第四章 非参数技术

5、证明当$\lim\limits_{n\to\infty} k_n \to \infty$和$\lim\limits_{n\to\infty} k_n/n \to 0$时，公式(30)收敛到$p(\mathbf{x})$。

证明：由公式(30)：

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

等式两边对n取极限：

$$\lim_{n\to\infty} p_n(\mathbf{x}) = \lim_{n\to\infty} \frac{k_n/n}{V_n} = \frac{\lim\limits_{n\to\infty} k_n/n}{\lim\limits_{n\to\infty} V_n}$$

因为$p_n(\mathbf{x}) \not\equiv 0$，而$\lim\limits_{n\to\infty} k_n/n = 0$，所以$\lim\limits_{n\to\infty} V_n = 0$。

定义样本点$\mathbf{x}$落在体积为V_n的区域D中的频率为P_n，则：

$$P_n = \int_D p(\mathbf{x}) d\mathbf{x} = \frac{k_n}{n}$$

等式两边对n取极限：

$$\lim_{n\to\infty} \int_D p(\mathbf{x}) d\mathbf{x} = \lim_{n\to\infty} \frac{k_n}{n}$$

由于当$n\to\infty$时，$V_n \to 0$，可以认为在区域D中$\mathbf{x}$的概率密度函数为一个常数，因此：

$$\lim_{n\to\infty} \int_D p(\mathbf{x}) d\mathbf{x} = p(\mathbf{x}) \lim_{n\to\infty} \int_D 1 d\mathbf{x} = p(\mathbf{x}) \lim_{n\to\infty} V_n = \lim_{n\to\infty} \frac{k_n}{n}$$

所以：

$$\lim_{n\to\infty} p_n(\mathbf{x}) = \lim_{n\to\infty} \frac{k_n/n}{V_n} = p(\mathbf{x})$$

（更严格的话，还可以证明方差收敛于0）

17、考虑一种分类问题，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$P(w_i) = 1/c$。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“零信息”的场合下取得。

证明：在本题中所限定的“零信息”可用如下条件表示：

每一个类别的概率分不相同，即对任意的$\mathbf{x}$，$p(\mathbf{x}|w_i)$相等，$i = 1, \cdots, c$；

每一个类别的先验概率相等，即$P(w_i) = 1/c$，$i = 1, \cdots, c$；

根据Bayes公式，$P(w_i|\mathbf{x})=\dfrac{p(\mathbf{x}|w_i)P(w_i)}{p(\mathbf{x})}$，后验概率 $P(w_i|\mathbf{x})$ 均相等，$i=1,\cdots,c$，因此：$P(w_i|\mathbf{x})=\dfrac{1}{c}$。

首先计算Bayes误差率 P^*：因为后验概率 $P(w_i|\mathbf{x})$ 均相等，因此根据Bayes决策准则，可以将 $\mathbf{x}$ 判别为任意的类别 w_m，而 $P(w_m|\mathbf{x})=\dfrac{1}{c}$，因此对 $\mathbf{x}$ 判别的错误率为：$P(e|\mathbf{x})=1-\dfrac{1}{c}$，因此：

$$P^*=\int P(e|\mathbf{x})\,p(\mathbf{x})\,d\mathbf{x}=1-\frac{1}{c}$$

然后计算最近邻分类规则的误差率 P：利用148页式(45)：

$$P=\int\left[1-\sum_{i=1}^{c}P^2(w_i|\mathbf{x})\right]p(\mathbf{x})\,d\mathbf{x}=\int\left(1-c\frac{1}{c^2}\right)p(\mathbf{x})\,d\mathbf{x}=1-\frac{1}{c}$$

因此在“零信息”场合，最近邻分类的误差率取得其上界。

19、考虑 d 维空间中的Euclid距离度量：

$$D(\mathbf{a},\mathbf{b})=\sqrt{\sum_{k=1}^{d}(a_k-b_k)^2}$$

假设我们对每一个坐标轴都进行尺度变换，也就是说 $x_k'=a_kx_k$，$k=1,2,\cdots,d$，其中 a_k 为非负实数。证明坐标变换后的空间为一个度量空间。并且讨论这一性质对标准的最近邻规则算法的重要性。

证明：令 $\mathbf{x},\mathbf{y},\mathbf{z}$ 为原 d 维空间中的三个矢量，定义 $d\times d$ 的矩阵

$$\mathbf{A}=\begin{bmatrix} a_i & 0 & \cdots & 0\\ 0 & a_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & a_d \end{bmatrix}$$

$\mathbf{x}',\mathbf{y}',\mathbf{z}'$ 为经过坐标变换之后空间中的三个矢量，$\mathbf{x}'=\mathbf{A}\mathbf{x},\mathbf{y}'=\mathbf{A}\mathbf{y},\mathbf{z}'=\mathbf{A}\mathbf{z}$，在变换后的空间中可以定义度量：

$$D'(\mathbf{x}',\mathbf{y}')=\sqrt{\sum_{i=1}^{d}(x_i'-y_i')^2}=\sqrt{\sum_{i=1}^{d}a_i^2(x_i-y_i)^2}$$

验证该度量满足距离度量的4个条件，根据 $D'(\mathbf{x}',\mathbf{y}')$ 的表达式，非负性、自反性和对称性显然成立，下面证明三角不等式。

方法一：应用Minkowski不等式（证明见20题）：

$$\left(\sum_{i=1}^{d}\left|t_i+s_i\right|^p\right)^{1/p}\le\left(\sum_{i=1}^{d}\left|t_i\right|^p\right)^{1/p}+\left(\sum_{i=1}^{d}\left|s_i\right|^p\right)^{1/p}$$

令其中 $p=2$， $t_i=x_i'-y_i'=a_i\left(x_i-y_i\right)$， $s_i=y_i'-z_i'=a_i\left(y_i-z_i\right)$，则：

$$t_i+s_i=x_i'-z_i'=a_i\left(x_i-z_i\right)$$

因此有不等式：

$$\sqrt{\sum_{i=1}^{d}\left(x_i'-y_i'\right)^2}+\sqrt{\sum_{i=1}^{d}\left(y_i'-z_i'\right)^2}=\sqrt{\sum_{i=1}^{d}a_i^2\left(x_i-y_i\right)^2}+\sqrt{\sum_{i=1}^{d}a_i^2\left(y_i-z_i\right)^2}$$

$$\ge\sqrt{\sum_{i=1}^{d}a_i^2\left(x_i-z_i\right)^2}=\sqrt{\sum_{i=1}^{d}\left(x_i'-z_i'\right)^2}$$

因此 $D'\left(\mathbf{x}',\mathbf{y}'\right)$ 为变换后的空间中的距离度量，而变换后的空间为度量空间。

方法二：直接证明 d 维空间中的欧氏距离度量满足三角不等式：

$$\sqrt{\sum_{k=1}^{d}\left(x_k-y_k\right)^2}+\sqrt{\sum_{k=1}^{d}\left(y_k-z_k\right)^2}\le\sqrt{\sum_{k=1}^{d}\left(x_k-z_k\right)^2}$$

首先证明不等式： $\sum_{k-1}^{d}\sum_{i=1}^{d}a_k^2b_i^2\ge\sum_{k-1}^{d}\sum_{i=1}^{d}a_kb_ka_ib_i$

因为：

$$\sum_{k=1}^{d}\sum_{i=1}^{d}\left(a_k^2b_i^2-a_kb_ka_ib_i\right)=\sum_{k=1}^{d}\sum_{i=1}^{d}a_kb_i\left(a_kb_i-a_ib_k\right)$$

$$=\sum_{k=1}^{d}\sum_{i=1,k<i}^{d}a_kb_i\left(a_kb_i-a_ib_k\right)+\sum_{k=1}^{d}\sum_{i=1,k=i}^{d}a_kb_i\left(a_kb_i-a_ib_k\right)+\sum_{k=1}^{d}\sum_{i=1,k>i}^{d}a_kb_i\left(a_kb_i-a_ib_k\right)$$

注意到中间一项等于0，最后一项交换 i,k 符号，则有：

$$\sum_{k=1}^{d}\sum_{i=1}^{d}a_k^2b_i^2-\sum_{k=1}^{d}\sum_{i=1}^{d}a_kb_ka_ib_i=\sum_{k=1}^{d}\sum_{i=1,k<i}^{d}\left[a_kb_i\left(a_kb_i-a_ib_k\right)+a_ib_k\left(a_ib_k-a_kb_i\right)\right]$$

$$=\sum_{k=1}^{d}\sum_{i=1,k<i}^{d}\left[\left(a_kb_i-a_ib_k\right)\left(a_kb_i-a_ib_k\right)\right]=\sum_{k=1}^{d}\sum_{i=1,k<i}^{d}\left(a_kb_i-a_ib_k\right)^2\ge 0$$

因此得证： $\sum_{k=1}^{d}\sum_{i=1}^{d}a_k^2b_i^2\ge\sum_{k=1}^{d}\sum_{i=1}^{d}a_kb_ka_ib_i$ ，亦即 $\left(\sum_{k=1}^{d}a_k^2\right)\left(\sum_{k=1}^{d}b_k^2\right)\ge\left(\sum_{k=1}^{d}a_kb_k\right)^2$

利用上述不等式：

$$\left(\sqrt{\sum_{k=1}^{d}a_k^2}+\sqrt{\sum_{k=1}^{d}b_k^2}\right)^2=\sum_{k=1}^{d}a_k^2+\sum_{k=1}^{d}b_k^2+2\sqrt{\sum_{k=1}^{d}a_k^2\bullet\sum_{k=1}^{d}b_k^2}$$

$$\geq\sum_{k=1}^{d}a_k^2+\sum_{k=1}^{d}b_k^2+2\sum_{k=1}^{d}a_kb_k=\sum_{k=1}^{d}\left(a_k+b_k\right)^2$$

不等式两边开方，即得：

$$\sqrt{\sum_{k=1}^{d}a_k^2}+\sqrt{\sum_{k=1}^{d}b_k^2}\geq\sqrt{\sum_{k=1}^{d}\left(a_k+b_k\right)^2}$$

令 $a_k=x_k-y_k,b_k=y_k-z_k$，则 $a_k+b_k=x_k-z_k$，代入上式：

$$\sqrt{\sum_{k=1}^{d}\left(x_k-y_k\right)^2}+\sqrt{\sum_{k=1}^{d}\left(y_k-z_k\right)^2}\geq\sqrt{\sum_{k=1}^{d}\left(x_k-z_k\right)^2}$$

因此 d 维空间中的欧氏距离度量满足距离度量的4个条件。

（三角不等式也可以用数学归纳法证）

该性质对于最近邻规则算法的重要性在于：最近邻算法的有效性会受到各维特征所选择的单位的影响，往往是单位选择较小的特征在欧氏距离的计算中起着较重要的作用，而单位较大的特征之间的差异在距离计算中往往被掩盖掉。为了解决上述问题，在分类之前可以分贝对每个特征进行一个尺度变换，使得特征的尺度均衡化，具体的可以选择变换系数：

$$a_i=\frac{1}{\max\left(x_i\right)-\min\left(x_i\right)},\quad i=1,\cdots,d$$

其中 $\max\left(x_i\right)$ 表示所有训练样本中第 i 维特征的最大值，$\min\left(x_i\right)$ 表示所有训练样本中第 i 维特征的最大值。然后在变换之后的空间中计算矢量之间欧氏距离。

20、证明 Minkowski 距离度量具有成为一种度量所需要的全部 4 种性质。

证明：作为度量必须满足 4 个性质，对于任意的向量 $\mathbf{a},\mathbf{b}$ 和 $\mathbf{c}$ 有

1) 非负性：$D\left(\mathbf{a},\mathbf{b}\right)\geq 0$；

2) 自反性：$D\left(\mathbf{a},\mathbf{b}\right)=0$ 当且仅当 $\mathbf{a}=\mathbf{b}$；

3) 对称性：$D\left(\mathbf{a},\mathbf{b}\right)=D\left(\mathbf{b},\mathbf{a}\right)$；

4) 三角不等式：$D\left(\mathbf{a},\mathbf{b}\right)+D\left(\mathbf{b},\mathbf{c}\right)\geq D\left(\mathbf{a},\mathbf{c}\right)$。

而 d 为空间中的 Minkowski 距离度量为：

$$L_k(\mathbf{a},\mathbf{b})=\left(\sum_{i=1}^{d}\left|a_i-b_i\right|^k\right)^{1/k}$$

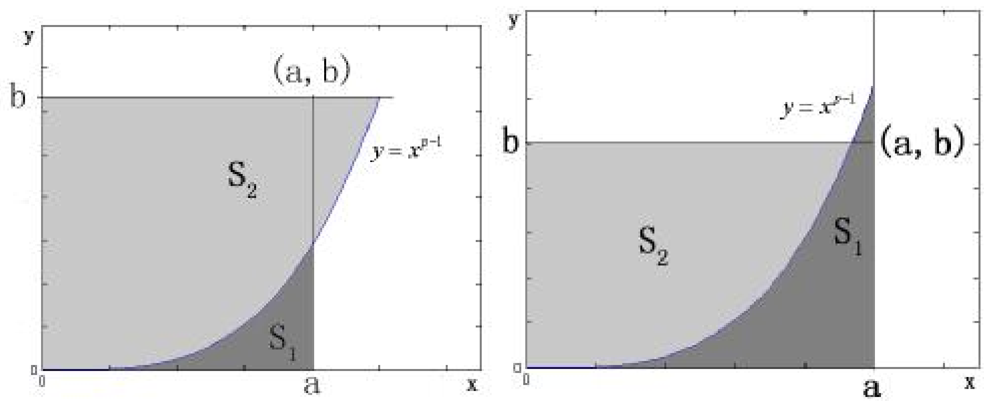
由 Minkowski 距离度量的表达式，性质 1~3 显然成立，下面证明三角不等式。

I、 证明辅助不等式：$ab\le\frac{a^p}{p}+\frac{b^q}{q}$，其中 $a\ge0$，$b\ge0$，$p>1$，$\frac{1}{p}+\frac{1}{q}=1$，$p,q$ 称为伴随数。

考察 2 维平面上的曲线 $y=x^{p-1}$，由如下图形可见，对任意的 a 和 b，必为下述两种情况之一，因此有：

$$ab\le S_1+S_2$$

其中 S_1 和 S_2 分别为两个阴影部分的面积，而 ab 为长方形部分的面积，当 $b=a^{p-1}$ 时上式取等号。



分别计算两部分的面积，其中 S_2 为反函数 $x=y^{q-1}$ 与 y 轴之间的面积（$q-1=\frac{1}{p-1}$）：

$$S_1=\int_0^a x^{p-1}dx=\left.\frac{x^p}{p}\right|_0^a=\frac{a^p}{p},\quad S_2=\int_0^b y^{q-1}dy=\left.\frac{y^q}{q}\right|_0^b=\frac{b^q}{q}$$

因此：$ab\le\frac{a^p}{p}+\frac{b^q}{q}$

II、证明 Hölder 不等式：$\sum_{i=1}^{n}\left|x_iy_i\right|\le\left(\sum_{i=1}^{n}\left|x_i\right|^p\right)^{1/p}\left(\sum_{i=1}^{n}\left|y_i\right|^q\right)^{1/q}$，其中 p,q 为伴随数，x_i,y_i 为实数。

令：$a_i=\frac{\left|x_i\right|}{\left(\sum_{k=1}^{n}\left|x_k\right|^p\right)^{1/p}}$，$b_i=\frac{\left|y_i\right|}{\left(\sum_{k=1}^{n}\left|y_k\right|^q\right)^{1/q}}$

利用辅助不等式有：$a_i b_i \le \frac{a_i^p}{p}+\frac{b_i^q}{q}$，因此：

$$\frac{|x_i|\bullet|y_i|}{\left(\sum_{k=1}^{n}|x_k|^p\right)^{1/p}\left(\sum_{k=1}^{n}|y_k|^q\right)^{1/q}} \le \frac{|x_i|^p}{p\left(\sum_{k=1}^{n}|x_k|^p\right)}+\frac{|y_i|^q}{q\left(\sum_{k=1}^{n}|y_k|^q\right)}$$

不等式两边对 i 求和：

$$\frac{\sum_{i=1}^{n}|x_i y_i|}{\left(\sum_{k=1}^{n}|x_k|^p\right)^{1/p}\left(\sum_{k=1}^{n}|y_k|^q\right)^{1/q}} \le \frac{\sum_{i=1}^{n}|x_i|^p}{p\left(\sum_{k=1}^{n}|x_k|^p\right)}+\frac{\sum_{i=1}^{n}|y_i|^q}{q\left(\sum_{k=1}^{n}|y_k|^q\right)}=\frac{1}{p}+\frac{1}{q}=1$$

因此：

$$\sum_{i=1}^{n}|x_i,y_i| \le \left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}\left(\sum_{i=1}^{n}|y_i|^q\right)^{1/q}$$

III、证明 Minkowski 不等式：$\left(\sum_{i=1}^{n}|x_i+y_i|^p\right)^{1/p} \le \left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|y_i|^p\right)^{1/p}$

当 $p=1$时，$\sum_{i=1}^{n}|x_i+y_i| \le \sum_{i=1}^{n}|x_i|+\sum_{i=1}^{n}|y_i|$ 显然成立，令 $p>1$，由下列恒等式成立：

$$(|a|+|b|)^p=(|a|+|b|)^{p-1}|a|+(|a|+|b|)^{p-1}|b|$$

这是因为等式右边：

$$(|a|+|b|)^{p-1}|a|+(|a|+|b|)^{p-1}|b|=\frac{(|a|+|b|)^p|a|+(|a|+|b|)^p|b|}{|a|+|b|}=(|a|+|b|)^p$$

令 $a=x_i, b=y_i$，带入恒等式，等式两边对 i 求和：

$$\sum_{i=1}^{n}(|x_i|+|y_i|)^p=\sum_{i=1}^{n}(|x_i|+|y_i|)^{p-1}|x_i|+\sum_{i=1}^{n}(|x_i|+|y_i|)^{p-1}|y_i| \tag{1}$$

由于：$\frac{1}{p}+\frac{1}{q}=1$，因此 $\frac{1}{p}=1-\frac{1}{q}=\frac{q-1}{q}$，$q=(q-1)p$，同理：$p=(p-1)q$

应用 Hölder 不等式：

$$\sum_{i=1}^{n}(|x_i|+|y_i|)^{p-1}|x_i|=\sum_{i=1}^{n}\left|x_i\bullet(|x_i|+|y_i|)^{p-1}\right|$$

$$\le\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}\left(\sum_{i=1}^{n}\left[(|x_i|+|y_i|)^{p-1}\right]^q\right)^{1/q}$$

$$=\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^{(p-1)q}\right)^{1/q}$$

$$=\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{1/q}\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}$$

同理：

$$\sum_{i=1}^{n}(|x_i|+|y_i|)^{p-1}|y_i|\le\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{1/q}\left(\sum_{i=1}^{n}|y_i|^p\right)^{1/p}$$

代入(1)式：

$$\sum_{i=1}^{n}(|x_i|+|y_i|)^p\le\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{1/q}\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{1/q}\left(\sum_{i=1}^{n}|y_i|^p\right)^{1/p}$$

$$=\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{1/q}\left[\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|y_i|^p\right)^{1/p}\right]$$

上式两边除以$\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{1/q}$：

$$\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{1-1/q}=\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{1/p}\le\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|y_i|^p\right)^{1/p}$$

而：$\left(\sum_{i=1}^{n}|x_i+y_i|^p\right)^{1/p}\le\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{1/p}$（$p>1$），所以有 Minkowski 不等式：

$$\left(\sum_{i=1}^{n}|x_i+y_i|^p\right)^{1/p}\le\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|y_i|^p\right)^{1/p}$$

IV、 证明 Minkowski 度量满足三角不等式：

令 $x_i=a_i-b_i, y_i=b_i-c_i$，则 $x_i+y_i=a_i-c_i$，分别代入 Minkowski 不等式：

$$\left(\sum_{i=1}^{d}|a_i-c_i|^p\right)^{1/p}\le\left(\sum_{i=1}^{d}|a_i-b_i|^p\right)^{1/p}+\left(\sum_{i=1}^{d}|b_i-c_i|^p\right)^{1/p}$$

因此 Minkowski 距离度量具有度量所需要满足的所有 4 个性质。