AUC的计算方法

摘要:

在机器学习的分类任务中，我们常用许多的指标，诸如召回率（Recall）、准确率（Precision）、F1值、AUC等。

那么，如果手动计算AUC应该要怎么计算呢？相信大家很多时候都是用写好的库直接计算，可能对AUC计算不太了解，下面这篇文章就简单的概述一下AUC的计算方法。

（注：本文的重点其实不在于阐述什么是AUC。因为网上关于这方面的文章实在太多了。但是对于AUC的计算的文章相对来说少一些）

1.什么是AUC?

相信这个问题很多玩家都已经明白了，简单的概括一下，AUC（are under curve)是一个模型的评价指标，用于分类任务。

那么这个指标代表什么呢？这个指标想表达的含义，简单来说其实就是随机抽出一对样本（一个正样本，一个负样本），然后用训练得到的分类器来对这两个样本进行预测，预测得到正样本的概率大于负样本概率的概率。

具体关于AUC含义的分析推荐以下回答：

2.如何计算AUC?

计算AUC时，推荐2个方法。

方法一：

在有M个正样本,N个负样本的数据集里。一共有M*N对样本（一对样本即，一个正样本与一个负样本）。统计这M*N对样本里，正样本的预测概率大于负样本的预测概率的个数。

这样说可能有点抽象，我举一个例子便能够明白。

ID	label	pro
A	0	0.1
B	0	0.4
C	1	0.35
D	1	0.8

假设有4条样本。2个正样本，2个负样本，那么M*N=4。即总共有4个样本对。分别是：

（D,B），（D,A）,(C,B),（C,A）。

在（D,B）样本对中，正样本D预测的概率大于负样本B预测的概率（也就是D的得分比B高），记为1

同理，对于（C,B）。正样本C预测的概率小于负样本C预测的概率，记为0.

在这个案例里，没有出现得分一致的情况，假如出现得分一致的时候，例如：

ID	label	pro
A	0	0.1
B	0	0.4
C	1	0.4
D	1	0.8

同样本是4个样本对，对于样本对（C,B）其I值为0.5。

方法二：

另外一个方法就是利用下面的公式：

$$AUC = \frac{\sum_{ins_i \in positiveclass} rank_{ins_i} - \frac{M \times (M+1)}{2}}{M \times N}.$$

这个公式看起来有点吓人，首先解释一下每一个符号的意思：

公式的含义见： 公式解释

同样本地，我们用上面的例子。

ID	label	pro
A	0	0.1
B	0	0.4
C	1	0.35
D	1	0.8

将这个例子排序。按概率排序后得到：

ID	label	pro	rank
A	0	0.1	1
C	1	0.35	2
B	0	0.4	3
D	1	0.8	4

按照上面的公式，只把正样本的序号加起来也就是只把样本C,D的rank值加起来后减去一个常数项

$$\frac{M(M+1)}{2}$$

即：

$$\frac{(4+2)-\frac{2*(2+1)}{2}}{2*2}=\frac{6-3}{4}=0.75$$

，这个答案和我们上面所计算的是一样的。

这个时候，我们有个问题，假如出现得分一致的情况怎么办？下面举一个例子说明：

ID	label	pro
A	1	0.8
B	1	0.7

C	0	0.5
D	0	0.5
E	1	0.5
F	1	0.5
G	0	0.3

在这个例子中，我们有4个取值概率为0.5，而且既有正样本也有负样本的情况。计算的时候，其实原则就是相等得分的rank取平均值。具体来说如下：

先排序：

ID	label	pro	rank
G	0	0.3	1
F	1	0.5	2
E	1	0.5	3
D	0	0.5	4
C	0	0.5	5
B	1	0.7	6
A	1	0.8	7

这里需要注意的是：相等概率得分的样本，无论正负，谁在前，谁在后无所谓。

由于只考虑正样本的rank值：

对于正样本A，其rank值为7

对于正样本B，其rank值为6

对于正样本E，其rank值为（5+4+3+2）/4

对于正样本F，其rank值为（5+4+3+2）/4

最后我们得到：

$$\frac{7+6+\frac{(5+4+3+2)}{4}+\frac{(5+4+3+2)}{4}-\frac{4*(4+1)}{2}}{4*3}=\frac{10}{12}$$

3.最后的最后，如何用程序验证?

为了方便，我们使用sk-learn里面自带的库来简单的验证一下我们的例子。

```
y_true is  [1 1 0 0 1 1 0]
y_scores is  [ 0.8  0.7  0.5  0.5  0.5  0.5  0.3]
AUC is 0.833333333333
y_true is  [0 0 1 1]
y_scores is  [ 0.1   0.4   0.35  0.8 ]
AUC is  0.75
```

其python的代码

```
from sklearn.metrics import roc_auc_score

y_true = np.array([1,1,0,0,1,1,0])

y_scores = np.array([0.8,0.7,0.5,0.5,0.5,0.5,0.3])

print "y_true is ",y_true
```

```
print "y_scores is ",y_scores

print "AUC is",roc_auc_score(y_true, y_scores)

y_true = np.array([0, 0, 1, 1])

y_scores = np.array([0.1, 0.4, 0.35, 0.8])

print "y_true is ",y_true

print "y_scores is ",y_scores

print "AUC is ",roc_auc_score(y_true, y_scores)
```