第七讲：机器学习基础

1.1 论述 True Positive Rate、False Positive Rate 与 Precision（查准率）、Recall（查全率）的联系。

一个真阳性（True Positive）是模型正确预测到阳性类别的结果。

同样地，一个真阴性（True Negative）是模型正确预测到阴性类别的结果。

一个假阳性（False Positive）是模型错误地预测到阳性类别的结果。

而一个假阴性（False Negative）是模型错误地预测到阴性类别的结果。

1.2 Precision（查准率）：

精确度的公式如下：

精确度 = TP/TP+FP

我们看一个例子把数据为：

- 真阳性（TP）：1
- 假阳性（FP）：1
- 假阴性（FN）：8
- 真阴性（TN）：90

精确度的计算将是：

精确度 = TP/TP+FP = 1/1+1 = 0.5

这意味着该模型的精确度为 0.5，表明当模型预测肿瘤为恶性时，它的正确率是 50%。

一个不产生任何假阳性的模型具有 1.0 的精确度。

精确度衡量的是被正确识别的阳性结果的比例。

精确度试图回答以下问题：

阳性识别的哪一部分实际上是正确的？

1.3 Precision （精确率）：

准确率是评估分类模型的一种指标。非正式地说，准确率是我们的模型做出正确预测的比例。准确率的正式定义是：

精确率 = TP+TN/TP+TN+FP+FN

其中 TP = 真阳性，TN = 真阴性，FP = 假阳性，FN = 假阴性

注：一个不产生任何假阴性的模型具有 1.0 的精确率。

我们看一个例子把数据为：

- 真阳性（TP）：1
- 假阳性（FP）：1
- 假阴性（FN）：8
- 真阴性（TN）：90

精确率 = 1+90/1+1+8+90 = 0.91 = 91% 百分点

1.4.Recall（查全率）的联系:

召回率试图回答以下问题：

正确识别的实际阳性占多大比例？

从数学角度定义，召回率如下：

Recall = TP/TP+TN

注：一个不产生任何假阴性的模型具有 1.0 的召回率。

让我们为我们的肿瘤分类器计算召回率：

- 真阳性（TP）：1
- 假阳性（FP）：1
- 假阴性（FN）：8
- 真阴性（TN）：90

召回率的计算是：

Recall = 1/1+90 = 0.11 = 11%百分点

2. 假阴性在新冠肺炎检测中出现的主要原因？

新冠肺炎检测中假阴性出现的主要原因可能包括：

样本收集的不当（如拭子采集技术不正确）

样本中病毒含量低于检测限（尤其是在感染的早期或晚期）

样本处理或运输过程中的污染或破坏

检测试剂或设备的灵敏度和特异性不足

个体生物学差异，如个人免疫反应

3. 数据集包含 1000 个样本，其中 500 个正例，500 个负例，将其划分为 70%训练，30%测试的留出法，估算多少种划分方式（给出式子，不需要具体计算数值）。

对于数据集包含 1000 个样本，500 个正例和 500 个负例，使用 70%训练集和 30%测试集的留出法，划分方式的估算可以使用组合数学中的组合公式来表示：

$$= \binom{500}{350} * \binom{500}{150}$$

其中 C（n，k）表示从 n 个不同元素中不重复地抽取 k 个元素的组合数。 这里我们分别计算正例和负例分别取出 70%作为训练样本的组合数，然后将两者相乘即为总的划分方式数。