**[机器学习算法中的准确率(Precision)、召回率(Recall)、F值(F-Measure)](https://www.cnblogs.com/Zhi-Z/p/8728168.html)**

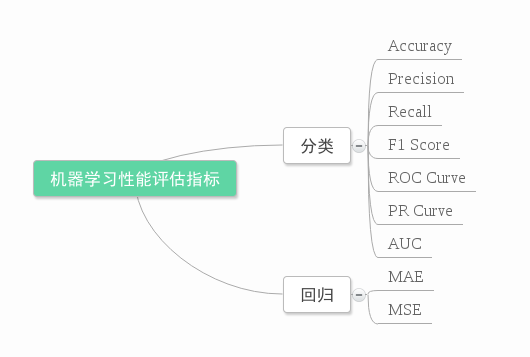
摘要：

　　数据挖掘、机器学习和推荐系统中的评测指标—准确率(Precision)、召回率(Recall)、F值(F-Measure)简介。

引言：

　　在机器学习、数据挖掘、推荐系统完成建模之后，需要对模型的效果做评价。

业内目前常常采用的评价指标有准确率(Precision)、召回率(Recall)、F值(F-Measure)等，下图是不同机器学习算法的评价指标。下文讲对其中某些指标做简要介绍。



**本文针对二元分类器！**   
**本文针对二元分类器！！**   
**本文针对二元分类器！！！**

**对分类的分类器的评价指标将在以后文章中介绍。**

**在介绍指标前必须先了解“混淆矩阵”：**

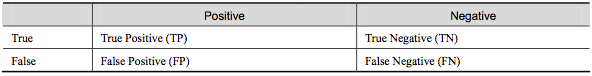
**混淆矩阵**

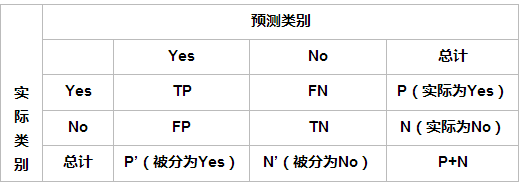
True Positive(真正，TP)：将正类预测为正类数

True Negative(真负，TN)：将负类预测为负类数

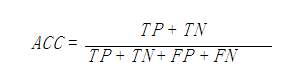
False Positive(假正，FP)：将负类预测为正类数误报 (Type I error)

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





**1、准确率（Accuracy）**

准确率(accuracy)计算公式为：   


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。

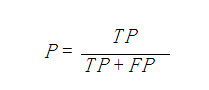
**3、灵敏度（sensitive）**

sensitive = TP/P，表示的是所有正例中被分对的比例，衡量了分类器对正例的识别能力。

**4、特效度（sensitive）**

specificity = TN/N，表示的是所有负例中被分对的比例，衡量了分类器对负例的识别能力。

**5、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，recall=TP/(TP+FN)=TP/P=sensitive，可以看到召回率与灵敏度是一样的。

**7、综合评价指标（F-Measure）**   
P和R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：   
这里写图片描述   
当参数α=1时，就是最常见的F1，也即   
这里写图片描述   
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。

**8、其他评价指标**

计算速度：分类器训练和预测需要的时间；

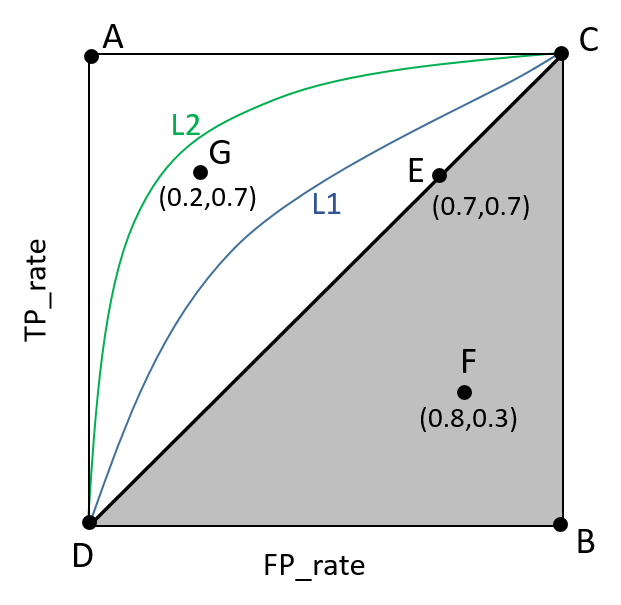
鲁棒性：处理缺失值和异常值的能力；

可扩展性：处理大数据集的能力；

可解释性：分类器的预测标准的可理解性，像决策树产生的规则就是很容易理解的，而神经网络的一堆参数就不好理解，我们只好把它看成一个黑盒子。

下面来看一下ROC和PR曲线（以下内容为自己总结）：

**1、ROC曲线：**   
ROC（Receiver Operating Characteristic）曲线是以假正率（FP\_rate）和假负率（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：



图片根据Paper：Learning from eImbalanced Data画出

其中：

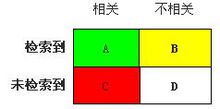
这里写图片描述

（1）曲线与FP\_rate轴围成的面积（记作AUC）越大，说明性能越好，即图上L2曲线对应的性能优于曲线L1对应的性能。即：曲线越靠近A点（左上方）性能越好，曲线越靠近B点（右下方）曲线性能越差。   
（2）A点是最完美的performance点，B处是性能最差点。   
（3）位于C-D线上的点说明算法性能和random猜测是一样的–如C、D、E点。位于C-D之上（即曲线位于白色的三角形内）说明算法性能优于随机猜测–如G点，位于C-D之下（即曲线位于灰色的三角形内）说明算法性能差于随机猜测–如F点。   
（4）虽然ROC曲线相比较于Precision和Recall等衡量指标更加合理，但是其在高不平衡数据条件下的的表现仍然过于理想，不能够很好的展示实际情况。

**2、PR曲线：**   
即，PR（Precision-Recall）曲线。   
举个例子（例子来自Paper：Learning from eImbalanced Data）：   
假设N\_c>>P\_c（即Negative的数量远远大于Positive的数量），若FP很大，即有很多N的sample被预测为P，因为这里写图片描述，因此FP\_rate的值仍然很小（如果利用ROC曲线则会判断其性能很好，但是实际上其性能并不好），但是如果利用PR，因为Precision综合考虑了TP和FP的值，因此在极度不平衡的数据下（Positive的样本较少），PR曲线可能比ROC曲线更实用。

百度百科的计算案例：

召回率和精度示意图

[](https://baike.baidu.com/pic/%E5%8F%AC%E5%9B%9E%E7%8E%87/560642/0/3790312e19e855674ec2262d?fr=lemma&ct=single)recall

假定：从一个大规模数据集合中检索文档时，可把文档分成四组：

- 系统检索到的相关文档（A）

- 系统检索到的不相关文档（B）

- 相关但是系统没有检索到的文档（C）

- 不相关且没有被系统检索到的文档（D）

则：

- 召回率R：用实际检索到相关文档数作为分子，所有相关文档总数作为分母，即R = A / ( A + C )

- 精度P：用实际检索到相关文档数作为分子，所有检索到的文档总数作为分母，即P = A / ( A + B )

举例来说：

一个数据库有500个文档，其中有50个文档符合定义。系统检索到75个文档，但是实际只有45个符合定义。则：

召回率R=45/50=90%

精度P=45/75=60%

本例中，系统检索是比较有效的，召回率为90%。但是结果有很大的噪音，有近一半的检索结果是不相关。 研究表明：在不牺牲精度的情况下，获得一个高召回率是很困难的。