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

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

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

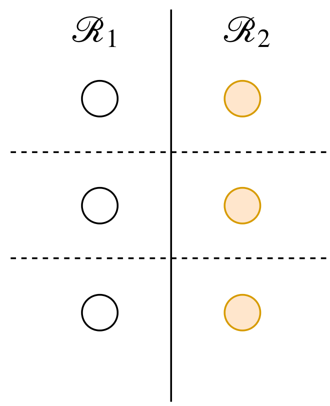
故原题得证。

问题 2 (§4 Q16). *考虑最近邻规则中的最简单的剪辑算法（算法3）。*

(a) 请给出一个反例，证明这个算法不能保证得到最小的样本点集。（可以考虑一个2类别问题，而其中的样本点都被限制在二维笛卡尔坐标网格的交点上。）

(b) 设计一种串行的剪辑算法，每一个训练样本点都被依次处理，并且在下一个点到达之前，或被保留，或被抛弃。并且说明，这样的算法产生的最后结果是否依赖于样本点的处理顺序。

解答. (a) 如下图所示，实线为判决边界，虚线为Voronoi边界。采用剪辑算法会将所有点留下，因为每一个点都有不同的Voronoi邻居。但实际上为了维持决策边界，每组左右一对点中只需留下一个，且保证$\mathcal{R}_1$和$\mathcal{R}_2$中都有点即可。

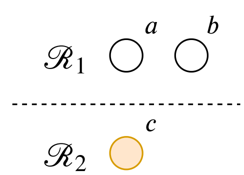


(b) 串行剪辑算法的伪代码如下

Algorithm 1 串行剪辑算法

初始化$\mathcal{D} \leftarrow$数据集，$n \leftarrow$原型点个数
for $\mathcal{D}$中每一个点$\mathbf{x}_i$ $(i \leftarrow 1$到$n)$ **do**
  if $\mathbf{x}_i$的最近邻与$\mathbf{x}_i$同类 **then**
    从$\mathcal{D}$中移除$\mathbf{x}_i$ ▷ 否则保留$\mathbf{x}_i$
return $\mathcal{D}$

考虑下面的例子，如果先处理a，则返回的$\mathcal{D} = \{b, c\}$；若先处理b，则返回$\mathcal{D} = \{a, c\}$。因此这种串行剪辑算法最后产生的结果依赖于样本点的处理顺序。



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

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

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。