# 准确度的度量

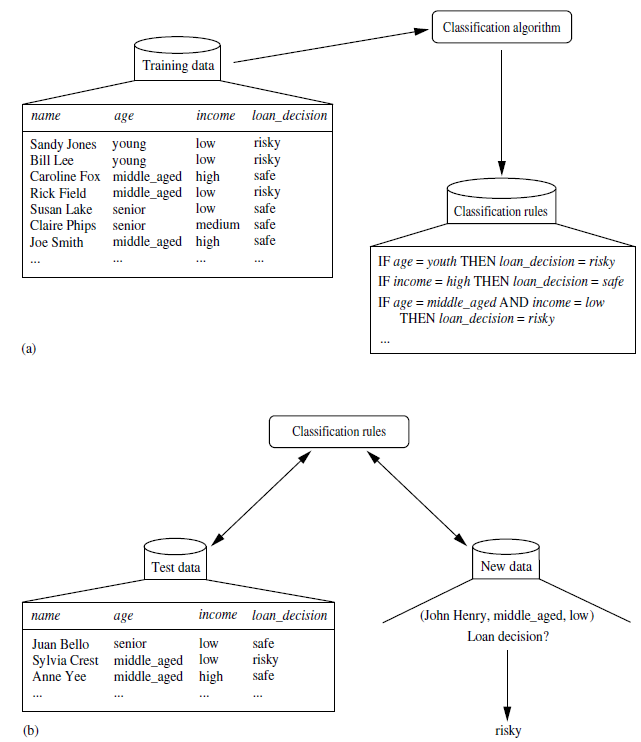
当训练的一个分类器后，会出现许多进一步的问题。例如，假设使用了销售数据训练一个分类器来预测有购买行为的客户。则我们随后会：

1. 期望估计一下该分类器用于预测未来客户具有购买行为的准确度有多高。（未来预测的准确度有多高？）

2. 另外可能会尝试不同方法来构建多个分类器，于是希望比较这些分类器的准确度。（两个模型哪个更准确？）

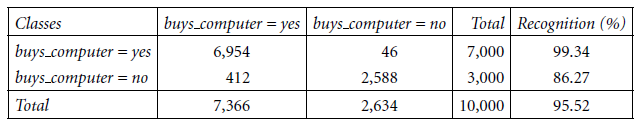
然而，什么是**准确度**？我们**如何估计准确度**？这里是否有一些策略来提升已有模型的准确度呢？

## 一、分类器准确度度量

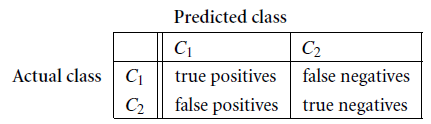


一般说来会将已知的数据集又分为两个集合：训练数据（training data）集合和试验数据（test data）集合。这里可以简称训练数据集合和试验数据集合分别为训练集（training set）和试验集（test set）。训练集用于得到分类器，而**试验集**就用于估计分类器的**准确度**。如果准确度可以接受，那么这个分类器就可以用于新数据（new data）集合的分类了。由此定义分类器针对一给定试验集的**准确度**（accurary）为由这个分类器正确分类的试验集记录（record）的百分比。在模式识别（pattern recognition）的范畴里，准确度又称作**识别率**（recognition rate）。分类器M的准确度简单表示为**Acc(M)**。我们也可以说分类器的差错率（error rate）或错分率（misclassification rate），简单表示为**1-Acc(M)=err(M)**。

混淆矩阵是分析分类器在识别不同类别时表现出的好坏程度的有用工具。下图是一个只有类别的混淆矩阵的实例。



给定m个类别，**混淆矩阵（confusion matrix）**是一个至少有m\*m大小的表格。给定两个类别，我们称作正面记录（positive tuples）（也就是感兴趣的主类别，例如上表中的buys\_computer=yes），和负面记录（negative tuples）（例如上表中的buys\_computer=no）。真实正面（true positive）指示那些由分类器正确标示的正面记录；而真实负面（true negative）为那些由分类器正确标示的负面记录。虚假正面（false positive）为没有正确标示的负面记录；而虚假负面（false negative）就是那些没有正确标示的正面记录。这些定义用于分析分类器能力时，很有用。其简单表示在下表中。



用以上定义，我们给出常用度量的公式表示：

sensitivity = TP/(TP+FP)

specificity = TN/(TN+FN)

precision = TP/(TP+FP) = sensitivity

accuracy=sensitivity\*(TP+FP)/ALL+specificity\*(TN+FN)/ALL=(TP+TN)/ALL

其中有：

ALL=TP+FP+TN+FN

这些度量分别就是**灵敏性（sensitivity）**，**特异性（specificity）**，**精密度（precision）**和**准确度（accuracy）**

参见前面的表，可以定义：

TP Rate = TP/(TP+FN)

FP Rate = FP/(FP+TN)

另外还定义了：

Recall = TP/(TP+FN) = TP Rate

Precision = TP/(TP+FP) = sensitivity

还有F-Measure定义为：

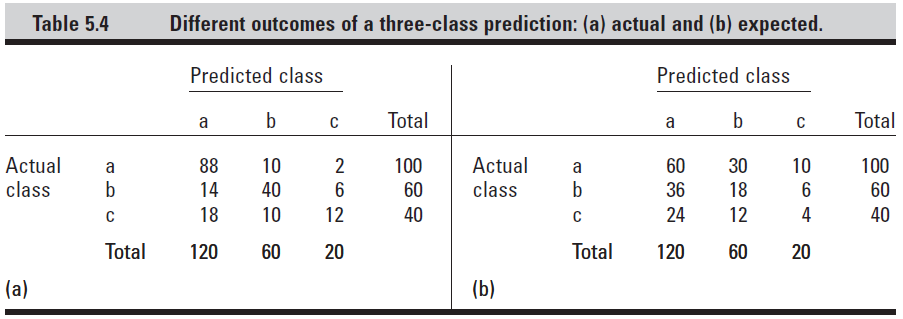
2\*recall\*precision/(recall + precision) = 2\*TP/(2\*TP + FN + FP)=(TP+TP)/((TP+FN) + (TP+FP))

如果考察的类别大于两个，例如为n个类别，则其混淆矩阵为n\*n的。这时可以认为对某个类别Ci，其：

* TP为矩阵中第i行和第i列交叉点单元的值；
* FP为第i列中除了TP单元外所有单元和的值；
* FN为第i行中除了TP单元外所有单元和的值；
* TN则为除了以上计算了的单元外，其它所有单元和的值；

这样就可以按照以上定义，求取该类别Ci相应的sensitivity, specificity, TP Rate, FP Rate, Recall, Precision；但准确度则是针对整个模型，为混淆矩阵中对角线相加除以总样本数！！！

此外，Kappa statistic的定义涉及到了期望混淆矩阵（expected confusion matrix）。如下图例子，



左边的表为进行了test set检验后得到的混淆矩阵。然后，根据各个预测类数目的比例，即120:60:20，来分配各个实际类的总书目，即100,60,40。这样就得到了期望混淆矩阵。设原来混淆矩阵对角单元数目相加的和为Do，而期望混淆矩阵中对角单元数目相加的和为De，所有单元的总和为ALL，则Kappa Statistic为：

Kappa = (Do-De)/(ALL-De)

例如，上表的例子中：

Do = 88+40+12 = 140

De = 60+18+4 = 82

ALL = 200

所以有：

Kappa = 58/118 = 0.49153

真实正面、真实负面、虚假正面和虚假负面在评估与分类模型相关的**代价与收益**（costs and benefits）（或称为**风险与增益**（risks and gains））时也很有用。与虚假负面相关的代价可能远远大于虚假正面的代价；类似的，与真实正面决策相关的收益可能与真实负面的收益并不相同。而在通常计算分类器的准确度时，我们都假设有相等的代价，即用试验集的总数来除真实正面和真实负面的总和。作为替代，我们可以计算每个决策的平均代价（或收益）来包含代价和收益。（我们所分析的一篇论文就是针对这个代价问题的！！！）

## 二、评测分类器准确度的方法

1. **保留（holdout）**方法和**随机子采样（random subsampling）**方法

保留方法就是随机将已有数据集分为两个独立的集合，分别称作训练集（training set）和试验集（test set）。通常2/3的数据放在训练集中，剩下的1/3放在试验集中。

随机子采样的方法是保留方法的一个改进，也就是重复进行k次保留方法。

2. **交叉验证（cross-validation）**方法

在k折交叉验证（k-fold cross-validation）方法中，初始数据被随机分为k个相互不包含的子集，即“折（fold）”，D1, D2,..., Dk，每个子集的大小近似相等。训练和试验共进行k次。在第i次迭代时，子集Di保留作为试验集，而剩下的其它数据作为训练集。对分类来说，准确度的估计为来自k迭代的正确分类的总数除以初始数据中记录的总数。（目前常用10折交叉验证！）

3. **Boostrap**

可替代地采样，.632 boostrap，(1-exp(-1))=0.632。boostrap在小数据集时有很好效果

## 三、集成方法（ensemble method）可提升准确度

包括**Bagging**（bootstrap aggregation）方法和**Boosting**方法。其基本思想就使用多个分类模型的组合。

## 四、模型选择

假设我们由数据生成了两个分类模型M1和M2。我们可以用10折交叉验证方法获得这两个模型的平均差错率（mean error rate）。但平均差错率只是当模型应用到未来真实数据时可能出现的差错的估计。当进行多次10折交叉验证后，各自的差错率很可能有差别。下面接受如何合理地比较和选择模型。

### 4.1 估计置信区间

我们可以使用统计学中的假设检验（hypothesis testing）来比较两个模型的准确度。这里使用了**t检验**（t-testing），又称作学生t检验。

1. 第一种情况，设进行k折交叉验证，每轮验证时两个模型使用同一个试验集，这称为成对比较（pairwise comparison）。设err(M1)i为模型M1在第i轮交叉验证得到的差错率，相应err(M2)i为模型M2在第i轮交叉验证得到的差错率。而用avgerr(M1)和avgerr(M2)表示两个模型各自平均差错率。这两个模型之间差异的方差表示为var(M1-M2)。t检验为k个样本计算自由度为k-1的t统计（t-statistic）。本成对比较情况下的t统计为：

t=(avgerr(M1)-avgerr(M2))/(sqrt(var(M1-M2)/k)),

其中：

var(M1-M2)=1/k \* SUM{[(err(M1)i - err(M2)i) - (avgerr(M1)-avgerr(M2))]^2}

得到t统计后，通过查表，可以得到原来假设的置信区间。

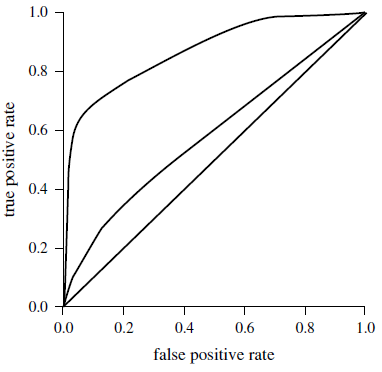
2. 第二种情况，每轮验证时两个模型有各自的试验集。而且可能两个模型分别进行了k1和k2折的交叉验证。这时t统计的方差为：

var(M1-M2)=sqrt(var(M1)/k1 + var(M2)/k2)

而我们把这两模型中最小的度数作为查表时用的自由度。

### 4.2 ROC折线

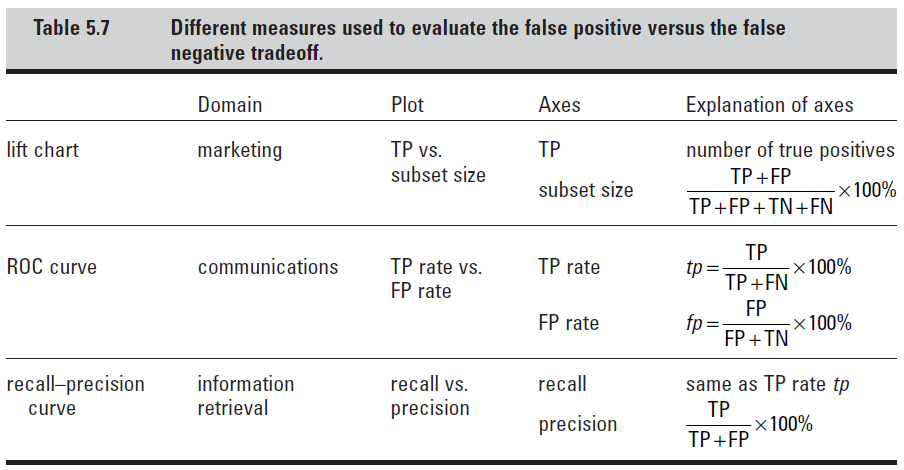
**ROC**能直观表示两个分类模型的比较。ROC是Receiver Operating Characteristic（接收机工作特性）的缩写，其源自于二战时为分析雷达图像而发展的信号检测理论。ROC折线显示了对一个模型，其真实正面率（灵敏性，正面记录中正确确认的部分）和虚假正面率（负面记录中被不正确地确认为正面的部分）之间的相互关系。真实正面率的任何增长都是以虚假正面率的增长为代价的，而ROC折线下面部分的面积是模型准确度的度量。



为了画给定分类模型M的ROC折线图，该模型必须能够返回每个试验记录的预测类别的一个概率。也就是，我们需要给试验记录按降序来分等级，其中分类器认为最可能属于正面的出现在这个等级列的顶部。ROC折线的横纵坐标分别是虚假正面率和真实正面率。下面开始画模型M的ROC折线。从左手底部的角开始，我们检查以上排序的记录列顶部记录的类别标签。如果我们得到一个真实正面，则在ROC折线上，我们向上移动，画一个点。而如果记录属于虚假正面，则我们向右边移动，并画一个点。这个过程针对列中的每个记录重复进行。

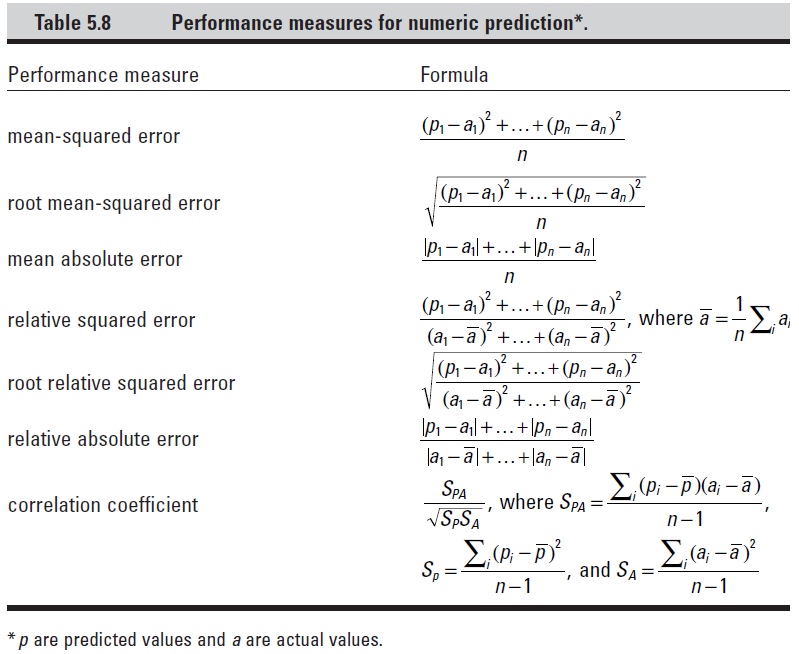
上图画出了两个分类模型的ROC折线。同时还画了对角线。为评估一个模型的准确度，我们可以度量折线下的面积。有软件包可以完成这样的工作。面积越接近0.5，则对应模型越不准确。而具有1.0面积的模型就有非常好的准确度。

另外还有两类折线来比较分类模型，分别是lift chart和recall-precision curve。其定义见下表中：



它们有各自通常使用的领域（marketing, communications, information retrieval）。

当我们预测的属性为数值变量的时候，需要有其它相应的对预测进行评估的参数，如下图：



这些公式中，pi和ai分别代表在进行了test set的检验后， test set中各个instances的预测值和实际值。