**1. 评估方法**

（1）留出法：

定义：直接将集合划分为两个互斥的集合，其中一个作为训练集合，一个作为测试集合

注意：

（a）测试集合和训练集合尽可能保持数据分布的一致性，比如要保证正反样本的比例不变（这是一种导致过拟合的原因）

（b）在给定了训练/测试集合的样本比例之后，仍要存在多种的划分方式，对数据集合D进行分割。毕竟单次的使用留出法的结果往往是不可靠的。一般要采用若干次的随即划分，重复进行实验评估后取得平均值作为留出法的最终评估结果

（c）训练/测试集合的大小比例问题。测试集合过小，会导致测评结果的方差变大；训练集合过小，会导致偏差过大，一般使用的都是2/3~4/5的样本用于训练

（2）交叉验证法（无放回的重采样）

交叉验证方法还有一个特例：留一法（Leave-one-out，LOO）。显然留一法不受随即样本划分的影响，因为m个样本只有唯一的方式划分为m个子集，每个子集包含一个样本。由于使用的训练数据集合和初始数据集合相比只是少了一个数据，那么相应训练出来的模型就会很相似，实践认为，留一法的评估结果往往是比较准确的。但是留一法存在性能上的缺陷。比如，m个样本就要训练m个模型，当m比较大的时候，那么就比较麻烦了

优点：

1. 在处理较小的数据集合的时候我们一般会使用较大的k值来增加，根据NX（1-1/k）较大的k使得在每次的迭代中会有更小的数据bias，但是同时会导致validation的时间变长，而且也会导致high Variance。也可以看出，LOO更适合小数据集上的训练

2. 那么数据较大的时候我们一般会使用较小的k值

（3）自助法Bootstrapping（有放回的重采样）

有放回的随机采样方法。我们可以做一个简单的估计，样本在m次采样的过程中始终不被采到的概率是（1-1/m）^m，对m取极限，得到1/e=0.368。也就是说，通过自助法，初始数据集合中大概有36.38%的样本未出现在采样数据集合D’中，于是我们可以将D‘作为训练集合，剩下的数据作为测试集合，那么我们对这中大概有1/3的样本没在测试集合中出现的测试称作包外估计（out-of-bag estimate）

优点：

1. 数据集合较小，难以有效划分训练/测试集时很有用

2. 能从训练集合中产生很多不同的训练集，这对集成学习等方法很有好处

缺点：

改变了初始数据的分布，会引入偏差。因此如果初始的数据量足够的时候还是希望使用留出法和交叉验证法比较好

2. 查准率，查全率，F1 Score

|  |  |  |
| --- | --- | --- |
| 真实情况 | 预测结果 | |
| 正例 | 反例 |
| 正例 | TP(真正例) | FN(假反例) |
| 反例 | FP(假正例) | TN(真反例) |

查准率P=TP/(TP+FP)

查准率(精确率)是针对我们预测结果而言的，它表示的是预测为正的样本中有多少是真正的正样本。那么预测为正就有两种可能了，一种就是把正类预测为正类(TP)，另一种就是把负类预测为正类(FP)

查准率更加重视正例的准确程度，在推荐系统中，我们更希望推荐的商品是真正对用户有吸引力的，而不想推荐一些不相关的产品，以免打扰用户；但是对于信息检索中的逃犯检索，我们就希望比较大的查全率。

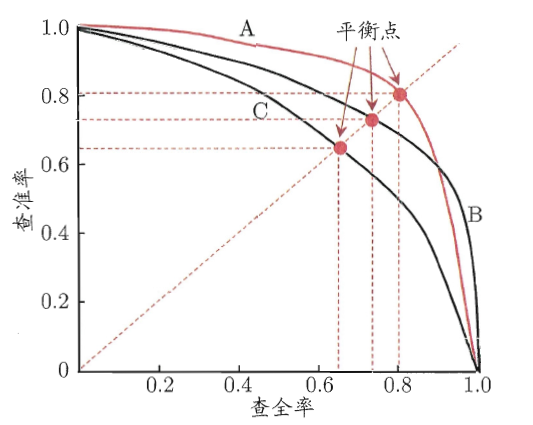
查全率R=TP/(TP+FN)

查全率(召回率)是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。那也有两种可能，一种是把原来的正类预测成正类(TP)，另一种就是把原来的正类预测为负类(FN)

查准率较高的时候，查全率往往偏低；查全率较高的时候，查准率往往比较低。我们可以通过一个例子来看，希望好瓜尽可能多的选择出来，则可以通过增加选瓜的数量来实现，如果将所有的西瓜都选择上，那么所有的好瓜也必然都选择上了，这样的查准率就会比较低；如希望选择出的好瓜的概率尽可能高，那么只挑选最有把握的瓜，但是这样就会导致漏掉不少的好瓜，使得查全率比较低。

在很多情况下，我们都的机器学习的模型对分类问题的预测结果都是以概率的形式呈现出来的，如果要计算准确率，那么我们就要设定一个阈值，手动将这些概率转换成一种类别。所以，这个阈值的设置在很大程度上都会影响到整个模型的概率准确率的计算，那么我们适用P-R曲线就可以不用这个转化的过程了（其实AUC和logloss都是可以实现这个过程）。我们根据机器学习得到结果排序，分别对样本作为正例进行预测，则每次就可以计算出当前的查准率和查全率，以查准率作为纵轴、查全率作为横轴，就得到了P-R曲线

**P-R曲线**



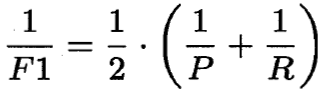
性能比较：

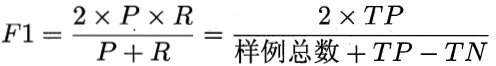
(1)如果一个机器学习的P-R曲线被另外的一个机器学习的曲线完全包住，那么可以断言后者的性能比前者的性能好，比如图中的A的性能就比B的性能好

(2)如果有交叉，那么只能在具体的查全率的状态下比较二者的查准率

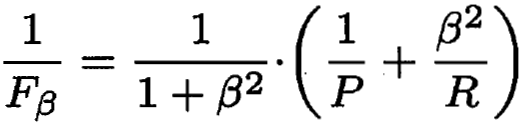
(3)当然，我们也可以通过平衡点（Break-Even Point）来判别二者的优劣。也就是“查准率=查全率”的时候的取值，值大，那么性能就好

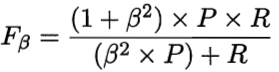
(4)上面的都是太简单了，我们可以使用F1来来度量，我们先看看F1是什么：F1是查全率和查准率的调和平均（与算术平均和几何平均相比，调和平均更加关注较小值）





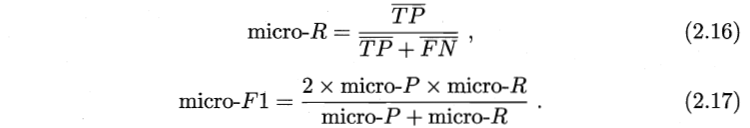
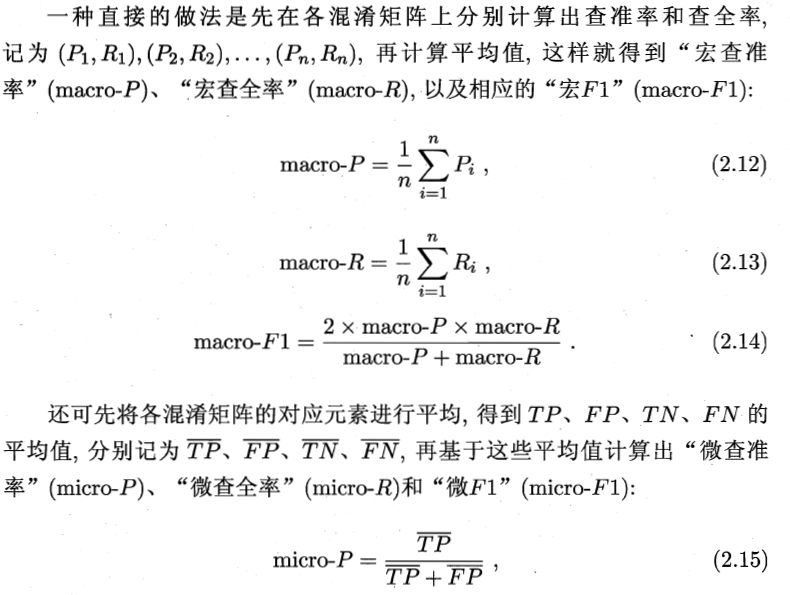
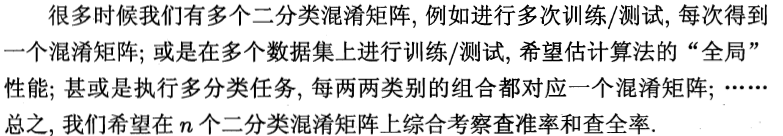
由于查准率和查全率对预测结果的关注度不同我们引入了一个新的评价标准，也就是更加一般化的F1 Score：



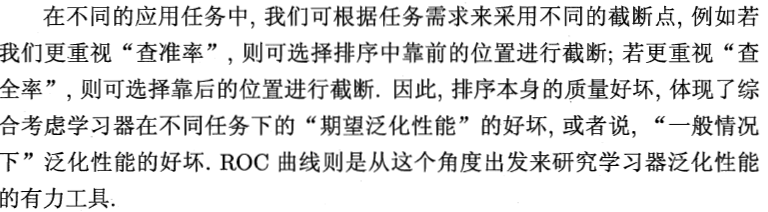


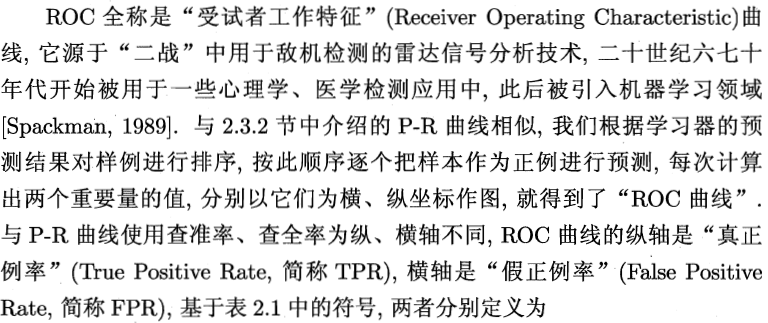
在上面的公式中，我们通过beta来调整查全率和产准率的权衡。Beta>1查全率具有更大影响；Beta<1查准率有更大的影响。

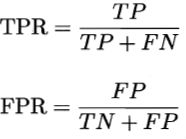
(5)上面的是针对单个二分类来说的，那么对于更一般的情况下，我们有如下方法：



**3. ROC与AUC**







AUC对样本类别是否均衡并不敏感，这也是不均衡样本通常用AUC评价分类器性能的一个原因。