

Part III-B: Medicine AI

Lecture by None

Note by THF

2024 年 11 月 2 日

目录

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 $MCC_1 > MCC_2$ ，而 Acc 可能接近

Notation. 查准率 P ，查全率 R ， F_1 度量：

◦ 查准率 (precision, P)：又叫精确率

$$\begin{aligned} P &= \frac{N_{TP}}{N_{P_p}} \\ &= \frac{TP}{TP + FP}. \end{aligned}$$

Notation. ◦ 查全率 (recall, R)：又叫召回率

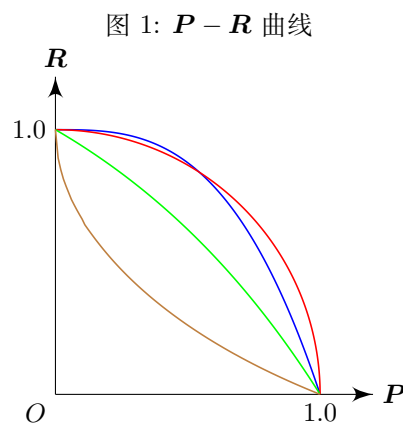
$$\begin{aligned} R &= \frac{N_{TP}}{N_{P_a}} \\ &= \frac{TP}{TP + FN}. \end{aligned}$$

一般情况下：查全率和查准率相矛盾

Learn 7

11.01

Notation. 在模型下，对样本阳/阴性的预测结果为一个概率 $p \in [0, 1]$ ，通过设定一个阈值 m 来区分由模型预测的结果；在该阈值下，计算查全率和查准率，绘制一个点；设定不同的阈值，将所有点连接，得到 $P-R$ 曲线



Learn 7