2.2 数据集包含 100 个样本, 其中正、反例各一半, 假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别(训练样本数相同时进行随机猜测), 试给出用 10 折交叉验证法和留一法分别对错误率进行评估所得的结果.

10折交叉验证法:

假设采用分层采样对数据集进行划分，则每次的训练集中正、反例样本数都相同，进行随机猜测，测试集中正、反例样本数也都相同，则错误率为50%

留一法:

① 留出的测试样本为正例，训练集中正样本：负样本 = 49:50，预测为负例，错误率100%

② 留出的测试样本为负例，训练集中正样本：负样本 = 50:49，预测为正例，错误率100%

故错误率是100%

2.4 试述真正例率(TPR)、假正例率(FPR)与查准率(P)、查全率(R)之间的联系.

表 2.1 分类结果混淆矩阵

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)

$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{TN+FP} \qquad P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN}$$

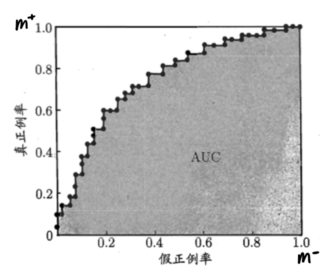
$TPR = R$，其他公没有直接联系。

2.5 试证明式(2.22).

$$\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1})\,. \tag{2.20}$$

$$\ell_{rank} = \frac{1}{m^+m^-}\sum_{\boldsymbol{x}^+\in D^+}\sum_{\boldsymbol{x}^-\in D^-}\left(\mathbb{I}\left(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\right)+\frac{1}{2}\mathbb{I}\left(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)\right)\right), \tag{2.21}$$

$$\mathrm{AUC} = 1-\ell_{rank}\,. \tag{2.22}$$



(b) 基于有限样例绘制的 ROC 曲线与 AUC

$$AUC = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)(y_i+y_{i+1}) = \sum_{i=1}^{m-1}\left[\frac{1}{2}\cdot(x_{i+1}-x_i)\cdot(y_i+y_{i+1})\right]$$

即AUC为ROC曲线与x轴围成的面积。

$$L_{rank} = \frac{1}{m^+m^-}\sum_{x\in D^+}\sum_{x\in D^-}\left[I\left(f(x^+)<f(x^-)\right)+\frac{1}{2}I\left(f(x^+)=f(x^-)\right)\right]$$

$$= \frac{1}{m^+m^-}\sum_{x\in D^+}\left[\sum_{x\in D^-}I\left(f(x^+)<f(x^-)\right)+\frac{1}{2}\sum_{x\in D^-}I\left(f(x^+)=f(x^-)\right)\right]$$

$$= \sum_{x^+\in D^+}\left[\frac{1}{m^+}\cdot\frac{1}{m^-}\sum_{x\in D^-}I\left(f(x^+)<f(x^-)\right)+\frac{1}{2}\cdot\frac{1}{m^+}\cdot\frac{1}{m^-}\sum_{x\in D^-}I\left(f(x^+)=f(x^-)\right)\right]$$

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

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

对ROC曲线与y轴围成的每个小梯形，每增加一个假正例的x坐标就新增一个单位 $\frac{1}{m^-}$

故上底 $= \frac{1}{m^-} \sum_{x^- \in D^-} \mathbb{I}\left(f(x^+) < f(x^-) \right)$，

下底 $= \frac{1}{m^-} \left(\sum_{x^- \in D^-} \mathbb{I}\left(f(x^+) < f(x^-) \right) + \sum_{x^- \in D^-} \mathbb{I}\left(f(x^+) = f(x^-) \right) \right)$，高y轴步长 $\frac{1}{m^+}$

$\sum_{x^+ \in D^+}$ 遍历所有小梯形。

即 ℓ_{rank} 为ROC曲线与y轴围成的面积.

$\therefore \quad AUC + \ell_{rank} = S_{ROC} = 1$

$AUC = 1 - \ell_{rank}$

2.9 试述 χ^2 检验过程.

① 提出原假设 H_0 与备择假设 H_1，检验水平设为 α

② 将总体X的取值范围K个互不相交的小区间 A_1、A_2、…、A_k

把落入第i个小区间 A_i 的样本个数记为 f_i，或为组频数(真实值)　$i=1,\cdots,k$

③ 当 H_0 为真时，根据原假设计算出总体X的值落入第i个小区间 A_i 的概率 p_i，于是 np_i 就是落入第i个小区间 A_i 的样本的理论频数(理论值)　$i=1,\cdots,k$

计算检验统计量 $\sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$，计算拒绝域，并判断结果.