# 评估指标的局限性

## 场景描述

模型评估过程中，分类问题、排序问题、回归问题往往需要使用不同的指标进行评估。在诸多的评估指标中，不仅不能发现模型本身的问题，而且会得出错误的结论。

## 问题1：准确率的局限性



准确率在分类问题中最简单也是最直观的评价指标，但存在明显的缺陷。比如，当负样本占99%时，分类器把所有样本都预测为负样本也可以获得99%的准确率。所以，当不同类别的样本不均衡的时候，占比大的类别往往成为影响准确率的最主要因素。

## 问题二：准确率与召回率的权衡

准确率是指分类正确的正样本个数占分类器判定为正样本的样本个数的比例；召回率是指分类正确的正样本个数占正真的正样本个数的比例。

在排序问题中，通常没有一个确定的阈值把得到的结果直接判定为正样本或负样本，而是采用TopN返回结果的Precision值和Recall值来衡量排序模型的性能，即认为模型返回的Top N的结果就是模型判定的正样本，然后计算前N个位置上的准确率Precision@N和前N个位置上的召回率Recall@N。

Precision值和Recall值即矛盾又统一的两个指标，为了提高Precision值，分类器需要尽量在‘更有把握’时才把样本预测为正样本，但此时往往会因为过于保守而漏掉很多‘没有把握’的正样本，导致Recall值降低。

PR曲线：横轴表示召回率，纵轴表示精确率。对于一个排序模型来说，其P-R曲线上的一个点代表着，在某一个阈值下，模型将大于该阈值的结果判定为正样本，小于该阈值的结果判定为负样本，此时返回结果对应的召回率和准确率。整条曲线是通过阈值从高到低移动而生成的。

F1-score和Roc曲线也能综合第反应一个排序模型的性能，F1 score是精确率和召回率的调和平均值，它定义为



## 问题3：平方根误差的‘意外’

一般情况下，RMSE能够很好地反应回归模型预测值与真实值的偏离程度，但实际问题中，如果存在个别偏离程度非常大的离群点时，即使离群点数量少，也会让RMSE指标变得很差。

**解决方案：**1.如果我们认定这些离群点是噪声点的话，就需要在数据预处理的阶段把这些噪声点过滤掉。2. 如果不认为这些离群点时噪声点的话，就需要进一步提高模型的预测能力，将离群点产生的机制建模进去。3. 可以找一个更合适的指标来评估模型。

# ROC曲线

## 场景描述

二值分类器是机器学习领域中最常见也是应用最广泛的分类器。评价二值分类器的指标很多，比如precision、recall、F1 score、P-R曲线。上面介绍的评估指标智能反应模型在某一个方面的性能。相比而言，ROC曲线则有很多的优点，经常作为评估二分类最重要的指标之一。

## 问题1：上面是ROC曲线

ROC曲线是Receiver Operating Characteristic Curve的简称（受试者工作特征曲线）。

ROC曲线的横坐标为假阳性（False Positive Rate, FPR）；纵坐标为真阳性（True Positive Rate, TPR）。



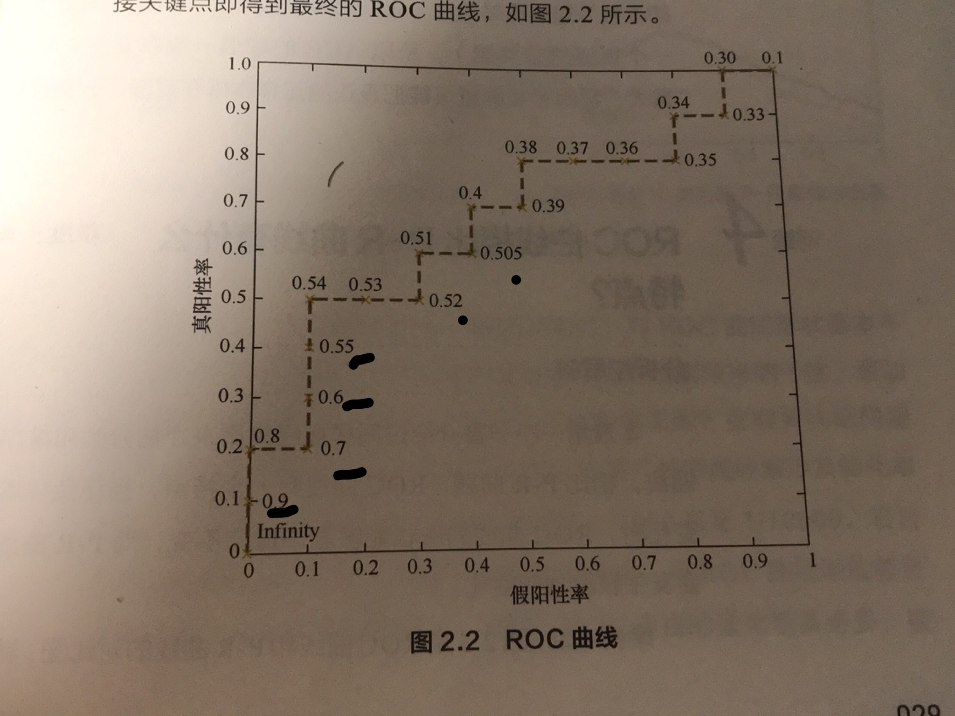


## 问题2：如何绘制ROC曲线

ROC曲线是通过不断移动分类器的‘截断点’来生成曲线上的一组关键点的。

在二值分类问题中，模型输出一般都是预测样本为正例的概率，样本按照概率从高到低排序，在输出最终的正例、负例之前，我们指定一个阈值，预测概率大于阈值的样本会被判定为正例；

每一个截断点对应一个FPR和TPR，在ROC图上绘制出每个截断点对应的位置，再连接所有点就得到最终的ROC曲线。

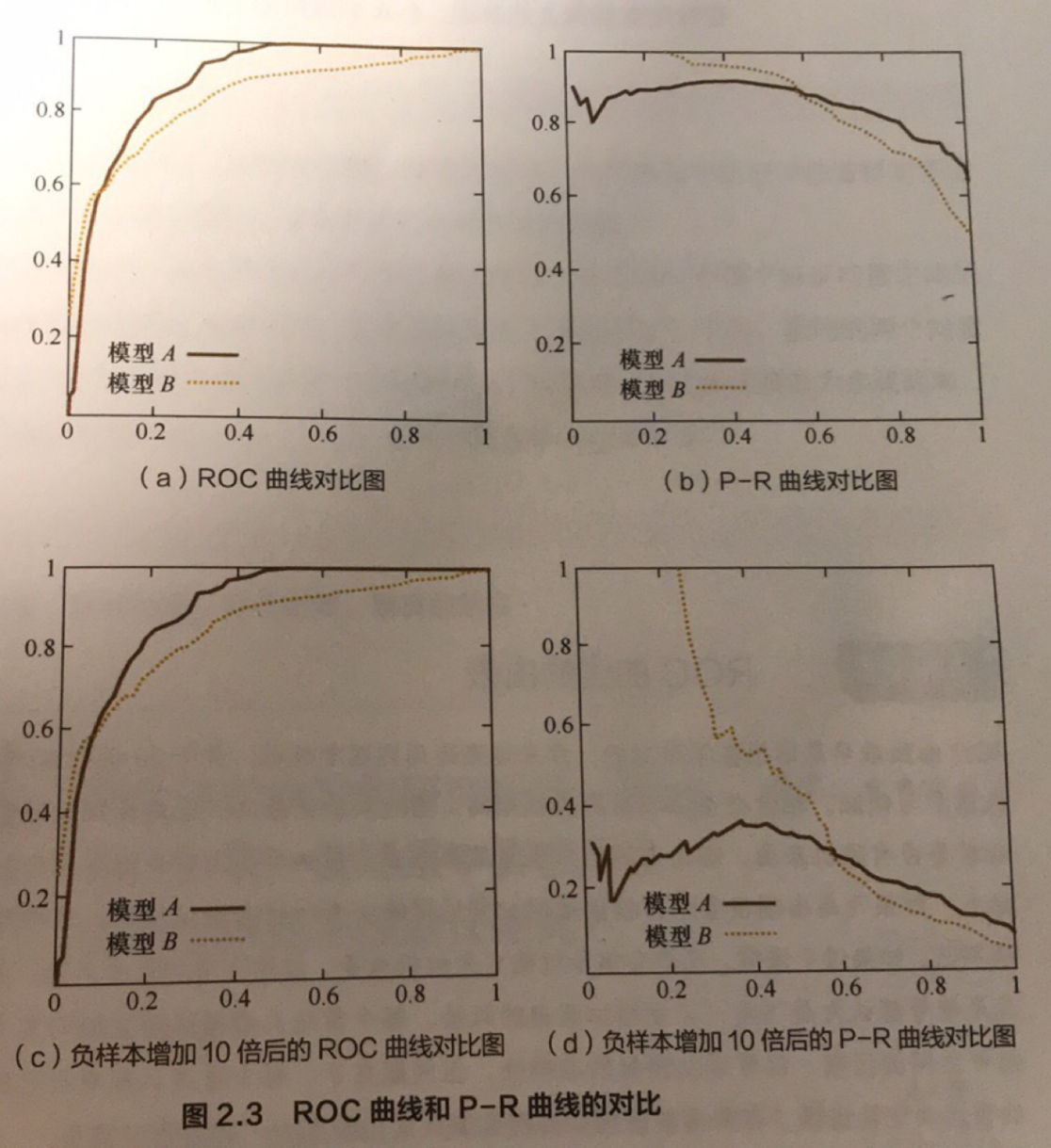


## 问题3：如何计算AUC

AUC指的是ROC曲线下面的面积，该值能够量化地反映基于ROC曲线衡量出的模型性能，计算AUC值只需要沿着ROC横轴做积分就可以了。取值一般在[0.5, 1]，AUC越大，说明分类器越可能把真正的正样本排在前面，分类性能越好。

## 问题4：ROC曲线相比PR曲线有什么特点

当负样本的分布发生变化时，相比PR曲线，ROC曲线的形状能够基本保持不变，而PR曲线的形状一般会发生较剧烈变化。



# 余弦距离的应用

## 场景描述：

在模型训练的过程中，我们也在不断地评估者样本间的距离，如何估计样本距离也是定义优化目标和训练方法的基础。

在机器学习中，通常将特征表示为向量的形式，所以在分析两个特征向量之间的相似性时，常使用余弦相似度来表示。余弦相似度范围[0, 1]，将1减去余弦相似度就是余弦距离，[0，2]。

## 问题1：结合你学习和研究经历，探讨为什么在一些情景中要使用余弦相似度而不是欧式距离



对于图像，研究对象的特征维度往往很高，余弦相似度在高维情况下依然保持“相同时为1，相交时为0，相反时为-1,”的性质，而欧式距离的数值则受维度的影响，范围不固定，并且含义比较模糊。

欧式距离体现**数值上**的**绝对差异**，余弦距离体现**方向**上的**相对差异**。

## 问题2：余弦距离是否是在一个严格定义的距离？

# A/B测试的陷阱

## 场景描述

在互联网公司中，A/B测试时验证新模块、新功能、新产品是否有效，新算法、新模型的效果是否有提升，新设计是否受到用户欢迎，新更改是否影响用户体验的主要测试方法。在机器学习领域，A/B测试时验证模型最终效果的主要手段。

## 问题1：在对模型进行过分的离线评估之后，为什么还要进行在线A/B测试?

（1）离线评估无法完全消除模型过拟合的影响，因此，得出的离线评估模型结果无法完全替代线上评估结果。

（2）离线评估无法完全还原线上的工程环境。一般来讲，离线评估往往不会考虑线上环境的延迟、数据丢失、标签数据缺失等情况，因此离线评估的结果是理想环境工程环境下的结果。

（3）线上系统的某些商业指标在离线模型中无法计算。离线评估一般是针对模型本身进行评估，而与模型相关的其他指标，特别是商业指标，往往

# 模型评估的方法

## 问题1：在模型评估过程中，有哪些主要的验证方法，它们的优缺点是什么

* **Holdout检验**：Holdout检验是最简单的也是最直接的验证方法，它将原始的样本集合随机划分为验证集和测试集两部分。**缺点**很明显，验证集上计算出来的最后模型指标与原始分组有很大关系。
* **交叉验证**：
  + **k-fold交叉验证**：首先将全部样本划分成k个大小相等的样本子集，依次遍历这k个子集，每次把当前子集作为验证集，其余所有子集作为训练集，进行模型的训练和评估，最后把k次评估指标的平均值作为最终的评估指标。
  + **留一验证**：每次留下1个样本作为验证集，其余所有样本作为测试集。样本总数为n，依次对n个样本进行遍历，进行n次，再将评估指标求平均值得到最终的评估指标。在样本总数较多的的情况下，留一验证法的时间开销极大。
* **自助法：**不管是Holdout检验还是交叉检验，都是基于划分验证集和测试集的方法进行模型评估的，然而，当样本规模比较小时，将样本集进行划分会让训练进一步减少，这可能会影响训练效果。自助法是基于自助采样法的检验方法，得到大小为n的训练集，n次采样过程中，有的样本会被重复采样，有的样本没有被抽样过，将这些没有被抽样过的样本作为验证集，进行模型验证。

## 问题2：在自助法的采样过程中，对n个样本进行n次自主抽样，当n趋于无穷大时，最终有多少数据从未被选择过

一个样本在一次抽样过程中未被抽样的概率是,n次抽样中未被抽到，当n趋于无穷大时，概率