

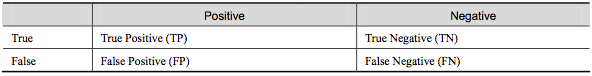
混淆矩阵

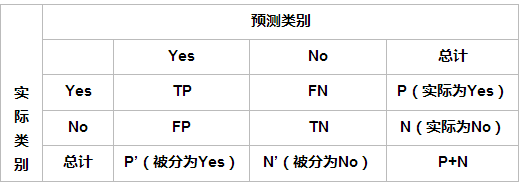
True Positive(真正，TP)：将正类预测为正类数

True Negative(真负，TN)：将负类预测为负类数

False Positive(假正，FP)：将负类预测为正类数误报 (Type I error)

False Negative(假负，FN)：将正类预测为负类数→漏报 (Type II error)





1.准确率（Accuracy）



注：准确率是我们最常见的评价指标，而且很容易理解，就是被分对的样本数除以所有的样本数，通常来说，正确率越高，分类器越好。

准确率确实是一个很好很直观的评价指标，但是有时候准确率高并不能代表一个算法就好。比如某个地区某天地震的预测，假设我们有一堆的特征作为地震分类的属性，类别只有两个：0：不发生地震、1：发生地震。一个不加思考的分类器，对每一个测试用例都将类别划分为0，那那么它就可能达到99%的准确率，但真的地震来临时，这个分类器毫无察觉，这个分类带来的损失是巨大的。

为什么99%的准确率的分类器却不是我们想要的，因为这里数据分布不均衡，类别1的数据太少，完全错分类别1依然可以达到很高的准确率却忽视了我们关注的东西。再举个例子说明下。在正负样本不平衡的情况下，准确率这个评价指标有很大的缺陷。比如在互联网广告里面，点击的数量是很少的，一般只有千分之几，如果用acc，即使全部预测成负类（不点击）acc也有 99% 以上，没有意义。因此，单纯靠准确率来评价一个算法模型是远远不够科学全面的。

2、错误率（Error rate）

错误率则与准确率相反，描述被分类器错分的比例，error rate = (FP+FN)/(TP+TN+FP+FN)，对某一个实例来说，分对与分错是互斥事件，所以accuracy =1 - error rate。

3、精确率、精度（Precision）

精确率(precision)定义为：



表示被分为正例的示例中实际为正例的比例。

6、召回率（recall）

召回率是覆盖面的度量，度量有多个正例被分为正例，

类别c的查全率(recall)为是所有真实标签为类别c的样本中，预测正确的比例



7、综合评价指标（F-Measure）

Precision和Recall指标有时候会出现的矛盾的情况，这样就需要综合考虑他们，最常见的方法就是F-Measure（又称为F-Score）。

F-Measure是Precision和Recall加权调和平均：



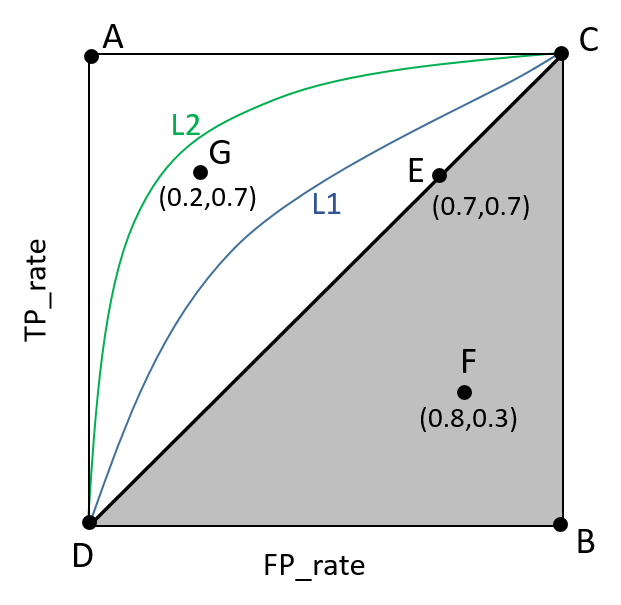
当参数α=1时，就是最常见的F1，也即



可知F1综合了P和R的结果，当F1较高时则能说明试验方法比较有效。

**1、ROC曲线：**

ROC（Receiver Operating Characteristic）曲线是以假正率（FP\_rate）和真正率（TP\_rate）为轴的曲线，ROC曲线下面的面积我们叫做AUC，如下图所示：



TPR：真实的正例中，被预测正确的比例

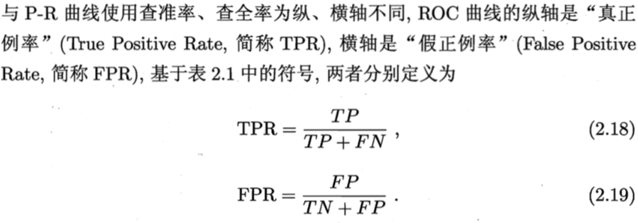
FPR：真实的反例中，被预测正确的比例

（1）Area Under Curve

AUC就是ROC曲线下方的面积。可以知道，TPR越大的情况下，FPR始终很小，才是好的，那么这条曲线就是很靠近纵轴的曲线，那么下方面积就大。所以AUC越大越好。

（2）若二者相等（CDE三点），意味着无论一个样本本身是正例还是负例，分类器预测其为正例的概率是一样的，这等同于随机猜测（注意这里的“随机”不是像抛硬币那样50%正面50%反面的那种随机）。G在这样的点上TPR总大于FPR，意为正例被判为正例的概率大于负例被判为正例的概率。

(4)虽然ROC曲线相比较与Precision Recall等衡量指标更加合理，但是其在高不平衡数据条件下的表现仍然过于理想，不能够很好的展示实际情况

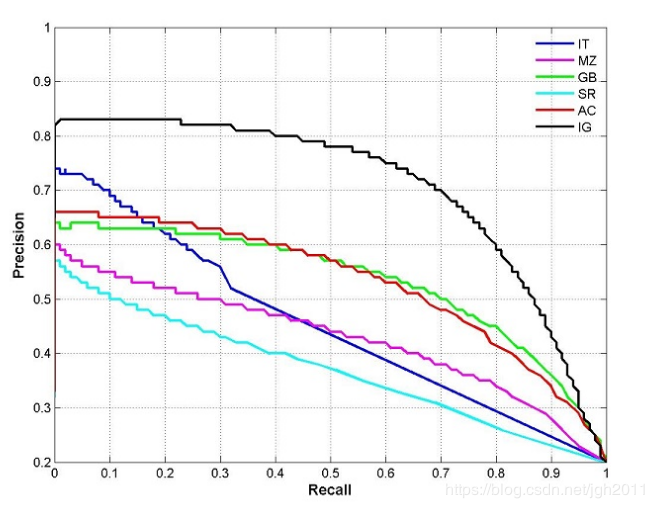


可以看出TPR和Recall的形式是一样的，就是查全率了，FPR就是保证这样的查全率你所要付出的代价，就是把多少负样本也分成了正的了。

当一个样本被分类器判为正例，若其本身是正例，则TPR增加；若其本身是负例，则FPR增加，因此ROC曲线可以看作是随着阈值的不断移动，所有样本中正例与负例之间的“对抗”。曲线越靠近左上角，意味着越多的正例优先于负例，模型的整体表现也就越好。

2、PR曲线：

即，PR（Precision-Recall）曲线。



P-R图直观地显示出学习器在样本总体上的查全率、查准率。在进行比较时，若一个学习器的P-R曲线被另一个学习器的曲线完全包住，则可断言后者的性能优于前者。