

# Part III-B: Medicine AI

Lecture by None

Note by THF

2024 年 10 月 31 日

## 目录

0.1 模型评估和性能度量 . . . . .	1
0.2 模型性能度量 . . . . .	2

## Learn 5

10.23

### 0.1 模型评估和性能度量

**Notation.** 留出法 (*hold-out*):

将原始数据集  $D$  分为两个互斥的子集  $S, T$ ,  $S$  作为训练数据集,  $T$  作为测试数据集:  $D = S \cup T, S \cap T = \emptyset$

在划分任务时要尽量保证  $S$  和  $T$  中的样本类别比例相似

**Example.**

$$D(a, b) \rightarrow S(\lambda a, \lambda b) \cup T((1 - \lambda)a, (1 - \lambda)b).$$

该过程称为分层采样法, 其中  $\lambda \in [\frac{2}{3}, \frac{4}{5}]$

使用  $S$  训练模型,  $T$  进行模型测试, 多次随机划分  $a, b$  在  $S$  和  $T$  内的内容, 多次实验取测试结果平均值

**Notation.** 交叉验证法/ $k$ 折交叉验证 (*cross validation/ $k$ -fold cross validation*):

$$D = D_1 \cup D_2 \cup \dots \cup D_k \text{ 且 } D_i \cap D_j = \emptyset (i \neq j).$$

此处  $\forall D_i$  由  $D$  分层采样得到

每次实验使用  $k-1$  个子集的并集训练，剩下的一个子集作为测试集：

$$S = \sum_{i=1}^{m-1} D_i + \sum_{i=m+1}^k D_i \quad T = D_m.$$

取不同的  $m$  值共可以得到  $k$  组“训练集-测试集”，得到  $k$  个结果，取  $k$  个结果的平均值

**Example.** 5 折交叉验证的数据划分：

$$\begin{array}{|c|c|c|c|c|} \hline D_1 & D_2 & D_3 & D_4 & \mathbf{D}_5 \\ \hline D_1 & D_2 & D_3 & \mathbf{D}_4 & D_5 \\ \hline D_1 & D_2 & \mathbf{D}_3 & D_4 & D_5 \\ \hline D_1 & \mathbf{D}_2 & D_3 & D_4 & D_5 \\ \hline \mathbf{D}_1 & D_2 & D_3 & D_4 & D_5 \\ \hline \end{array} \Rightarrow \begin{cases} Res_1 \\ Res_2 \\ Res_3 \\ Res_4 \\ Res_5 \end{cases} \xrightarrow{\text{Avg.}} \text{Result}$$

**Notation.** 若样本量  $m$  等于子集数  $k$ ，交叉验证法等同于留一法 (*leave one out*, LOO)

留一法的优点：训练结果更准确

缺点：样本量太大的时候消耗过多资源

## 0.2 模型性能度量

**Notation.** 错误率：

$$E = \frac{1}{m} N(f(x_i) \neq y_i).$$

准确率：

$$\text{Acc} = \frac{1}{m} N(f(x_i) = y_i).$$

$m$  为样本总数， $N(f(x_i) = y)$  表示符合特征  $f: x \rightarrow y$  的样本数量

## Learn 6

10.31

**Notation.** 二分类问题:

将一个样本分至两个类别的问题，如：鉴定邮件是否为垃圾邮件，预测某人是否会患上某种疾病等问题

对于二分类问题，真实结果有两种，使用模型预测也会产生两种结果，组合得到混淆矩阵：

$$\begin{bmatrix} \text{真阳性 (TP)} & \text{假阴性 (FN)} \\ \text{假阳性 (FP)} & \text{真阴性 (TN)} \end{bmatrix}.$$

其中：阳性/阴性为模型预测结果，真/假为真实结果

准确率 (Acc) 根据混淆矩阵的计算：

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

**Notation.** 马修斯相关系数 (Matthews Correlation Coefficient, MCC):

MCC 比 Acc 更加全面（正负数据不平衡）

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \in [-1, 1].$$

**Notation.** MCC 结果解读：

- $\text{FP} = \text{FN} = 0$ ：无误判结果，代入得： $\text{MCC} = 1$ ，表示模型完美
- $\text{TP} = \text{TN} = 0$ ：全部误判，代入得： $\text{MCC} = -1$ ，表示最差
- $\text{TP} \times \text{TN} = \text{FP} \times \text{FN}$ ，即  $\text{MCC} = 0$ ，表示模型完全随机判断

当样本中阴性样本远少于阳性样本时，Acc 计算不能涉及到假阴性与假阳性而 MCC 可以

若第一个模型对阳性和阴性样本判断接近，而第二个模型对阳性样本表现极佳但对阴性样本表现极差，则  $\text{MCC}_1 > \text{MCC}_2$ ，而 Acc 可能接近