auc含义详述

2015年7月27日 星期一　　下午10:31

一,简介

AUC的值是ROC曲线的面积, 是一个衡量二分类好坏的指标,值域在0~1之间, 但一般正常的分类器在0.5~1之间

ROC曲线是一 个衡量二分类好坏的曲线,

纵坐标是分类器将正样本分为正类占所有正样本的比率 true positive rate(TPR),

横坐标是分类器将负样本分为正类占所有负样本的比例false positive rate(FPR)

由于分类器的输出其实是一个得分,我们可以人工卡阈值,将得分转为预测的类别,我们可以通过调整阈值,达到调整TPR,FPR的效果

通过调阈值,能画出一个TPR和FPR的变化曲线,这个曲线就是ROC

二,详述

1,

AUC（Area Under Curve）被定义为ROC曲线下的面积，显然这个面积的数值不会大于1。又由于ROC曲线一般都处于y=x这条直线的上方，所以AUC的取值范围在0.5和1之间。使用AUC值作为评价标准是因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应AUC更大的分类器效果更好。

2, 详细解释TPR,FPR等 (详见参考资料2ROC曲线-csdn博客 和 参考资料4 FPR维基)

a, 类似准确率召回率的统计方式,但又有些区别

b, TP FN FP TN四个基本概念

如果一个实例是正类并且也被 预测成正类，即为真正类（True positive）,

如果实例是负类被预测成正类，称之为假正类（False positive）。

相应地，如果实例是负类被预测成负类，称之为真负类（True negative）,

正类被预测成负类则为假负类（false negative）。

TP：正确肯定的数目；

FN：漏报，没有正确找到的匹配的数目；

FP：误报，给出的匹配是不正确的；

TN：正确拒绝的非匹配对数；

		预测		
		1	0	合计
实际	1	True Positive（TP）	False Negative（FN）	Actual Positive(TP+FN)
	0	False Positive（FP)	True Negative(TN)	Actual Negative(FP+TN)
合计		Predicted Positive(TP+FP)	Predicted Negative(FN+TN)	TP+FP+FN+TN

c, TPR,FPR

从列联表引入两个新名词。

真正类率(true positive rate ,TPR), 计算公式为TPR=TP/ (TP+ FN)，刻画的是分类器所识别出的正实例占所有正实例的比例。

负正类率(false positive rate, FPR),计算公式为FPR= FP / (FP + TN)，计算的是分类器错认为正类的负实例占所有负实例的比例。

还有一个真负类率（True Negative Rate，TNR），也称为specificity,计算公式为TNR=TN/ (FP+ TN) = 1-FPR。

3, 从采样角度的理解

和Wilcoxon-Mann-Witney Test是等价的,Wilcoxon-Mann-Witney Test就是测试任意给一个正类样本和一个负类样本，正类样本的score有多大的概率大于负类样本的score有了这个定义，我们就得到了另外一中计 算AUC的办法：得到这个概率。我们知道，在有限样本中我们常用的得到概率的办法就是通过频率来估计之。这种估计随着样本规模的扩大而逐渐逼近真实值

三, 评价

1， 问题:

其实他衡量的是, 对于分为正类之后， 的分类效果的好坏. 不关注分为负类的样本

其实: 不是的， 由于统计的是比率, 负类的情况会在分母中得到体现， 或者说 ， 由于是二分类, 对于负类的分类好坏, 可以直接推出来

2， 相比准确率/召回率, 少了个超参数:阈值, 这样的话, 说明更能体现分类器实际的能力， 而不是通过人工调整阈值, 避开一些分类器处理不好的case， 只看准确率/召回率, 有一种可能是你通过调整阈值, 让分类器拟合你的测试集, 导致掩盖了一些问题.

3， 缺点:相对于准确率/召回率, 少一道工序, 所体现的是分类器的实际能力, 而不是在实际产品应用中的效果.

4， 准确率/召回率这些指标, 其实是把分类器输出的连续的打分, 硬性的映射为了一个二分类的输出0/1

这样其实不能很好的刻画，分类器输出的那连续的得分的好坏,也就是他的"置信度"的好坏

比如说，我用准召率衡量,卡了一个阈值之后，分类器打分在这个阈值下的输出,不论值是多少,都被分为了负类,反之正类

参考资料:

1，AUC 详细的新浪博客 http://blog.sina.com.cn/s/blog_814f5e700100z9cz.html

2，ROC曲线-csdn博客 很赞 http://blog.csdn.net/abcjennifer/article/details/7359370

3,AUC维基 http://en.wikipedia.org/wiki/Area_under_the_curve_(pharmacokinetics)

4，FPR维基 http://en.wikipedia.org/wiki/False_positive_rate

https://www.google.com.hk/search?safe=strict&biw=1366&bih=643&q=auc%E8%AE%A1%E7%AE%97&revid=1206219716&sa=X&ei=J9H6U6nEGoTg8AWBuoH4Bg&ved=0CHoQ1QIoBw

http://alexkong.net/2013/06/introduction-to-auc-and-roc/