机器学习基本概念

1.训练过程

训练分为：训练集，测试集，验证集

为什么要测试集?

答:

1. **模型性能评估：** 测试集用于评估机器学习模型的性能。在训练阶段，模型学习了数据的模式和关系。测试集包含了模型之前未见过的数据，因此可以用来测试模型对新数据的泛化能力。这有助于确定模型是否能够在实际应用中表现良好。
2. **防止过拟合：** 测试集有助于检测模型是否过拟合训练数据。如果模型在训练数据上表现良好但在测试数据上表现糟糕，这可能是过拟合的迹象。测试集可以帮助识别这种情况，从而有助于优化模型，使其在未来数据上更稳健。
3. **超参数调优：** 在模型训练之前，通常需要选择一些超参数（如学习率、树的深度等）。测试集用于评估不同超参数组合的性能，帮助选择最佳超参数设置。
4. **模型选择：** 当有多个候选模型可供选择时，测试集用于比较它们的性能，以确定哪个模型最适合特定任务。
5. **验证模型假设：** 测试集可用于验证关于数据的模型假设。例如，如果模型假设特定特征与目标之间存在线性关系，可以使用测试集来验证这种假设是否成立。

测试集怎么保留?

K折交叉验证： K折交叉验证是一种通过多次划分数据来评估模型性能的方法。数据集被分成K个折，然后模型被训练和测试K次，每次使用一个不同的折作为测试集，其余的作为训练集。最后，结果平均以获得更稳健的性能评估。

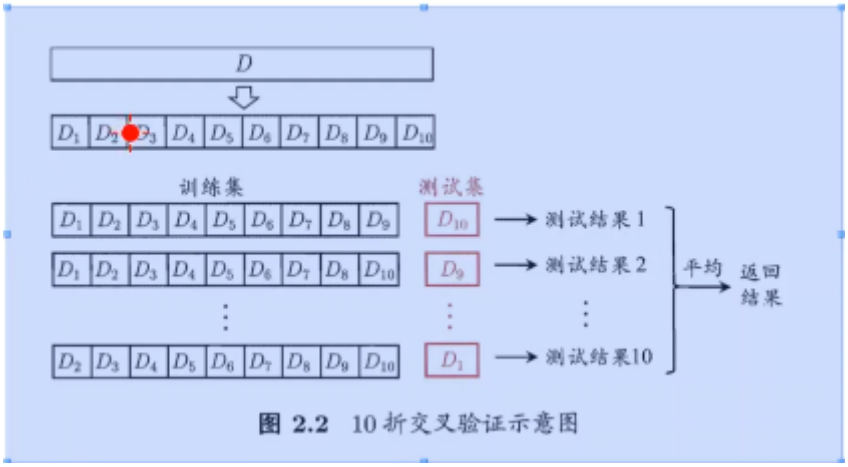


图 2.2 10 折交叉验证示意图

留一法交叉验证： 留一法是K折交叉验证的一个特例，其中K等于数据点的数量。每个数据点被用作一次测试集，其余的作为训练集。这对于小型数据集非常有用，但计算成本较高。

留出验证： 留出验证是将数据集分成训练集和验证集的方法。通常，验证集用于调整超参数，而测试集用于最终性能评估。这可以避免在超参数调优中泄漏有关测试集的信息。

验证集有什么用，怎么划分?

验证集用于在模型训练过程中评估模型性能、进行超参数调整和模型选择。比如需要多次调参的训练中就会划分验证集，进行超参数选择，等到结果不错时，再使用测试集最终评估模型的性能。

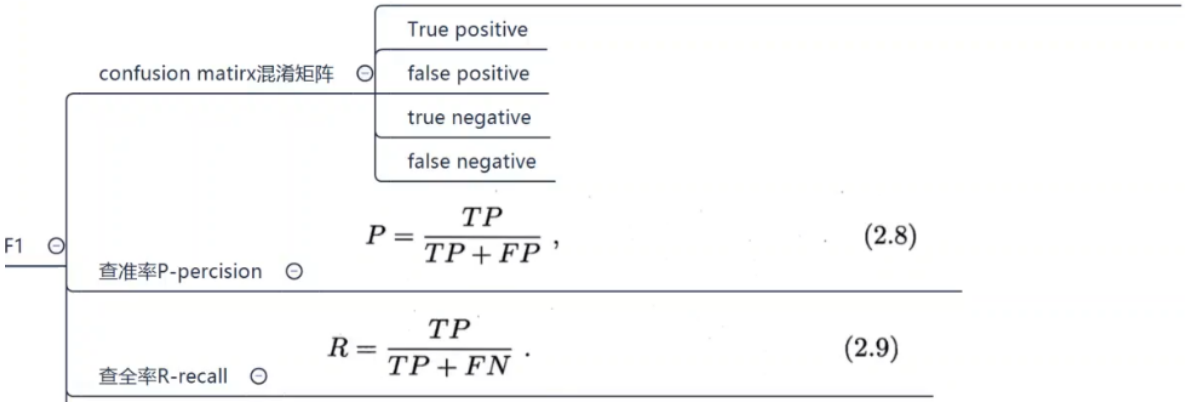
- 训练集：约60-80%的数据用于模型的训练。
- 验证集：约10-20%的数据用于模型性能评估和超参数调整。
- 测试集：约10-20%的数据用于最终性能评估，确保模型未在测试前接触到这些数据。

2.评估指标

查准率与查全率：

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)



$$P = \frac{TP}{TP + FP}, \tag{2.8}$$

$$R = \frac{TP}{TP + FN}. \tag{2.9}$$

F1 Score

针对精准率和召回率都有其自己的缺点；如果阈值较高，那么精准率会高，但是会漏掉很多数据；如果阈值较低，召回率高，但是预测的会很不准确。

例子一

假设总共有10个好苹果，10个坏苹果。针对这20个数据，模型只预测了1个好苹果，对应结果如下表：

	预测为好苹果	预测为坏苹果
标签为好苹果	1	9
标签为坏苹果	0	10

$$Precision = \frac{1}{1+0} = 1$$

$$Recall = \frac{1}{1+9} = 0.1$$

虽然精确率很高，但是这个模型的性能并不好。

例子二

同样总共有10个好苹果，10个坏苹果。针对这20个数据，模型把所有的苹果都预测为好苹果，对应结果如下表

	预测为好苹果	预测为坏苹果
标签为好苹果	10	0
标签为坏苹果	10	0

$$Precision = \frac{10}{10+10} = 0.5$$

$$Recall = \frac{10}{10+0} = 1$$

虽然召回率很高，但是这个模型的性能并不好。

从上述例子中，可以看到精确率和召回率是此消彼长的，如果要兼顾二者，就需要F1 Score。

$$F1 = \frac{2 \times P \times R}{P+R}$$

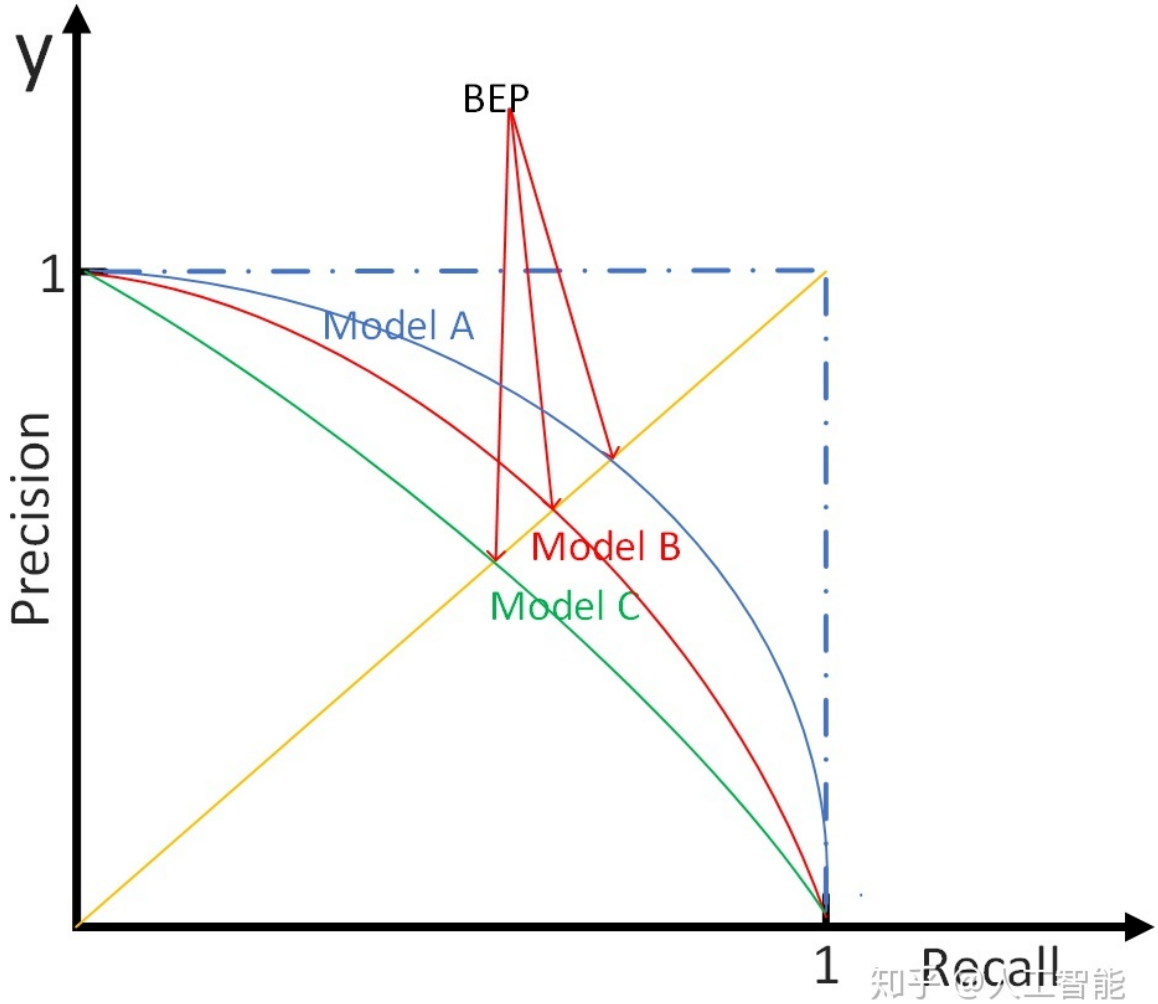
F1 Score是一种**调和平均数。**

P-R曲线

P-R曲线是描述精确率和召回率变化的曲线。对于所有的正样本，

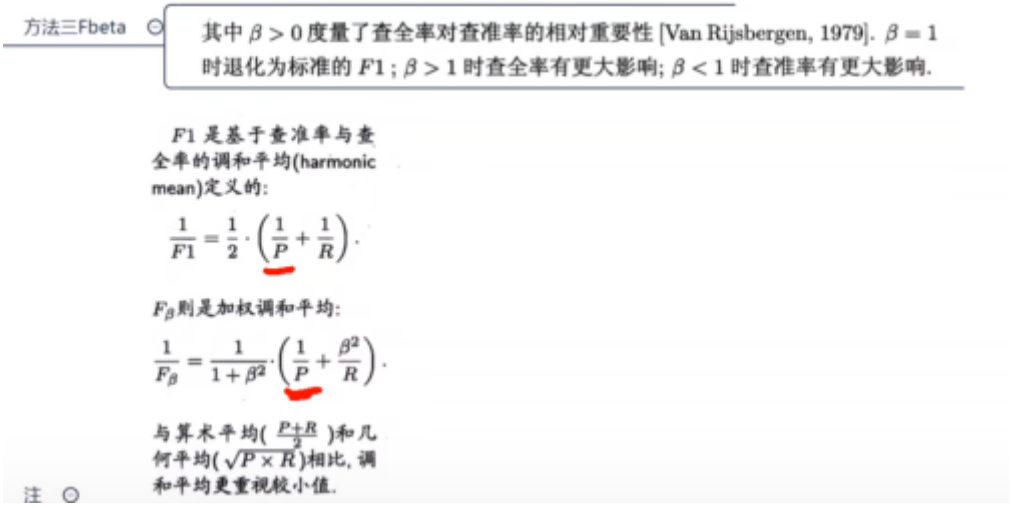
绘制P-R曲线?

设置不同的阈值，模型预测所有的正样本，计算对应的精准率和召回率。



模型与坐标轴围成的面积越大，则模型的性能越好。但一般来说，曲线下的面积是很难进行估算的，所以衍生出了“**平衡点**”（Break-Event Point，简称**BEP**），即当P=R时的取值，平衡点的取值越高，性能更优。

F_beta

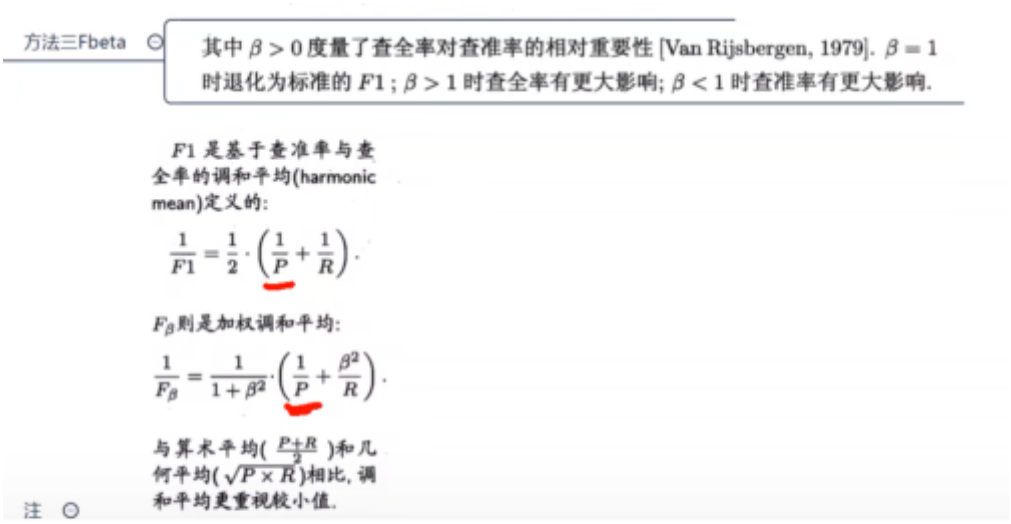


ROC与AUC

1.为什么会有ROC?

例子三

有好苹果9个，坏苹果1个，模型把所有的苹果均预测为好苹果。



$$Accuracy = \frac{9}{9+1} = 0.9$$
$$Precision = \frac{9}{9+1} = 0.9$$
$$Recall = \frac{9}{9+0} = 1$$
$$F1 = \frac{2 \times P \times R}{P+R} = \frac{2 \times 0.9 \times 1}{1+0.9} = \frac{1.8}{1.9} \approx 1$$

我们能够得出，尽管 Precision、Recall、F1都很高，但是模型效果却不好**（对坏苹果分类效果）**。所以针对样本不均衡，以上指标很难区分模型的性能，就需要用到ROC和AUC。

预测标签 / 真实标签	正例	反例
正例	TP（真正类）	FN（假反类）
反例	FP（假正类）	TN（真反类）

在介绍ROC和AUC之前，我们需要明确以下三个概念：

真正类率（true positive rate, TPR），也称为**灵敏度(sensitivity)**，等同于召回率。刻画的是被分类器正确分类的正实例占所有正实例的比例。

$$TPR = \frac{\text{正样本预测正确量}}{\text{正样本总量}} = \frac{TP}{TP+FN}$$

真负类率（true negative rate, TNR），也称为**特异度(specificity)**，刻画的是被分类器正确分类的负实例占所有负实例的比例。

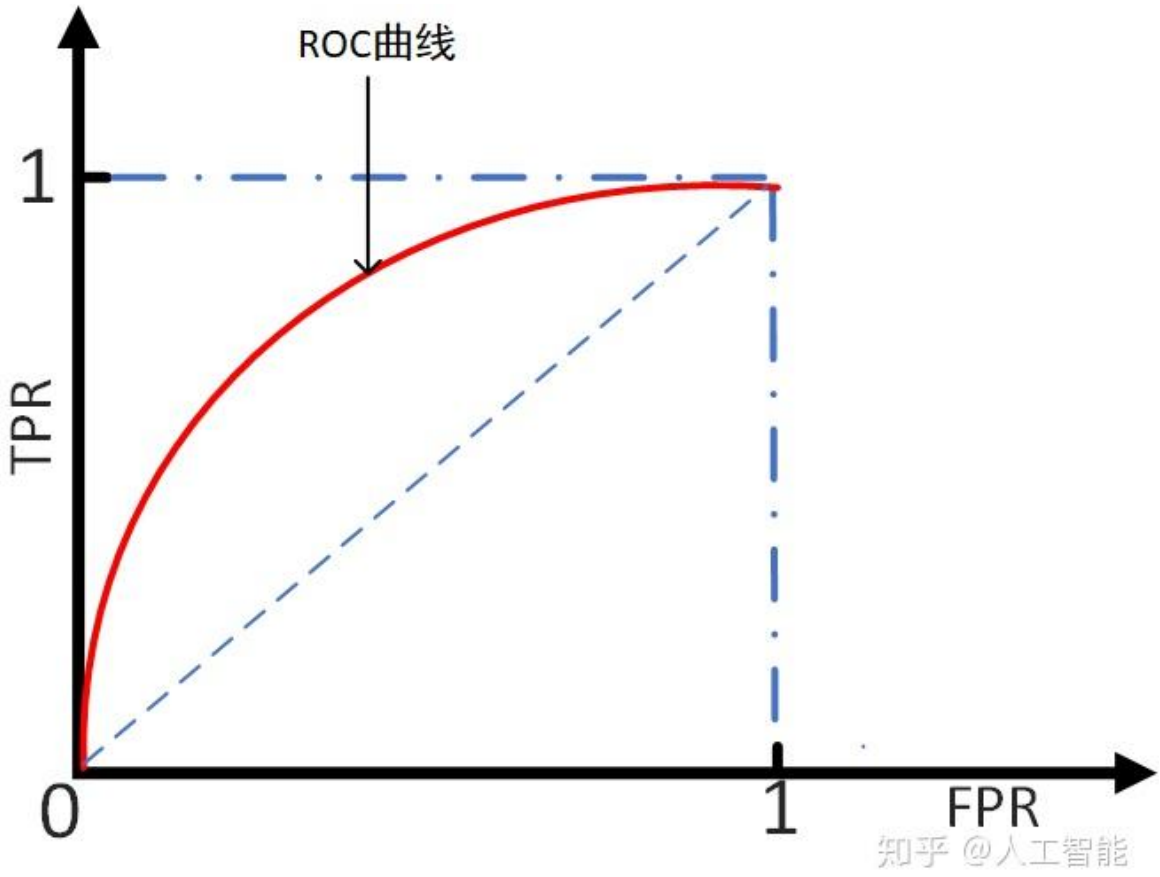
$$TNR = \frac{\text{负样本预测正确量}}{\text{负样本总量}} = \frac{TN}{FP+TN}$$

负正类率（false positive rate, FPR），也称为1-specificity，计算的是被分类器错认为正类的负实例占所有负实例的比例。

$$FPR = 1 - TNR = \frac{\text{负样本预测错误量}}{\text{负样本总量}} = \frac{FP}{FP+TN}$$

ROC曲线

ROC（Receiver Operating Characteristic）曲线，又称接受者操作特征曲线。曲线对应的纵坐标是TPR，横坐标是FPR。



理想目标： TPR=1, FPR=0，即图中(0,1)点。故ROC曲线越靠拢(0,1)点，即，越偏离45度对角线越好。对应的就是TPR越大越好，FPR越小越好。

AUC

AUC(Area Under Curve)是处于ROC曲线下方的那部分面积的大小。AUC越大，代表模型的性能越好。

对于例子三的样本不均衡，对应的TPR=1，而FPR=1，能够判断模型性能不好。