Homework 02

最大似然参数估计、非参数技术、多维正态分布

Homework 2.1

设 x 为一个 d 维的二值向量（即其分量取值为 0 或 1），服从多维伯努利分布

$$p(\mathbf{x} \mid \Theta) = \prod_{i=1}^{d} \Theta_i^{x_i} (1-\Theta_i)^{1-x_i}$$

其中 $\Theta = (\Theta_1, \cdots, \Theta_d)^T$ 是未知的参数向量，而 Θ_i 为 x_i=1 的概率。证明，对于 Θ 的最大似然估计为

$$\hat{\Theta} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$$ ，n 为样本数

Homework 2.1

证明：假设数据集 $D=\{\vec{X_1},\vec{X_2},\cdots,\vec{X_n}\}$，n为样本数

由题可知，$\vec{X_k}$，$k=1,2,\cdots,n$ 服从多维伯努利分布.

即 $p(\vec{X_k}\mid\vec{\Theta}) = \prod_{i=1}^{d} \Theta_i^{x_{ki}} (1-\Theta_i)^{1-x_{ki}}$，其中d为$X_k$的维度d，$x_{ki}$为第k个样本的第i维

而Θ_i为 $x_{ki}=1$的概率.

则样本观测值的概率为：

$$p(D\mid\vec{\Theta}) = \prod_{k=1}^{n} p(\vec{X_k}\mid\vec{\Theta})$$

那么$\vec{\Theta}$的极大似然估计即求使得$p(D\mid\vec{\Theta})$最大的$\vec{\Theta}$.

对$p(D\mid\vec{\Theta})$求对数得：

$$\ln p(D\mid\vec{\Theta}) = \sum_{k=1}^{n} \ln p(\vec{X_k}\mid\vec{\Theta})$$

∵对数函数是单调的，∴求对数后极值保持一致. 那么我们求得$p(D\mid\vec{\Theta})$最大的$\vec{\Theta}$即求$\ln p(D\mid\vec{\Theta})$的极值.

因此，对 $\ln p(D|\vec{\theta})$ 求导可得：

$$\frac{\partial \ln p(D|\vec{\theta})}{\partial \vec{\theta}} = \frac{\partial \sum_{k=1}^{n} \ln p(\vec{X}_k|\vec{\theta})}{\partial \vec{\theta}}$$

$\because p(\vec{X}_k|\vec{\theta}) = \prod_{i=1}^{d} \theta_i^{x_{ki}} (1-\theta_i)^{1-x_{ki}}$

$$\therefore \frac{\partial \ln p(D|\vec{\theta})}{\partial \vec{\theta}} = \frac{\partial \sum_{k=1}^{n}\sum_{i=1}^{d} x_{ki}\ln\theta_i + (1-x_{ki})\ln(1-\theta_i)}{\partial \vec{\theta}}$$

$$= \sum_{i=1}^{d}\left[\frac{\partial \sum_{k=1}^{n} x_{ki}\ln\theta_i + (1-x_{ki})\ln(1-\theta_i)}{\partial \theta_i}\right]$$

$$= \sum_{i=1}^{d}\sum_{k=1}^{n}\left(\frac{x_{ki}}{\theta_i} - \frac{1-x_{ki}}{1-\theta_i}\right)$$

$$= \sum_{i=1}^{d}\sum_{k=1}^{n} \frac{x_{ki} - x_{ki}\theta_i - \theta_i + x_{ki}\theta_i}{\theta_i(1-\theta_i)}$$

$$= \sum_{i=1}^{d}\sum_{k=1}^{n} \frac{x_{ki} - \theta_i}{\theta_i(1-\theta_i)}$$

$$= \sum_{k=1}^{n} \frac{\vec{X}_k - \vec{\theta}}{\vec{\theta}(1-\vec{\theta})}$$

令导数为零可得：$\sum_{k=1}^{n}(\vec{X}_k - \vec{\theta}) = 0$

$$\therefore \sum_{k=1}^{n}\vec{X}_k - n\vec{\theta} = 0$$

$$\Rightarrow \vec{\theta} = \frac{1}{n}\sum_{k=1}^{n}\vec{X}_k$$

因此得证，$\vec{\theta}$ 的最大似然估计为 $\vec{\theta} = \frac{1}{n}\sum_{k=1}^{n}\vec{X}_k$，n为样本数

Homework 2.2

令 $D=\{x_1,\cdots,x_n\}$ 为 n 个独立的已标记的集合。令 $D_k(x)=\{x_1^{'},\cdots,x_k^{'}\}$ 为样本 x 的 k 个最邻近。根据 k-近邻规则，x 将归入 $D_k(x)$ 中出现次数最多的那个类别。考虑一个 2 类别问题，先验概率为 $P(\omega_1)=P(\omega_2)=1/2$ 。进一步假设类条件概率密度 $P(X|\omega_i)$ 在 10 单位超球体内为均匀分布。

（a）证明如果 k 为奇数，那么平均误差率为

$$P_n(e)=\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j}$$

Homework 2.2

(a) 证明：设样本 x 的类别为 w_1，那么 x 若在 k-近邻中被归入 w_2 则出现误差。

而根据 k-近邻规则，x 将归入 $D_k(x)$ 中出现次数最多的那个类别。

假设 $D_k(x)$ 中出现 j 个 w_1 类别的样本，当样本数 n 较大时，可以认为是随机抽取 j 个 w_1 类别样本。(因为类概率密度 $P(X|w_i)$ 在10单位超球体内为均匀分布。)

而已知 $P(w_1)=P(w_2)=\frac{1}{2}$ 且 k 为奇数，则平均误差率为：

$$P_n(e)=\sum_{j=0}^{(k-1)/2}\binom{n}{j}P(w_1)^j P(w_2)^{n-j}$$

$$=\sum_{j=0}^{(k-1)/2}\binom{n}{j}\left(\frac{1}{2}\right)^n$$

$$=\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j}$$

即 $D_k(x)$ 中有少于等于 $(k-1)/2$ 的为 w_1 类别则样本 x 会被误归入 w_2 类别。同理，若设样本 x 的类别为 w_2，那样本被

误归入 w_1 类别的概率也相同，因此当k为奇数，平均误差率为 $P_n(e)=\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j}$ 得证.

（b）证明在这种情况下，如果 k>1，那么最近邻规则比 k-近邻规则有更低的误差率。

(b) 最近邻规则的平均误差率即将 $k=1$ 代入(a)式：

$$P_n(e)=\frac{1}{2^n}\sum_{j=0}^{0}\binom{n}{j}=\frac{1}{2^n}$$

而 $k>1$ 时，当k为奇数时，平均误差率为：

$$\begin{aligned}P_n(e)&=\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j}\\&=\frac{1}{2^n}\left[\sum_{j=0}^{0}\binom{n}{j}+\sum_{j=1}^{(k-1)/2}\binom{n}{j}\right]\\&=\frac{1}{2^n}\left[1+\sum_{j=1}^{(k-1)/2}\binom{n}{j}\right]\\&=\frac{1}{2^n}+\frac{1}{2^n}\sum_{j=1}^{(k-1)/2}\binom{n}{j}\\&>\frac{1}{2^n}\end{aligned}$$

由(a)中同理，当k为偶数时平均误差率为：

$$\begin{aligned}P_n(e)&=\frac{1}{2^n}\sum_{j=0}^{k/2-1}\binom{n}{j}\\&=\frac{1}{2^n}\left[\sum_{j=0}^{0}\binom{n}{j}+\sum_{j=1}^{k/2-1}\binom{n}{j}\right]\\&=\frac{1}{2^n}+\frac{1}{2^n}\sum_{j=1}^{k/2-1}\binom{n}{j}\\&>\frac{1}{2^n}\end{aligned}$$

因此，$k>1$，最近邻规则比k-近邻规则有更低的误差率得证.

(c) 如果 k 随着 n 的增加而增加，同时又受 $k < a\sqrt{n}$ 的限制，那么证明：

当 $n \to \infty$ 时 $P_n(e) \to 0$ 。

(c)证明：首先展开错误率计算公式：(在偶数情况下，奇数同理)

$$P_n(e) = \frac{1}{2^n}\sum_{j=0}^{k/2-1}\binom{n}{j}$$

$$= \frac{1}{2^n}\sum_{j=0}^{k/2-1}\frac{n!}{j!(n-j)!}$$

$$= \frac{1}{2^n}\binom{n}{0} + \frac{1}{2^n}\binom{n}{1} + \cdots + \frac{1}{2^n}\binom{n}{k/2-1}$$

对于组合数而言，当 $j < \frac{n}{2}$ 的时候，$\binom{n}{j}$ 递增，同时题中 $k < a\sqrt{n}$

$$\therefore \frac{1}{2^n}\binom{n}{0} < \frac{1}{2^n}\binom{n}{1} < \cdots < \frac{1}{2^n}\binom{n}{a\sqrt{n}/2-1}$$

对于任意 $\varepsilon > 0$，我们可以有：

$$\frac{1}{2^n} \times \frac{n!}{(a\sqrt{n}/2-1)!\,(n-a\sqrt{n}/2+1)!} < \varepsilon$$

$$\therefore \frac{1}{2^n}\binom{n}{0} < \frac{1}{2^n}\binom{n}{1} < \cdots < \frac{1}{2^n}\binom{n}{a\sqrt{n}/2-1} < \varepsilon$$

即：$\lim\limits_{n \to +\infty} \frac{1}{2^n}\binom{n}{j} = 0$

$$\therefore \lim_{n \to +\infty} P_n(e) = \sum_{j=0}^{k/2-1}\lim_{n \to +\infty}\frac{1}{2^n}\binom{n}{j}$$

$$= \sum_{j=0}^{k/2-1}(0)$$

$$= 0$$

因此，当 $n \to \infty$ 时，$P_n(e) \to 0$ 得证。

Yubei

Homework 2.3

考虑三维正态分布$p(\mathbf{x}|\omega) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,其中

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \quad 及 \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 5 \end{pmatrix}$$

(a) 求点$\mathbf{x_0} = (0.5, 0, 1)^t$处的概率密度;

(a) 点 $\vec{x}_0 = \begin{pmatrix} 0.5 \\ 0 \\ 1 \end{pmatrix}$ 处的概率密度为：（下面的上标t均为转置 参照的是题目写法）

$$p(\vec{x}_0|\omega) = \frac{1}{(2\pi)^{\frac{3}{2}} |\vec{\Sigma}|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(\vec{x}_0 - \vec{\mu})^t \vec{\Sigma}^{-1} (\vec{x}_0 - \vec{\mu})\right]}$$

$\because \vec{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 5 \end{pmatrix} \quad \therefore |\vec{\Sigma}|^{\frac{1}{2}} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 5 \end{vmatrix}^{\frac{1}{2}} = \left(1 \begin{vmatrix} 5 & 2 \\ 2 & 5 \end{vmatrix}\right)^{\frac{1}{2}} = \sqrt{21}$

而 $\vec{\Sigma}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 5 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}^{-1} \\ 0 & \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{5}{21} & -\frac{2}{21} \\ 0 & -\frac{2}{21} & \frac{5}{21} \end{pmatrix}$

$\therefore -\frac{1}{2}(\vec{x}_0 - \vec{\mu})^t \vec{\Sigma}^{-1} (\vec{x}_0 - \vec{\mu})$

$= -\frac{1}{2} \times \left[\begin{pmatrix} 0.5 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}\right]^t \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{5}{21} & -\frac{2}{21} \\ 0 & -\frac{2}{21} & \frac{5}{21} \end{pmatrix} \times \left[\begin{pmatrix} 0.5 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}\right]$

$= -\frac{1}{2} \times \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix}^t \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{5}{21} & -\frac{2}{21} \\ 0 & -\frac{2}{21} & \frac{5}{21} \end{pmatrix} \times \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix}$

$= -\frac{1}{2} \times \begin{pmatrix} -0.5 \\ -\frac{8}{21} \\ -\frac{1}{21} \end{pmatrix}^t \times \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix}$

$= -\frac{1}{2} \times (0.25 + \frac{16}{21} + \frac{1}{21}) = -0.53$

∴ $p(\vec{x}_0|\omega) = \dfrac{1}{(2\pi)^{\frac{3}{2}} \times \sqrt{21}} e^{-0.53}$

≈ 0.00816

(b) 构造白化变换矩阵$\mathbf{A_\omega}$($\mathbf{A_\omega} = \mathbf{\Phi\Lambda^{-1/2}}$), 计算分别表示本征向量和本征值的矩阵$\mathbf{\Phi}$和$\mathbf{\Lambda}$; 然后, 将此分布转换为以原点为中心协方差矩阵为单位阵的分布, 即$p(\mathbf{x}|\omega)\sim N(\mathbf{0},\mathbf{I})$;

(b) 令特征方程(本征方程)为0, 即$|\vec{\Sigma} - \lambda\vec{I}| = 0$可得:

$$\begin{vmatrix} 1-\lambda & 0 & 0 \\ 0 & 5-\lambda & 2 \\ 0 & 2 & 5-\lambda \end{vmatrix} = (1-\lambda)\left[(5-\lambda)^2-4\right] = 0$$

解得: $\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 7$.

那么特征矩阵(本征矩阵)即为: $\vec{\Lambda} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 7 \end{pmatrix}$

那么将特征值$\lambda_1, \lambda_2, \lambda_3$分别代入$(\vec{\Sigma}-\lambda\vec{I})\vec{x} = 0$可得:

特征向量 $\vec{\phi}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $\vec{\phi}_2 = \begin{pmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}$, $\vec{\phi}_3 = \begin{pmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$

∴特征向量矩阵为 $\vec{\Phi} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$

且 $\vec{A_w} = \vec{\Phi}\vec{\Lambda}^{-\frac{1}{2}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{6}}{6} & \frac{\sqrt{14}}{14} \\ 0 & -\frac{\sqrt{6}}{6} & \frac{\sqrt{14}}{14} \end{pmatrix}$

~~那~~同时, 构造变换 $\vec{Y} = \vec{A_w}^t(\vec{x}-\vec{\mu}) \sim N(0, \vec{I})$ 可将原分布转换为 $N(0, \vec{I})$.

(指$\vec{A_w}$的转置)

(c) 将整个同样的转换过程应用于点$\mathbf{x_0}$以产生一变换点$\mathbf{x_\omega}$；

(c) 通过(b)的转换同样:

$$\vec{x}_w = \vec{A}_w^t(\vec{x}_0 - \vec{\mu})$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{6}}{6} & \frac{\sqrt{14}}{14} \\ 0 & -\frac{\sqrt{6}}{6} & \frac{\sqrt{14}}{14} \end{pmatrix}^t \times \left[\begin{pmatrix} 0.5 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}\right]$$

$$= \begin{pmatrix} -0.5 \\ -\frac{\sqrt{6}}{6} \\ -\frac{3\sqrt{14}}{14} \end{pmatrix}$$

(d) 通过详细计算，证明原分布中从$\mathbf{x_0}$到均值$\boldsymbol{\mu}$的 Mahalanobis 距离与变换后的分布中从$\mathbf{x_\omega}$到 **0** 的 Mahalanobis 距离相等；

(d) 原分布中$\vec{x}_0$到均值$\vec{\mu}$的Mahalanobis距离为:

$$(\vec{x}_0 - \vec{\mu})^t \Sigma^{-1} (\vec{x}_0 - \vec{\mu}) \quad (\Sigma^{-1}\text{已在(a)中算出})$$

$$= \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix}^t \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{5}{21} & -\frac{2}{21} \\ 0 & -\frac{2}{21} & \frac{5}{21} \end{pmatrix} \times \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix}$$

$$= \begin{pmatrix} -0.5 \\ -\frac{8}{21} \\ -\frac{1}{21} \end{pmatrix}^t \times \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix} = 0.25 + \frac{16}{21} + \frac{1}{21} \approx 1.06$$

而变换后分布中$\vec{x}_w$到原点的Mahalanobis距离为:

$\vec{x_w}^t \Sigma^{-1} \vec{x_w} = 0.5^2 + \frac{1}{6} + \frac{9}{14} \approx 1.06$

从而得证原分布和变换后分布的距离相等.

(e) 概率密度在某个一般线性变换下是否保持不变？换句话说，对于某线性变换 **T**，是否有$p(\mathbf{x_0}|N(\boldsymbol{\mu},\boldsymbol{\Sigma}))=p(\mathbf{T^t x_0}|N(\mathbf{T^t}\boldsymbol{\mu},\mathbf{T^t}\boldsymbol{\Sigma}\mathbf{T}))$？解释原因；

(e) 对于线性变换$\vec{T}$，有

$$p(\vec{x_0}|N(\vec{\mu},\vec{\Sigma})) = p(\vec{T}^t\vec{x_0}|N(\vec{T}^t\vec{\mu},\vec{T}^t\vec{\Sigma}\vec{T}))$$

证明如下：

设原分布通过一个$\vec{T}$的线性变换，则变换后均值为：

$$\vec{\mu}' = \frac{1}{n}\sum_{i=1}^{n}\vec{x_i}'$$

$$= \frac{1}{n}\sum_{i=1}^{n}\vec{T}^t\vec{x_i}$$

$$= \vec{T}^t\frac{1}{n}\sum_{i=1}^{n}\vec{x_i} = \vec{T}^t\vec{\mu}$$

其中n为样本数，$\vec{\mu}, \vec{x}, \vec{\mu}', \vec{x}'$分别为变换前后的均值和样本

而变换后的协方差矩阵为：

$$\vec{\Sigma}' = \sum_{i=1}^{n}(\vec{x_i}'-\vec{\mu}')(\vec{x_i}'-\vec{\mu}')^t$$

$$= \sum_{i=1}^{n}(\vec{T}^t\vec{x_i}-\vec{T}^t\vec{\mu}) \times (\vec{T}^t\vec{x_i}-\vec{T}^t\vec{\mu})^t$$

$$= \vec{T}^t\left[\sum_{i=1}^{n}(\vec{x_i}-\vec{\mu})(\vec{x_i}-\vec{\mu})^t\right]\vec{T}$$

$$= \vec{T}^t\Sigma\vec{T}$$

因此，可证通过线性变换$\vec{T}$后有

$$p(\vec{x_0}|N(\vec{\mu},\vec{\Sigma})) = p(\vec{T}^t\vec{x_0}|N(\vec{T}^t\vec{\mu},\vec{T}^t\vec{\Sigma}\vec{T}))$$

(也就是概率密度在某个一般线性变换下保持不变)

(f) 证明当把一个一般的白化变换$\mathbf{A}_{\omega} = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2}$应用于一个高斯分布时可保证最终分布的协方差与单位阵 $\mathbf{I}$ 成比例，检查变换后的分布是否仍具有归一化特性。

(f) 由(e)已得通过白化变换后的协方差矩阵为

$\vec{\Sigma}' = \vec{A_w}^t \vec{\Sigma} \vec{A_w}$

而由题知 $\vec{A_w} = \vec{\Phi}\vec{\Lambda}^{-\frac{1}{2}}$，设 $\vec{\Phi} = k\vec{\Omega}$，其中 $\vec{\Omega}$ 为正交矩阵，k为常数

∵ $\vec{\Sigma}$ 为对称矩阵则有 $\vec{\Sigma} = \vec{\Omega}\Lambda\vec{\Omega}^t$

又∵ $\vec{\Lambda}$ 是对角矩阵 ∴ 也有 $(\vec{\Lambda}^{-\frac{1}{2}})^t = \vec{\Lambda}^{-\frac{1}{2}}$

且∵ $\vec{\Omega}$ 是正交矩阵 ∴ $\vec{\Omega}^t = \vec{\Omega}^{-1}$

因此 $\vec{\Sigma}' = \vec{A_w}^t \vec{\Sigma} \vec{A_w}$

$= (\vec{\Phi}\vec{\Lambda}^{-\frac{1}{2}})^t \vec{\Sigma} (\vec{\Phi}\vec{\Lambda}^{-\frac{1}{2}})$

$= \vec{\Lambda}^{-\frac{1}{2}} \vec{\Phi}^t \vec{\Omega} \vec{\Lambda} \vec{\Omega}^t \vec{\Phi} \vec{\Lambda}^{-\frac{1}{2}}$

$= k^2 \vec{\Lambda}^{-\frac{1}{2}} \vec{\Omega}^t \vec{\Omega} \vec{\Lambda} \vec{\Omega}^t \vec{\Omega} \vec{\Lambda}^{-\frac{1}{2}}$

$= k^2 \vec{\Lambda}^{-\frac{1}{2}} \vec{\Lambda} \vec{\Lambda}^{-\frac{1}{2}}$

$= k^2 \vec{I}$

∴ 变换后分布的协方差与单位阵 $\vec{I}$ 成比例.

另外.变换后的分布仍具有归一化特性.

∵ 已由(e)证明在线性变换后 $p(\vec{x_0}|N(\vec{\mu},\vec{\Sigma})) = p(\vec{T}^t\vec{x_0}|N(\vec{T}^t\vec{\mu}, \vec{T}^t\vec{\Sigma}\vec{T}))$

即概率密度在变换后保持不变.

因此.变换后的分布仍然保持了变换前归一化的特性.