# 模式识别作业五

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** (§4 Q5)**.** 证明当$\lim_{n\to\infty} k_n = \infty$和$\lim_{n\to\infty} k_n/n = 0$时，公式

$$p_n(\mathbf{x}) = \frac{k_n/n}{V_n}$$

收敛到$p(\mathbf{x})$。

**解答.** 由概率论中的结论，我们有当

$$\lim_{n\to\infty} \mathbb{E}\left(|p_n(\mathbf{x}) - p(\mathbf{x})|^2\right) = 0$$

时，$p_n(\mathbf{x})$依概率收敛到$p(\mathbf{x})$。而

$$\mathbb{E}\left(|p_n(\mathbf{x}) - p(\mathbf{x})|^2\right) = \mathbb{D}\left(p_n(\mathbf{x})\right) - \left(\mathbb{E}\left(|p_n(\mathbf{x}) - p(\mathbf{x})|\right)\right)^2$$

故欲证原题，只需证在

$$\lim_{n\to\infty} k_n = \infty$$

$$\lim_{n\to\infty} k_n/n = 0$$

的前提下，满足

$$\mathbb{E}\left(p_n(\mathbf{x})\right) \to p(\mathbf{x}), \quad n \to \infty$$

$$\mathbb{D}\left(p_n(\mathbf{x})\right) \to 0, \quad n \to \infty$$

- 下证$\mathbb{E}\left(p_n(\mathbf{x})\right) \to p(\mathbf{x}),\ n \to \infty$

  由公式(11)，

  $$p_n(\mathbf{x}) = \frac{k_n/n}{V_n} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

  其中$h_n$为包含$\mathbf{x}$的$k_n$个近邻的球的半径，$V_n$为该球的体积，且

  $$\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) = \begin{cases} 1 & \|\mathbf{x}-\mathbf{x}_i\| \leq h_n \\ 0 & \|\mathbf{x}-\mathbf{x}_i\| > h_n \end{cases}$$

  故由公式(23)，

  $$\mathbb{E}\left(p_n(\mathbf{x})\right) = \int \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{v}}{h_n}\right) p(\mathbf{v})\,\mathrm{d}\mathbf{v} = \int \delta_n(\mathbf{x}-\mathbf{v}) p(\mathbf{v})\,\mathrm{d}\mathbf{v} \to p(\mathbf{x}),\ n \to \infty$$

当且仅当

$$\frac{1}{V_n}\varphi\left(\frac{\mathbf{x}}{h_n}\right)=\delta(\mathbf{x}),\ n\to\infty$$

其中$\delta(\mathbf{x})$为狄利克函数。

因为$\lim\limits_{n\to\infty} k_n/n=0$，故可以得到$k_n$个点与$\mathbf{x}$的距离都小于$h_n$。进而当$V_n\to 0, h_n\to 0$时，有

$$\varphi\left(\frac{\mathbf{x}}{h_n}\right)=\begin{cases}1 & \mathbf{x}=\mathbf{0}\\ 0 & \mathbf{x}\neq\mathbf{0}\end{cases}$$

故$\dfrac{1}{V_n}\varphi\left(\dfrac{\mathbf{x}}{h_n}\right)=\delta(\mathbf{x})$

- 下证$\mathbb{D}\left(p_n(\mathbf{x})\right)\to 0,\ n\to\infty$

  由公式(24)，有

$$\begin{aligned}\lim_{n\to\infty}\operatorname{Var}\left[p_n(\mathrm{x})\right]&=\lim_{n\to\infty}\frac{1}{nV_n}\int\frac{1}{V_n}\varphi^2\left(\frac{\mathbf{x}-\mathbf{u}}{h_n}\right)p(\mathbf{u})d\mathbf{u}\\&=\frac{\displaystyle\int\lim_{n\to\infty}\left[\frac{1}{V_n}\varphi^2\left(\frac{\mathbf{x}-\mathbf{u}}{h_n}\right)\right]p(\mathbf{u})d\mathbf{u}}{\lim\limits_{n\to\infty}nV_n}\\&=\frac{\displaystyle\int\delta(\mathbf{x}-\mathbf{u})p(\mathbf{u})d\mathbf{u}}{\lim\limits_{n\to\infty}nV_n}\qquad\text{公式(23)}\\&=\frac{p(\mathrm{x})}{\lim\limits_{n\to\infty}nV_n}\end{aligned}$$

  又$\lim\limits_{n\to\infty}\dfrac{k_n}{nV_n}=P(\mathbf{x})$，且$\lim\limits_{n\to\infty}k_n=\infty$，进而$\lim\limits_{n\to\infty}nV_n=\infty$。最终得到$\mathbb{D}\left(p_n(\mathbf{x})\right)\to 0,\ n\to\infty$。

故原题得证。

**问题 2** (§4 Q6)**.** 令$\mathcal{D}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$为$n$个独立的已标记的样本的集合。令$\mathcal{D}_k(\mathbf{x})=\{\mathbf{x}'_1,\ldots,\mathbf{x}'_k\}$为样本$\mathbf{x}$的$k$个最近邻。回忆根据$k$-近邻规则，$\mathbf{x}$将被归入$\mathcal{D}_k(\mathbf{x})$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$p(\mathbf{x}\mid\omega_i)$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

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

*(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。

*(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。

**解答.** (a) 由于只有两个类别，故

$$\text{平均误差率}P_n(e) = P(\text{标记为}\omega_1\text{但真实类别为}\omega_2) + P(\text{标记为}\omega_2\text{但真实类别为}\omega_1)$$

又$\omega_1$和$\omega_2$的对称性，

$$P_n(e) = 2P(\text{标记为}\omega_1\text{但真实类别为}\omega_2)$$

因$k$为奇数，故$k$-近邻中若至少有$(k+1)/2$个点标记为$\omega_i$，则该点也被标记为$\omega_i$；或至多有$(k-1)/2$个点标记为$\omega_i$，则该点被标记为$\overline{\omega_i}$（这里$\bar{\cdot}$代表取反）。则

$$\begin{aligned}
P(\text{标记为}\omega_1\text{但真实类别为}\omega_2) &= P(\omega_2)P(\mathcal{D}\text{中至多有}(k-1)/2\text{个点为}\omega_2 \mid \omega_2) \\
&= \frac{1}{2}\sum_{j=1}^{(k-1)/2}\binom{n}{j}\frac{1}{2^j}\frac{1}{2^{n-j}} \\
&= \frac{1}{2^{n+1}}\sum_{j=0}^{(k-1)/2}\binom{n}{j}
\end{aligned}$$

故

$$P_n(e) = 2P(\text{标记为}\omega_1\text{但真实类别为}\omega_2) = \frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j}$$

(b) 当$k=1$时，$P_n(e) = \frac{1}{2^n}$

当$k>1$时，

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

故最近邻规则比$k$-近邻规则有更低的误差率。

(c) 由二项式系数的性质，当$j \le \lfloor n/2 \rfloor$时，$\binom{n}{j}$单调递增，而$j = 0, \ldots, (k-1)/2$，故

$$\binom{n}{j} \le \binom{n}{\frac{k-1}{2}}$$

进而

$$
\begin{aligned}
P_n(e) &= \frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j} \\
&\le \frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{\frac{k-1}{2}} \\
&= \frac{1}{2^n}\frac{k+1}{2}\binom{n}{\frac{k-1}{2}} \\
&= \frac{1}{2^n}\frac{k+1}{2}\frac{n!}{\left(\frac{k-1}{2}\right)!\left(n-\frac{k-1}{2}\right)!} \\
&\le \frac{1}{2^n}\frac{n!}{\left(n-\frac{k-1}{2}\right)!} \qquad k\ge 7 \\
&\le \frac{1}{2^n}\frac{n!}{(n-k)!} \\
&= \frac{1}{2^n}\cdot n\cdot(n-1)\cdots(n-k+1) \\
&\le \frac{1}{2^n}n^k \\
&< \frac{1}{2^n}n^{a\sqrt{n}} \\
&= \left(\frac{n^a}{2\sqrt{n}}\right)^{\sqrt{n}}
\end{aligned}
$$

因$\lim_{n\to\infty}\frac{n^a}{2^{\sqrt{n}}}=0$，故由夹逼定理$\lim_{n\to\infty}P_n(e)=0$。

**问题 3** (§4 Q17)**.** 考虑一种分类问题，总共有$c$个不同的类别，每一个类别的概率分布相同，并且每一个类别的先验概率都是$P(\omega_i)=1/c$。证明公式*(52)*所给出的误差率上界

$$
P \le P^*\left(2-\frac{c}{c-1}P^*\right)
$$

在本题中的“原信息”的场合下能够取到。

**解答.** 由题设有$p(\mathbf{x}\mid\omega_i)=p(\mathbf{x}), i=1,\ldots,c$，进而

$$
p(\mathbf{x})=\sum_{i=1}^{c}p\left(\mathbf{x}\mid\omega_i\right)P\left(\omega_i\right)=\sum_{i=1}^{c}p(\mathbf{x}\mid\omega)\frac{1}{c}=p(\mathbf{x}\mid\omega)
$$

而由条件概率

$$
P\left(\omega_i\mid\mathbf{x}\right)=\frac{p\left(\mathbf{x}\mid\omega_i\right)P\left(\omega_i\right)}{p(\mathbf{x})}=\frac{p(\mathbf{x}\mid\omega)1/c}{p(\mathbf{x}\mid\omega)}=\frac{1}{c} \qquad (*)
$$

将($*$)式代入公式(45)，有

$$
\begin{aligned}
P &= \lim_{n\to\infty} P_n(e) \\
&= \int \left[1-\sum_{i=1}^{c} P^2\left(\omega_i \mid \mathbf{x}\right)\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \int \left[1-\sum_{i=1}^{c} \frac{1}{c^2}\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \left(1-\frac{1}{c}\right)\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1-\frac{1}{c}
\end{aligned}
$$

又将($*$)式代入贝叶斯误差

$$
\begin{aligned}
P^* &= \int P^*(\text{error } \mid \mathbf{x}) p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \sum_{i=1}^{c} \int_{\mathcal{R}_i} \left[1-P\left(\omega_i \mid \mathbf{x}\right)\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \sum_{i=1}^{c} \int_{\mathcal{R}_i} \left(1-\frac{1}{c}\right) p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \left(1-\frac{1}{c}\right)\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= 1-\frac{1}{c}
\end{aligned}
$$

因此有$P=P^*$，即误差率上界

$$
P^*\left(2-\frac{c}{c-1}P^*\right) = 1-\frac{1}{c}
$$

在本题中的“原信息”的场合下能够取到。