1. **模型评估**

在不失一般性的情况下，本文以二分类的情况介绍相关模型评估指标。以1代表正类，0代表负类。一般以1代表人们更关注的类别，比如信用欺诈问题中，将发生信用欺诈行为的的类别设为1。

**1）精度（accuracy）：**

又称正确率，指模型预测正确的样本占全部样本的比率，一般情况下，模型的精度越高，说明模型的效果越好，该指标在分类问题中经常使用。尽管精度在用途十分广泛，但其仍有不小的缺陷。例如，收集到一个信用欺诈的样本，其中类别为1（即发生信用欺诈行为）的样本占5个，类别为0（即未发生信用欺诈行为）的样本占95个，假如存在一个学习器，将这些样本的类别全部预测为0，其精度就高达95%，但这样的学习器是无意义的，它无法准确分辨出类别为1的样本。

**2）混淆矩阵。**

在进行模型评估时，混淆矩阵包含大量的信息，相对于精度而言，它更能体现出一个学习器的综合性能。表1为混淆矩阵常见形式。

表1 混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion matrix | | Predicted condition | |
| Positive | Negative |
| True condition | Positive | TP | FN |
| Negative | FP | FN |

矩阵四个元素的含义：

* 真正例（True Positive，TP）：指被学习器正确分类的正例数据（即真实类别为1且预测类别为1的样本）
* 真负例（True Negative，TN）：指被学习器正确分类的负例数据（即真实类别为0且预测类别为0的样本）
* 假正例（False Positive，FP）：被错误地标记为正例数据的负例数据（即真实类别为0且预测类别为1的样本）
* 假负例（False Negative，FN）：被错误地标记为负例数据的正例数据（即真实类别为1且预测类别为0的样本）

混淆矩阵常见的指标解释：

* 查准率（precision）：指学习器预测为正例的样本中，正例所占的比例，计算公式如下：
* 查全率（recall）：指实际类别为正例的样本中，被预测为正例的样本所占的比例，计算公式如下：



为什么引入查准率和查全率指标呢？

在一些常见的问题中，仅仅依靠精度并不能解决实际生活中的需求，例如在信息检索过程中，用户经常会关心“检索出的信息中有多少比例是用户感兴趣的”“用户感兴趣的信息中有多少被检索出来啦”’。查准率和查全率就比较适合此类问题。

然而，遗憾的是，一般情况下，查准率与查全率呈现此消彼长的状态。查准率高时，查全率经常偏低；查全率高时，查准率经常偏低。这种现象也是比较容易理解的，为了提高查准率，学习器倾向于把更多的样本归为负类，导致FP减小，FN增大，所以导致查全率降低；同理为了提高查全率，学习器倾向于把更多的样本归为正类，导致TP增大，FP也增大，但FP得增量将会比TP得增量大许多，所以导致查准率降低。



为了平衡两个指标在模型评估过程中的作用，F1度量指标被引入，其基于查准率和查全率两个指标产生，计算公式如下：



但有时对于不同的分类问题，其侧重点也有所不同。比如，搜索引擎在为用户提供信息时，基于用户输入的关键字，搜索引擎将要判断将哪些信息提供给用户，此时查准率显然更重要，因为用户更希望自己获得的信息中，更多是自己想要的；利用机器学习判断蘑菇是否有毒时，查全率更为重要，因为毒蘑菇被错分为无毒蘑菇的后果更为严重，将会导致食用者中毒。

因此Fβ被引入，其计算公式如下：

其中β > 0,当β > 1时，查全率有更大的影响；当β < 1时，查准率有更大的影响。

1. **ROC曲线与AUC值**

ROC（Receiver Operating Characteristic）:全称为“受试者工作特征”，它源于“二战”中的雷达信号分析技术，后来被引入机器学习领域。为什么引入ROC呢？首先看一下一些学习器的特点，许多学习器能够基于样本产生一个在[0.0,1.0]之间的一个实值，这个实值可以作为样本属于类别1的概率，可以将这些实值进行从大到小排序，将“最可能”属于类别1



的样本放在最前面，“最不可能”的样本放在最后面，然后设置一个阈值，大于这个阈值，样本类别归为1，小于这个阈值，样本类别归为0。从这里可以看出，学习器产生的一系列的实值对学习器预测效果的好坏至关重要，一个好的学习器应该使得实际类别为1的样本对应的实值尽可能大，以确保其被分类为1。ROC曲线正是为了评估学习器这方面特性的优劣。

ROC曲线的绘制：ROC曲线的横轴为“假正例率”（FPR）（即实际为负例的样本中被预测为正例的比率），纵轴为“真正例率”（TPR）（即实际为正例的样本中被预测为正例的比率），其计算方式分别如下：



上面谈到，一些学习器能够能够产生一系列类似于概率的实值，基于这些值，即可绘制ROC



曲线，将这些实值按照从大到小的顺序排列，首先设置阈值为最大的实值，将样本分类，可以得到一组（FPR，TPR），这相当于产生了ROC曲线上的一个点，然后将阈值设置为第二大的实值，又得到一组（FPR，TPR），以此类推，将得到许多（FPR，TPR），将这些点用曲线连接，就得到ROC曲线。图1中左图为理想的ROC曲线，因为实际问题中，样本数量有限，右图是更常见的情形。

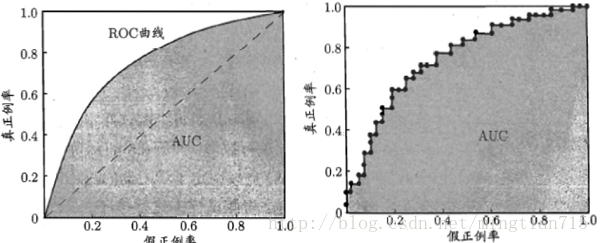


图1 ROC曲线与AUC

ROC曲线分析：在ROC曲线中，点（0,1）时最理想的情形，此时TPR为1，FPR为0，即所有的正例都被排在负例前面，点（1,0）时最糟糕的情形，此时TPR为0，FPR为1，即所有的负例都被排在正例前面。对角线时随机猜测的情形，即样本被分类正确的概率为0.5，一般学习器的效果优于随机猜测，其ROC曲线能够包住对角线。同样在比较两个学习器的性能时，如果一个学习器的ROC曲线能包住另一个学习器的ROC曲线，则认为其性能较好。

AUC：ROC曲线下的面积。通过相关文献可以得知：从所有1样本中随机选取一个样本，从所有0样本中随机选取一个样本，使用学习器对这两个样本进行预测，把1样本预测为1的概率为p1，把0样本预测为1的概率为p0，p1大于p0的概率就是AUC。AUC的值越大，说明学习器对样本的分类排序能力越好，所以对于ROC曲线发生交叉的学习器，便可以通过AUC进行比较，AUC值较大的分类器性能相对较好。