取值的，服从均值为$\mathbf{m}_0$，协方差为Σ_0的高斯分布。

(a) 均值$\boldsymbol{\mu}$的*MAP*估计是什么？

(b) 假设我们用线性变换来变换坐标$\mathbf{x}' = A\mathbf{x}$，其中$A$为非奇异矩阵。那么，*MAP*能够对变换以后的$\boldsymbol{\mu}'$做出正确的估计吗？并加以解释。

解答. (a) 对于高斯分布有对数似然函数$\ell(\boldsymbol{\mu})$和概率密度函数$p(\boldsymbol{\mu})$

$$\begin{aligned}\ell(\boldsymbol{\mu}) &= \sum_{k=1}^{n} \ln\left[p\left(\mathbf{x}_k|\boldsymbol{\mu}\right)\right] \\ &= -\frac{n}{2}\ln\left[(2\pi)^d|\Sigma|\right] - \sum_{k=1}^{n}\frac{1}{2}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{x}_k-\boldsymbol{\mu}\right) \\ p(\boldsymbol{\mu}) &= \frac{1}{(2\pi)^{d/2}\left|\Sigma_0\right|^{1/2}}\exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}-\mathbf{m}_0\right)^{\mathrm{T}}\Sigma_0^{-1}\left(\boldsymbol{\mu}-\mathbf{m}_0\right)\right]\end{aligned}$$

进而MAP估计为

$$\begin{aligned}\hat{\boldsymbol{\mu}} =& \arg\max_{\boldsymbol{\mu}}\left(\ell(\boldsymbol{\mu})p(\boldsymbol{\mu})\right) \\ =& \arg\max_{\boldsymbol{\mu}}\left\{\left[-\frac{n}{2}\ln\left[(2\pi)^d|\Sigma|\right] - \sum_{k=1}^{n}\frac{1}{2}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)\right]\right. \\ &\left.\cdot\left[\frac{1}{(2\pi)^{d/2}\left|\Sigma_0\right|^{1/2}}\exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}-\mathbf{m}_0\right)^{\mathrm{T}}\Sigma_0^{-1}\left(\boldsymbol{\mu}-\mathbf{m}_0\right)\right]\right]\right\}\end{aligned}$$

MAP最好前后都取对数，但这里按照的是课本的公式，即最大似然取对数，概率不取。

(b) 由均值和协方差的性质有

$$\begin{aligned}\boldsymbol{\mu}' &= \mathbb{E}\left[\mathbf{x}'\right] = \mathbb{E}[A\mathbf{x}] = A\mathbb{E}[\mathbf{x}] = A\boldsymbol{\mu} \\ \Sigma' &= \mathbb{E}\left[\left(\mathbf{x}'-\boldsymbol{\mu}'\right)\left(\mathbf{x}'-\boldsymbol{\mu}'\right)^{\mathrm{T}}\right] \\ &= \mathbb{E}\left[\left(A\mathbf{x}'-A\boldsymbol{\mu}'\right)\left(A\mathbf{x}'-A\boldsymbol{\mu}'\right)^{\mathrm{T}}\right] \\ &= \mathbb{E}\left[A\left(\mathbf{x}'-\boldsymbol{\mu}'\right)\left(\mathbf{x}'-\boldsymbol{\mu}'\right)^{\mathrm{T}}A^{\mathrm{T}}\right] \\ &= A\mathbb{E}\left[\left(\mathbf{x}'-\boldsymbol{\mu}'\right)\left(\mathbf{x}'-\boldsymbol{\mu}'\right)^{\mathrm{T}}\right]A^{\mathrm{T}} \\ &= A\Sigma A^{\mathrm{T}}\end{aligned}$$

进而有$\boldsymbol{\mu}'$的对数似然函数

$$\begin{aligned}\ell(\boldsymbol{\mu}') &= \ln\left(\prod_{k=1}^{n}p\left(\mathbf{x}_k'|\boldsymbol{\mu}'\right)\right) \\ &= \sum_{k=1}^{n}\ln\left[p\left(A\mathbf{x}_k|A\boldsymbol{\mu}\right)\right] \\ &= -\frac{n}{2}\ln\left[(2\pi)^d\left|A\Sigma A^{\mathrm{T}}\right|\right] - \sum_{k=1}^{n}\frac{1}{2}\left(\left(\mathbf{x}_k-\boldsymbol{\mu}\right)^{\mathrm{T}}A^{\mathrm{T}}\right)\left(A\Sigma A^{\mathrm{T}}\right)^{-1}\left(A\left(\mathbf{x}_k-\boldsymbol{\mu}\right)\right) \\ &= -\frac{n}{2}\ln\left[(2\pi)^d\left|A\Sigma A^{\mathrm{T}}\right|\right] - \sum_{k=1}^{n}\frac{1}{2}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)^{\mathrm{T}}\left(A^{\mathrm{T}}\left(A^{-1}\right)^{\mathrm{T}}\right)\Sigma^{-1}\left(A^{-1}A\right)\left(\mathbf{x}_k-\boldsymbol{\mu}\right) \\ &= -\frac{n}{2}\ln\left[(2\pi)^d\left|A\Sigma A^{\mathrm{T}}\right|\right] - \sum_{k=1}^{n}\frac{1}{2}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)\end{aligned}$$

类似地可以得到$\boldsymbol{\mu}'$的高斯密度

$$\begin{aligned}
p(\boldsymbol{\mu}') &= \frac{1}{(2\pi)^{d/2}\left|\Sigma_0'\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}'-\mathbf{m}_0'\right)^{\mathrm{T}}\Sigma_0'^{-1}\left(\boldsymbol{\mu}'-\mathbf{m}_0'\right)\right] \\
&= \frac{1}{(2\pi)^{d/2}\left|\Sigma_0'\right|^{1/2}} \exp\left[-\frac{1}{2}\left(A\boldsymbol{\mu}-A\mathbf{m}_0\right)^{\mathrm{T}}\left(A\Sigma_0A^{\mathrm{T}}\right)^{-1}\left(A\boldsymbol{\mu}-A\mathbf{m}_0\right)\right] \\
&= \frac{1}{(2\pi)^{d/2}\left|\Sigma_0'\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}-\mathbf{m}_0\right)^{\mathrm{T}}A^{\mathrm{T}}\left(A^{-1}\right)^{\mathrm{T}}\Sigma_0^{-1}A^{-1}A\left(\boldsymbol{\mu}-\mathbf{m}_0\right)\right] \\
&= \frac{1}{(2\pi)^{d/2}\left|\Sigma_0'\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}-\mathbf{m}_0\right)^{\mathrm{T}}\Sigma_0^{-1}\left(\boldsymbol{\mu}-\mathbf{m}_0\right)\right]
\end{aligned}$$

新的MAP估计为

$$\begin{aligned}
\hat{\boldsymbol{\mu}}' =& \arg\max_{\boldsymbol{\mu}}\left\{\left[-\frac{n}{2}\ln\left[(2\pi)^d\left|A\Sigma A^{\mathrm{T}}\right|\right]-\sum_{k=1}^{n}\frac{1}{2}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)\right]\right. \\
&\left.\cdot\left[\frac{1}{(2\pi)^{d/2}\left|A\Sigma_0A^{\mathrm{T}}\right|^{1/2}}\exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}-\mathbf{m}_0\right)^{\mathrm{T}}\Sigma_0^{-1}\left(\boldsymbol{\mu}-\mathbf{m}_0\right)\right]\right]\right\}
\end{aligned}$$

与(a)比较知，$\hat{\boldsymbol{\mu}}$和$\hat{\boldsymbol{\mu}}'$的差异均在常数部分，因此MAP可以对变换后的$\boldsymbol{\mu}'$做出正确估计。

问题 3 (§3 Q38). 令$p_{\mathbf{x}}(\mathbf{x}\mid\omega_i), i=1,2$为任意的概率密度函数，均值为$\boldsymbol{\mu}_i$，协方差矩阵为$\Sigma_i$，其中并不要求$p_{\mathbf{x}}(\mathbf{x}\mid\omega_i)$必须为正态概率密度。令$y=\mathbf{w}^{\mathrm{T}}\mathbf{x}$表示投影，并且设投影后的结果的概率密度函数为$p(y\mid\omega_i)$，其均值为$\mu_i$，方差为$\sigma_i^2$。

(a) 证明准则函数

$$J_1(\mathbf{w})=\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}$$

当

$$\mathbf{w}=(\Sigma_1+\Sigma_2)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

时取得最大值。

(b) 如果$P(\omega_i)$为ω_i的先验概率，证明

$$J_2(\mathbf{w})=\frac{(\mu_1-\mu_2)^2}{P(\omega_1)\sigma_1^2+P(\omega_2)\sigma_2^2}$$

当

$$\mathbf{w}=[P(\omega_1)\Sigma_1+P(\omega_2)\Sigma_2]^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

时取得最大值。

(c) 在*(a)*和*(b)*之间，哪个与公式*(96)*的联系更密切，请解释为什么。

解答. (a) 对于$i=1,2$有

$$\mu_i = \frac{1}{n_i}\sum_{y\in\mathcal{Y}_i} y = \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i} \mathbf{w}^{\mathrm{T}}\mathbf{x} = \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_i$$

$$\sigma_i^2 = \sum_{y\in\mathcal{Y}_i}(y-\mu_i)^2 = \mathbf{w}^{\mathrm{T}}\left[\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\right]\mathbf{w} = \mathbf{w}^{\mathrm{T}}\Sigma_i\mathbf{w}$$

$$\Sigma_i = \sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}$$

有总类内散度矩阵S_W及总类间散度矩阵S_B

$$S_W = \Sigma_1 + \Sigma_2$$

$$S_B = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}$$

进而

$$J_1(\mathbf{w}) = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} = \frac{\mathbf{w}^{\mathrm{T}}S_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}S_W\mathbf{w}}$$

即广义瑞利商。由课本公式(106)，可以得到$J_1(\mathbf{w})$在$\mathbf{w}=S_W^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$得到最大值，即在

$$\mathbf{w} = (\Sigma_1+\Sigma_2)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

时取得最大值。

(b) 同(a)理，用广义瑞利商表示有

$$J_2(\mathbf{w}) = \frac{(\mu_1-\mu_2)^2}{P(\omega_1)\sigma_1^2+P(\omega_2)\sigma_2^2} = \frac{\mathbf{w}^{\mathrm{T}}S_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}S'_W\mathbf{w}}$$

其中$S'_W = P(\omega_1)\sigma_1^2+P(\omega_2)\sigma_2^2$。同理有$J_2(\mathbf{w})$在

$$\mathbf{w} = S_W'^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = [P(\omega_1)\Sigma_1+P(\omega_2)\Sigma_2]^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

时取得最大值。

(c) 在公式(96)中，

$$J(\mathbf{w}) = \frac{|\tilde{m}_1-\tilde{m}_2|^2}{\tilde{s}_1^2+\tilde{s}_2^2}$$

令$\tilde{m}_i=\mu_i$和$\tilde{s}_i^2=\sigma_i^2$，可以得到(a)中的式子，即(96)与(a)联系更密切。