# 评估指标

# 分类任务

## 欠拟合与过拟合

我们实际希望的，是在新样本上能表现得很好的学习器.为了达到这个目的，应该从训练样本中尽可能学出适用于所有潜在样本的"普遍规律"，这样才能在遇到新样本时做出正确的判别.然而，当学习器把训练样本学得"太好"了的时候，很可能巳经把训练样本自身的一些特点当作了所有潜在样本都会具有的一般性质，这样就会导致泛化性能下降这种现象在机器学习中称为"过拟合" (overfitting). 与"过拟合"相对的是"欠拟合" (underfitting) ，这是指对训练样本的一般性质尚未学好.图1.1 给出了关于过拟合与欠拟含的一个便于直观理解的类比.

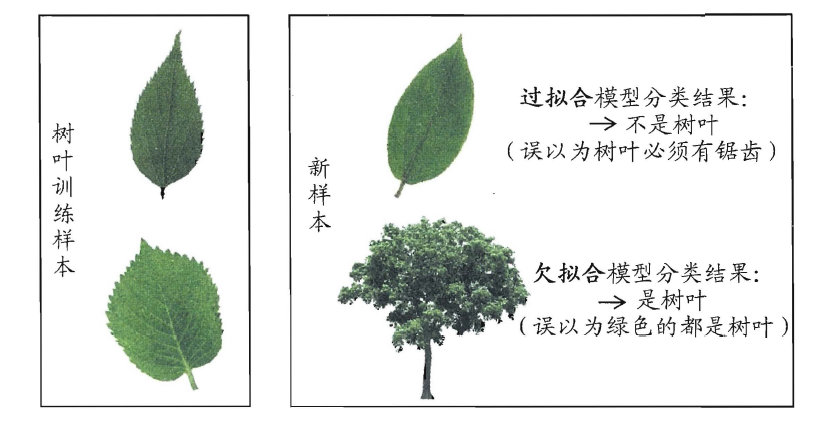


图1.1 过拟合与欠拟合（**源自机器学习周志华**）

有多种因素可能导致过拟合，其中最常见的情况是由于学习能力过于强大，以至于把训练样本所包含的不太一般的特性都学到了，而欠拟合则通常是由于学习能力低下而造成的欠拟合比较容易克服，例如在决策树学习中扩展分支、在神经网络学习中增加训练轮数等，而过拟合则很麻烦.在后面的学习中我们将看到，过拟合是机器学习面临的关键障碍，各类学习算法都必然带有一些针对过拟合的措施;然而必须认识到，过拟合是无法彻底避免的，我们所能做的只是"缓解'气或者说减小其风险。

## 1.2准确度指标

### 1.2.1混淆矩阵（Confusion Matrix）

混淆矩阵也称误差矩阵，是表示精度评价的一种标准格式，用n行n列的矩阵形式来表示。

对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划分为真阳例(true positive，TP) 、假阳例(false positive， FP) 、真阴例(true negative， TN) 、假阴例(false negative， FN) 四种情形，令TP 、FP 、TN 、FN 分别表示其对应的样例数，则显然有TP+FP+TN+FN=样例总数.分类结果的"混淆矩阵" (confusion matrix) 如图1.2所示



图1.2 混淆矩阵

#### 1.TP,FP,FN,TN

TP（True Positives）：

预测结果是True（对），真实值为Positives（正例），模型认为它也是Positives（正例）；在一般情况下，会认为IOU>0.5时为真正例，在一些数据集上的指标也是为0.5；

FP（False Positives）

预测结果是False（错），真实值为Negative（负例），模型却认为它是Positives（正例）；同理，一般情况下会认为 IOU<0.5 时为假正例；

FN（ False Negatives）

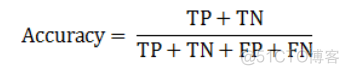
预测结果是False（错），真实值为Positives（正例），模型预测它为Negative（负例）；

TN（True Negatives）

预测结果为True（对），真实值为Negative（负例），模型预测它为Negative（负例）；一般用不到这个

#### 2.准确率（精度）（Accuracy）

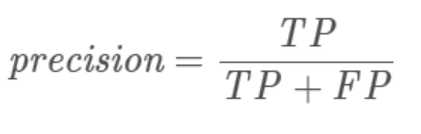
即表示分类模型所判断的所有结果中，预测正确的结果占比，公式为



#### 3.精准率或者查准率（precision）

**预测为正的样本中真正为正的的比例，**计算公式为：

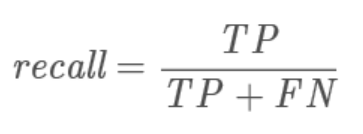
预测结果中，预测结果中真阳占预测的所有阳{真阳+假阳（原本是真阴被预测为阳）}的比例



#### 4.召回率（Recall）

**所有正样本中被正确预测的比例**，计算公式为：

预测的结果中，真阳占{真阳+假阴（原本是阳被预测为阴）}的比例。

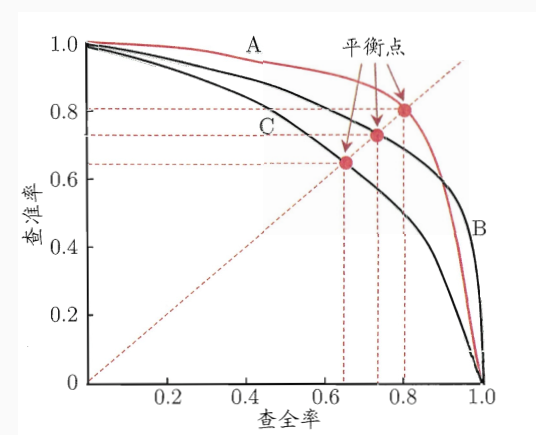


在一些特定场景中，会将其作为模型评估标准；

#### 5.平均精度 AP（Average Precision）

P-R曲线PR曲线（红线）以下与横轴、纵轴之间的面积。PR曲线是由Precision（精准率或者查准率）与Recall（召回率或者查全率）构成的曲线，横轴为Recall，纵轴为Precision。

假设有100张图片，要分成猫，狗，鸡三类，100张图片对应100个真实值，模型分类后我们会得到对应的100个预测值。这里我们可以只取前10个预测值出来，计算10个值中猫预测出几张，预测对几张，从而能计算出猫的precison和recall；接着我们可以取前20个预测值同样能计算出一组猫的precison和recall；这样一直增加到取100个预测值，就能得到猫的10组（recall, precision）值来绘制曲线。这里要注意的是：随着选取预测值增加，recall肯定是增加或不变的（选取的预测值越多，预测出来的猫越多，即查全率肯定是在增加或不变），若增加选取预测值后，recall不变，一个recall会对应两个precison值，一般选取较大的那个precision值。 如果我们每次只增加一个预测值，就会得到大约100对（recall， precisoin）值，然后就能绘制猫的PR曲线，计算出其下方的面积，就是猫对应的AP值(Average Precision)。 如果我们接着对狗和鸡也采用相同方法绘制出PR曲线， 就能得到猫，狗， 鸡三个AP值，取平均值即得到了整个模型最终的mAP（mean Average Precsion）。如下图中A， B， C三条PR曲线



#### 6. mAP（mean Average Precision）

AP衡量的是对一个类检测好坏，mAP就是对多个类的检测好坏。在多类多目标检测中，计算出每个类别的AP后，再除以类别总数，即所有类别AP的平均值，比如有两类，类A的AP值是0.5，类B的AP值是0.2，那么mAP=（0.5+0.2）/2=0.35。

F1 score

#### 7.ROC曲线

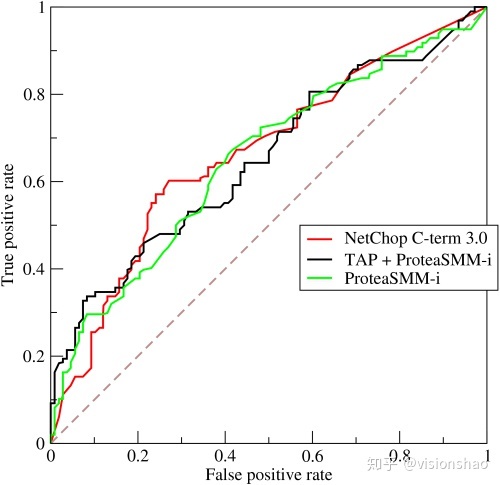
ROC曲线（Receiver Operating Characteristic）全称：受试者工作特征曲线

提到ROC曲线就要先说明一下两个概念：FPR（假阳率），TPR（真阳率）它们都是对分类任务的一个评测指标。

FPR（False Positive Rate）= FP /（FP + TN），即负类数据被分为正类的比例

TPR（True Positive Rate）= TP /（TP + FN），即正类数据被分为正类的比例

ROC曲线如下，横坐标是假阳率，纵坐标是真阳率：



#### 8.AUC曲线

AUC，（Area Under Curve），被定义为ROC曲线下的面积，显然这个面积小于1，又因为ROC曲线一般都处于y=x这条直线的上方，所以AUC一般在0.5到1之间。使用AUC值作为评价标准是因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应AUC更大的分类器效果更好。

AUC的含义为，当随机挑选一个正样本和一个负样本，根据当前的分类器计算得到的score将这个正样本排在负样本前面的概率。

从AUC判断分类器（预测模型）优劣的标准：

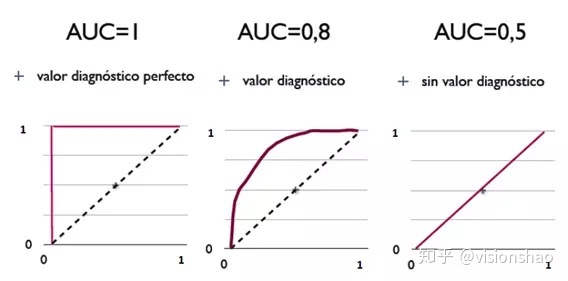
AUC = 1，是完美分类器，采用这个预测模型时，存在至少一个阈值能得出完美预测。绝大多数预测的场合，不存在完美分类器。

0.5 < AUC < 1，优于随机猜测。这个分类器（模型）妥善设定阈值的话，能有预测价值。

AUC = 0.5，跟随机猜测一样（例：丢铜板），模型没有预测价值。

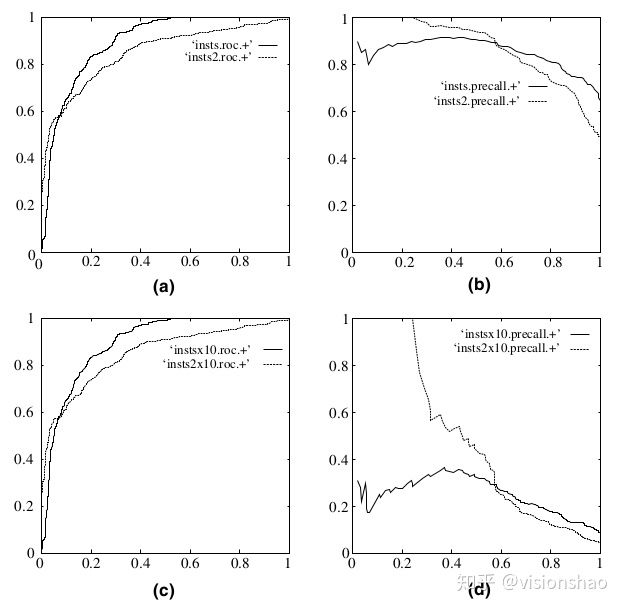
AUC < 0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。

Kappa系数



#### 3.为什么使用ROC曲线？

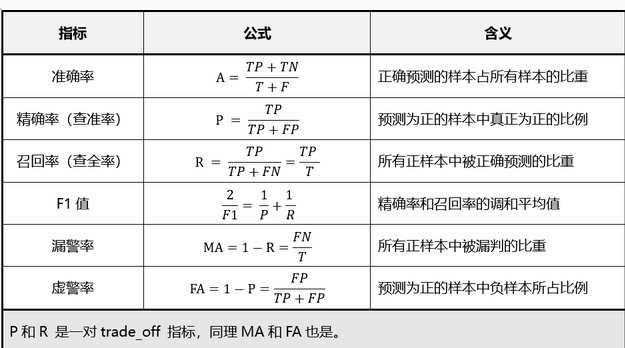
既然已经这么多评价标准，为什么还要使用ROC和AUC呢？因为ROC曲线有个很好的特性：当测试集中的正负样本的分布变化的时候，ROC曲线能够保持不变。在实际的数据集中经常会出现类不平衡(class imbalance)现象，即负样本比正样本多很多(或者相反)，而且测试数据中的正负样本的分布也可能随着时间变化。下图是ROC曲线和Precision-Recall曲线的对比：



在上图中，(a)和(c)为ROC曲线，(b)和(d)为Precision-Recall曲线。(a)和(b)展示的是分类其在原始测试集(正负样本分布平衡)的结果，(c)和(d)是将测试集中负样本的数量增加到原来的10倍后，分类器的结果。可以明显的看出，ROC曲线基本保持原貌，而Precision-Recall曲线则变化较大。

#### 4.总结

用一张图总结上面所说的



### 1.3性能指标

单张样本平均处理时间/FPS

单张样本处理时间方差

## 2目标检测任务

### 2.1准确度指标

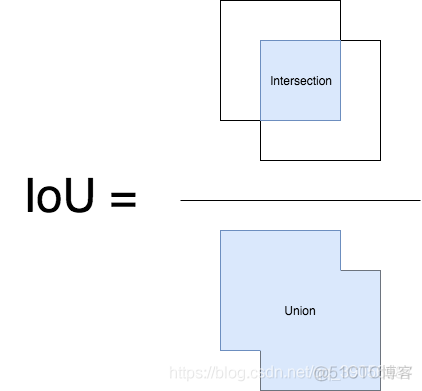
分类准确度指标

AP50，AP75，APS，APM，APL，mAP，mMR（mean Miss Rate）。

2.2.1 IOU计算

IoU 的全称为交并比（Intersection over Union），在目标检测中即计算预测边界框与真实边界框的重叠程度，重叠程度越高，说明越接近真实框。IoU 计算的是 “预测的边框” 和 “真实的边框” 的交集和并集的比值，即公式等于：

IOU = 相交的面积 / 相并的面积



#### 2.2.1 AP

在目标检测中还有一个IoU（交并比、Intersection over Union、IoU）, 通过比较检测bbox和真实bbox的IoU来判断是否属于TP（True Positive），例如设置IoU阈值为**0.7**，则IoU大于0.7的则判定为TP，否则为FP。因此当我们设置不同的IoU阈值时，也会得到不同的mAP值，再将这些mAP值进行平均就会得到mmAP，一般不做特别说明mmAP即指通常意义上的mAP。

因此目标检测mAP计算方法如下：给定一组IOU阈值，在每个IOU阈值下面，求所有类别的AP，并将其平均起来，作为这个IOU阈值下的检测性能，称为mAP(比如mAP@0.5就表示IOU阈值为0.5时的mAP)；最后，将所有IOU阈值下的mAP进行平均，就得到了最终的性能评价指标：mmAP。

### 2.2性能指标

单张样本平均处理时间/FPS

单张样本处理时间方差

## 3关键点检测任务

### 3.1准确度指标

PCK

PCKh-0.5

OKS（Object Keypoint Similarity）

AP

mAP

Recall

### 3.2性能指标

单张样本平均处理时间/FPS

单张样本处理时间方差