assignment1

Xiaoma

2022.09.19

2.2

10 折交叉验证

由于交叉验证过程中要保证子集数据分布尽可能一致，则每次训练中正例和反例的比例都相同，那么结果随机猜测，错误率为 50

留一法

若留出正例，则训练中反例个数大于正例，预测一定错误，留出为反例同类，则错误率为 100

2.4

真正例率：$\frac{TP}{TP+FP}$，真正例占预测正例的比例。
假正例率：$\frac{FP}{FP+TN}$，真反例被预测为正例的比例。
查准率：$\frac{TP}{TP+FP}$，真正例占预测正例的比例。
查全率：$\frac{TP}{TP+FN}$，真正例被预测为正例的比例。

2.5

将 AUC 面积分割为若干个四边形，对于每一条横线或斜线，对于每一个四边形，设左侧高为 m_i^+，右侧高为 m_j^+，则

$$m_i^+ = \frac{\sum_{x_+ \in \mathcal{D}^+} \mathbb{I}(f(x^+) > f(m_j))}{m^+}$$

$$m_j^+ = \frac{\sum_{x_+ \in \mathcal{D}^+} \mathbb{I}(f(x^+) > f(m_j))}{m^+} + \frac{\sum_{x_+ \in \mathcal{D}^+} \mathbb{I}(f(x^+) = f(m_j))}{m^+}$$

$$m_j - m_i = \frac{\sum_{x^- \in \mathcal{D}^-} \mathbb{I}(f(x^-) = f(m_j))}{m^-}$$

$$S = \frac{(\sum_{x^- \in \mathcal{D}^-} \mathbb{I}(f(x^-) = f(m_j))) * (\sum_{x_+ \in \mathcal{D}^+} \mathbb{I}(f(x^+) > f(m_j)) + \frac{1}{2} \sum_{x_+ \in \mathcal{D}^+} \mathbb{I}(f(x^+) = f(m_j))}{m^- m^+}$$

$$AUC = \frac{\sum_{x_+ \in \mathcal{D}^+} \sum_{x_- \in \mathcal{D}^-} \mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))}{m^- m^+}$$

则

$$AUC = 1 - \mathcal{L}_{rank}$$

2.9

(1) 提出原假设：
H_0：总体 X 的分布函数为 $F(X)$。如果总体分布为离散型，则假设具体为
H_0：总体 X 的分布律为 $PX = x_i = p_i, i = 1, 2, \ldots$

(2) 将总体 X 的取值范围分成 k 个互不相交的小区间 $A_1, A_2, ..., A_k$，如可取区间的划分视具体情况而定，但要使每个小区间的样本值个数不小于 5，而区间个数 k 不要太大也不要太小。

(3) 把落入第 i 个区间的 A_i 的样本值的个数记作 fi，成为组频数，所有组频数之和等于样本容量 n。

(4) 当 H_0 为真时，根据所假设的总体理论分布，可算出总体 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 个小区间的样本值的理论频数。

(5) 当 H_0 为真时，n 次试验中样本值落入第 i 个小区间的频率 $\frac{f_i}{n}$ 与概率 p_i 应该很接近，当 H_0 不真时，则 $\frac{f_i}{n}$ 与概率 p_i 相差很大。