Chapter 2:

2.1 留出法 (hold-out) { 采样 (sampling)
分层采样 (stratified sampling)

sampling: $\binom{1000}{700}$

stratified sampling: ~~$\binom{700}{350}$~~ $\binom{500}{350} \cdot \binom{500}{150}$

2.2 10折验证法 (10-fold cross validation):

card(S)=90 里 45 个正, 45 个误, 随机猜测, 正确率 50%

留一法: (Leave-One-Out)

计 99 里 49 正, 50 误, 猜为误, 正确率 0%.

2.3 $F_1 = \frac{2P \times R}{P+R}$, Bep = value(P=R).

2.4 预测结果:

实际:	正例	反例
正例	TP	FN
反例	FP	TN

(confusion matrix)

recall: $R = \frac{TP}{TP+FN}$

precision: $P = \frac{TP}{TP+FP}$

真正率 $TPR = \frac{TP}{TP+FN} = R$

假正率 $FPR = \frac{FP}{TN+FP} = \frac{FP}{\text{actual } N}$

2.5 ① ~~证明：假设每个预测~~

我们把每个预测的预测值从小到大设为 threshold.

得到坐标 $\{(x_1, y_1) \cdots (x_m, y_m)\}$.

若 (x_{k-1}, y_{k-1}) 预测正确，则 $x_k = x_{k-1}$，$y_k = y_{k-1} + \frac{1}{m^+}$

反之，$x_k = x_{k-1} + \frac{1}{m^-}$，$y_k = y_{k-1}$.

② 我们考虑 1-trank，x 轴正间长为 $\frac{1}{m^-}$

~~1-trank = $\sum_{x^- \in D^-}$~~.

每个预测设为 $c_1, \cdots c_m$，

$\{x^-\} + \{x^+\} = \{c_i\}_{i=1}^{m}$，每个预测值有 $f(c_i) < f(c_j)$ $(\forall i$

我们得到 ~~(x_{k-1}, y_{k-1}) 与 (x_k, y_k) 之间~~ 的方柱 ~~面积为~~ 底为 x_i^- 高度为

(设此时横坐标为 x_i^-)

即在 x_i^- 的阈值时，所有 TP 的比例.

$$h(x_i^-) = \sum_{x^+ \in D^+} \mathbb{I}(f(x^+) > f(x_i^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x_i^-))$$

(此时 ~~取~~ 阈值为 $f(x_i^-)$，所有 $x_j^+ \in$ ~~自~~ D^+，~~要判断~~

若 $f(x_j^+) > f(x_i^-)$，则 x_j^+ 为 TP. 中间取 = 的就取 $\frac{1}{2}$.

得 $1\text{-trank} = \sum_{i \in D^-} \cdot$ ~~面积~~ $\left($ 底为 $\frac{1}{m^-}$ ~~面积~~ $\cdot\, h(x_i^-)\right)$

$$= \sum_{x_i^- \in D^-} \frac{1}{m^-} \cdot \frac{1}{m^+} \cdot \sum_{x^+ \in D^+} \left(\mathbb{I}(f(x^+) > f(x_i^-)) + \frac{1}{2} \cdots\right)$$

$$= \frac{1}{D^-} \cdot \frac{1}{D^+} \sum_{x^- \in D^-} \sum_{x^+ \in D^+} \left(\mathbb{I} f(x^+) > f(x^-) + \frac{1}{2}\mathbb{I} \cdots\right)$$

trank 可以显然的得出. 口

2.6 错误率: $E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)\neq y_i)$

$$=\frac{FN+FP}{TP+TN+FP+FN}=\frac{FN+FP}{P+N}$$

ROC 曲线: $\left(\frac{FP}{TN+FP},\ \frac{TP}{TP+FN}\right)=\left(\frac{FP}{N},\ \frac{TP}{P}\right)$

N denotes negative test sampling,

P denotes positive test sampling.

我们得到，若 $\frac{FP}{N}<\frac{TP}{P}$ $\frac{FP}{N}<\frac{FN+FP}{P+N}<\frac{TP}{P}$,

即 $\frac{FN+FP}{P+N}$ 在 $y=x$ 的上方，反之在下方.

2.7 给定 (x_1,y_1) —— (x_m,y_m) 顺序，预测值从小到大排列

ROC 与上述预测结果一一对应

代价曲线（cost curve）显然也与预测结果一一对应.

（看不懂代价曲线的 p 是什么?）

瞎猜的证明.

2.8 ① Min-max 规范化 把取值范为从 $[x_{min}, x_{max}]$

线性变换为 $[x'_{min}, x'_{max}]$，适用于平均分布的全函数

② z-score: $x'=\frac{x-\bar{x}}{\sigma_x}$, $E(x')=0$, $Var(x')=1$.

把分布正规化，适用于接近正态分布或其他类似分布.