ML-01

一、概率论

(1)

$$F_X(x)=\begin{cases}0 & , & x\le 0\\ \frac{x}{4} & , & 0<x<1\\ \frac{1}{4} & , & 1\le x\le 3\\ \frac{3x-7}{8} & , & 3<x<5\\ 1 & , & x\ge 5\end{cases}$$

(2)

由于$Y=g(X)=1/X$单调可导，可利用概率密度公式

$$f_Y(y)=\begin{cases}f_X(h(y))|h'(y)| & y\in(\alpha,\beta)\\ 0 & 其它,\end{cases}$$

其中$h(X)=g^{-1}(X)=1/X$

求得$f_Y(y)=\frac{f_X(1/y)}{y^2}$，再带入$f_X(x)$

$$f_Y(y)=\begin{cases}0 & , & 0<y<1/5\\ \frac{3}{8y^2} & , & 1/5<y<1/3\\ 0 & , & 1/3\le y<1\\ \frac{1}{4y} & , & y\ge 1\end{cases}$$

(3)

证明：

$$\begin{aligned}\Leftrightarrow\quad & \int_{z=0}^{\infty} zf(z)dz & - & \int_{z=0}^{\infty} P[Z\ge z]dz &= 0\\ \Leftrightarrow\quad & \int_{z=0}^{\infty}\int_{t=0}^{z} f(z)dtdz & - & \int_{z=0}^{\infty}\int_{t=z}^{\infty} f(t)dtdz &= 0\\ \Leftrightarrow\quad & \int_{z=0}^{\infty}\int_{t=0}^{z} f(z)dtdz & - & \int_{t=0}^{\infty}\int_{z=0}^{t} f(t)dzdt &= 0\end{aligned}$$

z,t对称，显然成立

验证：

1. $$\begin{aligned}E[X] &= \int_{x=0}^{\infty} xf(x)dx\\ &= \int_{x=0}^{1} x/4\,dx+\int_{x=3}^{5} 3x/8\,dx=\frac{25}{8}\end{aligned}$$

2.

$$\begin{aligned} E[X] &= \int_{x=0}^{\infty} P[X \geq x]dx \\ &= \int_{x=0}^{\infty}[1-F(x)]dx \\ &= \int_{x=0}^{1} \frac{4-x}{4}dx + \int_{x=1}^{3} \frac{3}{4}dx + \int_{x=3}^{5} \frac{15-3x}{8}d \\ &= \frac{7}{8} + \frac{3}{2} + \frac{15}{4} - 3 = \frac{25}{8} \end{aligned}$$

3.

$$\begin{aligned} E[Y] &= \int_{y=0}^{\infty} yf(y)dy \\ &= \int_{\frac{1}{5}}^{\frac{1}{3}} \frac{3y}{8y^2}dy + \int_{1}^{\infty} \frac{1}{4y}dy \\ &= \infty \end{aligned}$$

4.

$$\begin{aligned} E[Y] &= \int_{y=0}^{\infty} P[Y \geq y]dy \\ &= \int_{y=0}^{\infty}[1-F(y)]dy \\ &= \infty \end{aligned}$$

二、自助法评估

(1)&(2)

$$\begin{aligned} 1.E[\bar{x}_m] &= \frac{1}{m}E[\sum_{i=1}^{m} x_i] = \frac{1}{m}m\mu = \mu \\ \text{由方差定义}E[x_i^2] &= \sigma^2 + \mu^2 \\ 2.E[\bar{x}_m^2] &= \frac{1}{m^2}E[(\sum_{i=1}^{m} x_i)^2] \\ &= \frac{1}{m^2}E[\sum_{i=1}^{m} x_i^2] + \frac{1}{m^2}E[\sum_{i\neq j} x_i x_j] \\ &= \frac{m\sigma^2 + m\mu^2 + m(m-1)\mu^2}{m^2} \\ &= \frac{\sigma^2}{m} + \mu^2 \\ 3.E[\bar{\sigma}_m^2] &= \frac{1}{m-1}E[\sum_{i=1}^{m}(x_i - \bar{x}_m)^2] \\ &= \frac{m}{m-1}E[x_i^2] - \frac{m}{m-1}E[\bar{x}_m^2] \\ &= \frac{m\sigma^2 + m\mu^2 - \sigma^2 - m\mu^2}{m-1} \\ &= \sigma^2 \\ 4.Var(\bar{x}_m) &= E[\bar{x}_m^2] - E^2[\bar{x}_m] = \frac{\sigma^2}{m} \end{aligned}$$

(3)

$$
\begin{aligned}
E[x_i^*|x_1,\cdots,x_m] &= \frac{1}{m}\sum_{i=1}^{m} x_i = \bar{x}_m \\
Var\left[x_i^* \mid x_1,\cdots,x_m\right] &= E\left[\left(x_i-\bar{x}_m\right)^2 \mid x_1,\cdots,x_m\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(x_i-\bar{x}_m\right)^2 \\
&= \frac{m-1}{m}\bar{\sigma}_m^2 \\
Var\left[\bar{x}_m^* \mid x_1,\cdots,x_m\right] &= \frac{1}{m^2}\cdot\sum_{i=1}^{m} Var\left[x_i^* \mid x_1,\cdots,x_m\right] = \frac{m-1}{m^2}\bar{\sigma}_m^2
\end{aligned}
$$

(4)

$$
\begin{aligned}
E\left[x_i^*\right] &= \sum_{i=1}^{m}\frac{1}{m}E\left[x_i\right] = \mu \\
Var\left[x_i^*\right] &= E\left[x_i^{*2}\right] - E[x_i^*]^2 \\
&= \sum_{i=1}^{m}\frac{1}{m}E\left[x_i^2\right] - \mu^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(E\left[x_i^2\right] - E[x_i]^2\right) \\
&= \frac{1}{m}\sum_{i=1}^{m} Var\left[x_i\right] \\
&= \sigma^2 \\
Var\left[\bar{x}_m^*\right] &= \frac{1}{m^2}\cdot\sum_{i=1}^{m} Var\left[x_i^*\right] = \frac{1}{m}\sigma^2
\end{aligned}
$$

(5)

自助法

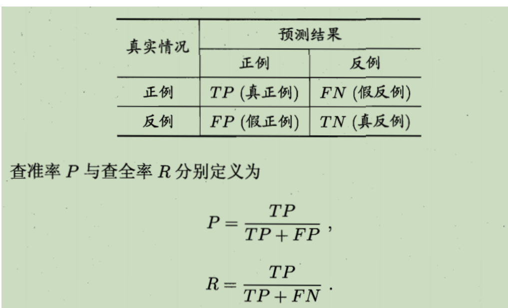
可以维持期望方差和原数据集相同，但会改变初始数据集的分布，不可避免重复抽取，会引入估计偏差。在数据集较小、难以有效划分训练测试集时很有用

交叉验证法

同一个样例不会被多次抽取，在大数据集上相对准确，但计算开销大

三、性能度量

以每个分类器输出值为阈值，判定正例和负例，得出每一个点对应的P-R值

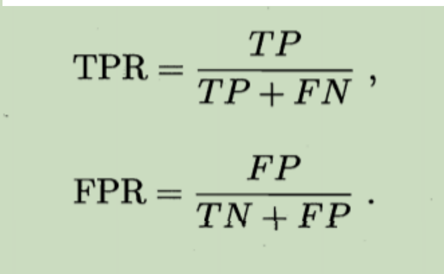


真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

查准率 P 与查全率 R 分别定义为

$$P = \frac{TP}{TP+FP},$$

$$R = \frac{TP}{TP+FN}.$$



$$\mathrm{TPR} = \frac{TP}{TP+FN},$$

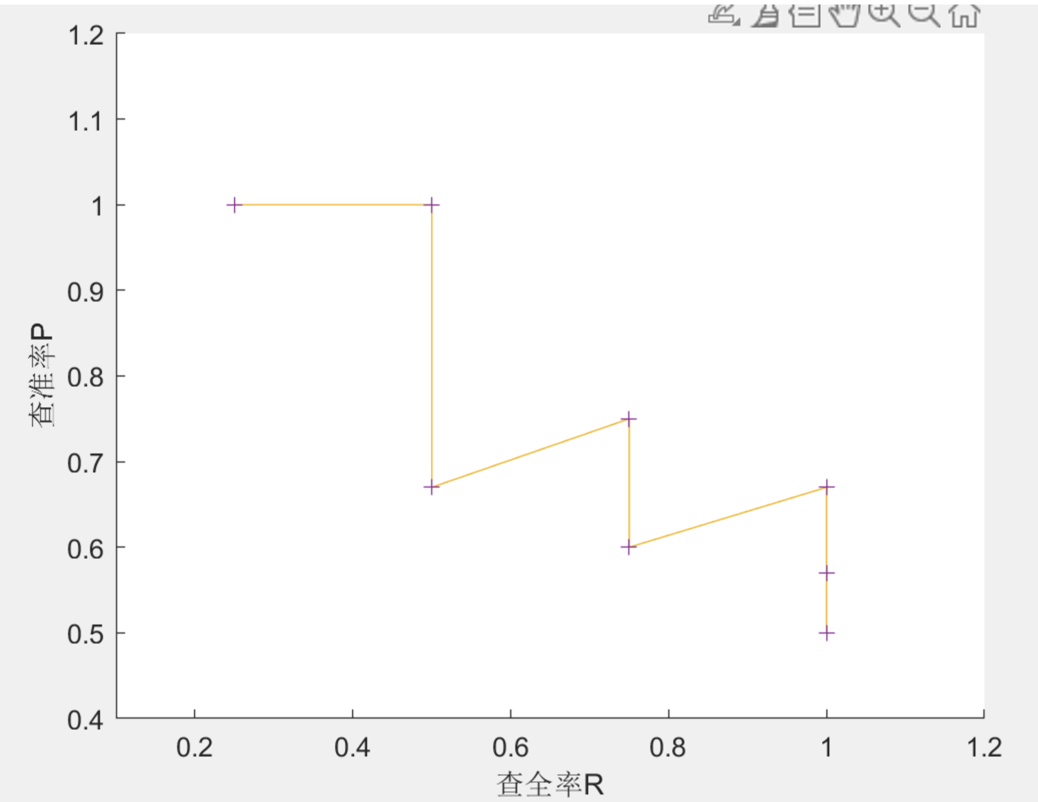
$$\mathrm{FPR} = \frac{FP}{TN+FP}.$$

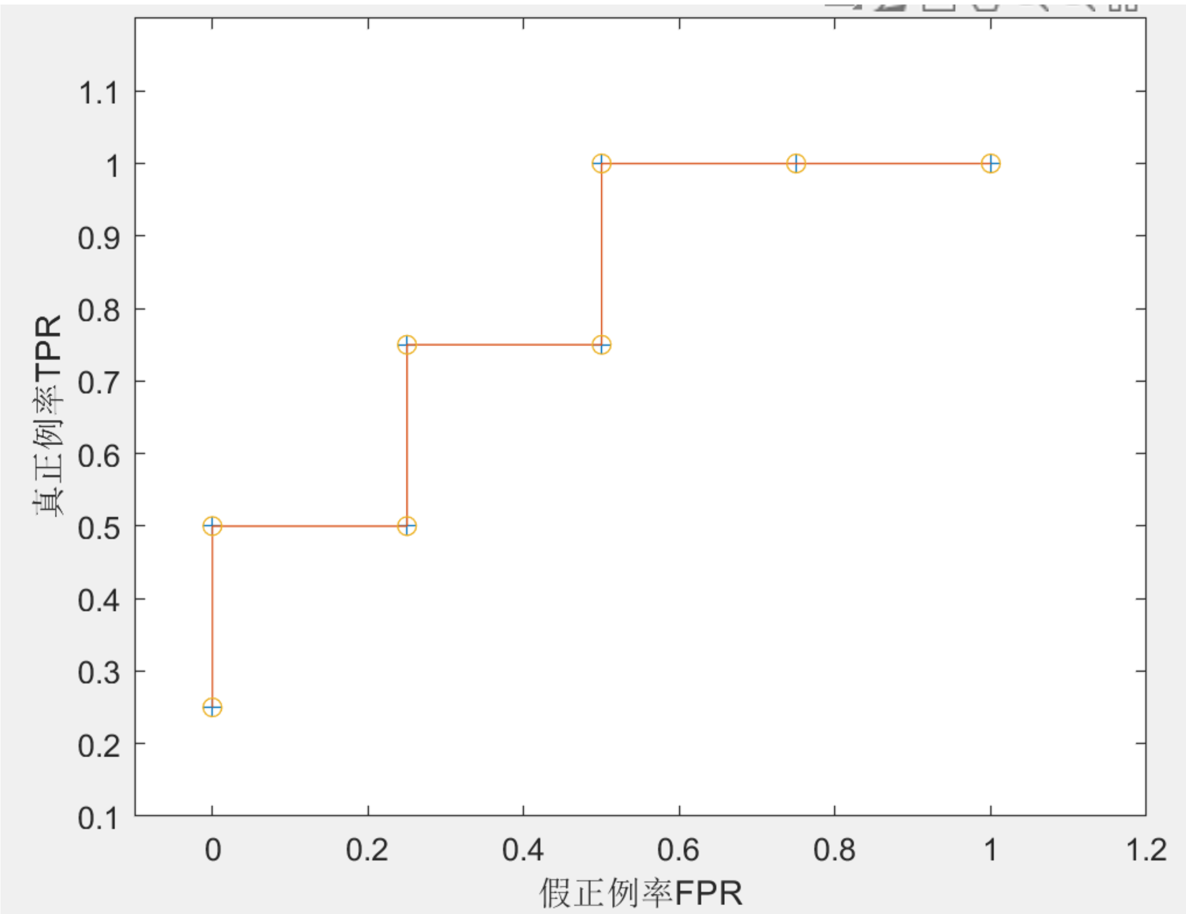
Table 1: 样例表

样例	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
标记	1	1	0	1	0	1	0	0
分类器输出值	0.81	0.74	0.62	0.55	0.44	0.35	0.25	0.21

点	1	2	3	4	5	6	7	8
TP	1	2	2	3	3	4	4	4
FP	0	0	1	1	2	2	3	4
FN	3	2	2	1	1	0	0	0
TN	4	4	3	3	2	2	1	0

点	1	2	3	4	5	6	7	8
P	1.0	1.0	0.67	0.75	0.6	0.67	0.57	0.5
R	0.25	0.5	0.5	0.75	0.75	1.0	1.0	1.0
TPR	0.25	0.5	0.5	0.75	0.75	1.0	1.0	1.0
FPR	0	0	0.25	0.25	0.5	0.5	0.75	1.0





AUC 面积为 $S=\frac{1}{2}\sum_{i=1}^{m-1}\left(x_{i+1}-x_{i}\right)\cdot\left(y_{i}+y_{i+1}\right)=0.8125$