机器学习评价指标AUC

原创 CD CoreDumper 2019-04-06

AUC（Area Under Curve）常被用来评价一个二值分类器的优劣，我们首先看一下 AUC 的定义：AUC 值是一个概率值，随机挑选一个正样本和一个负样本，使用分类器分别对这两个样本计算 score，根据 score 的大小将正样本排在负样本前面的概率就是 AUC 值。AUC 值越大，分类器越有可能将正样本排在负样本前面，即能够更好的分类。

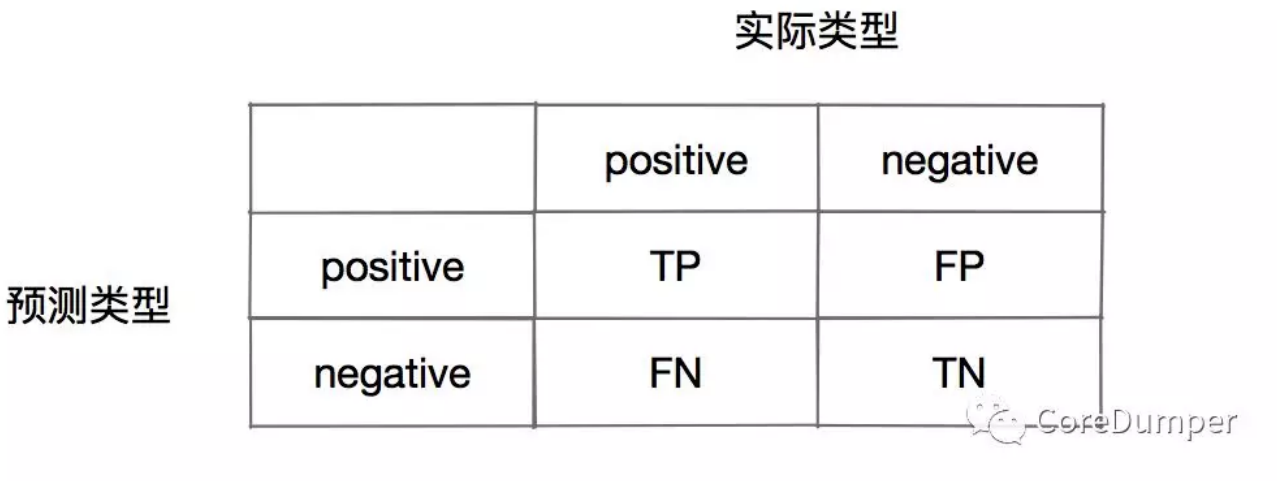
ROC 曲线

上面只是 AUC 概念上的定义，接下来我们看看 AUC 是如何计算的。首先需要了解另外一个概念 ROC（Receiver Operating Characteristic）曲线。

针对一个二分类问题，我们将实例分成正类（positive）和负类（negative）两种。使用分类器对一个实例进行分类预测时，会得到如下四种结果：

- True Positive（TP）：实际为正类，预测为正类；
- False Positive（FP）：实际为负类，预测为正类；
- True Negative（TN）：实际为负类，预测为负类；
- False Negative（FN）：实际为正类，预测为负类。

详细图解如下所示：

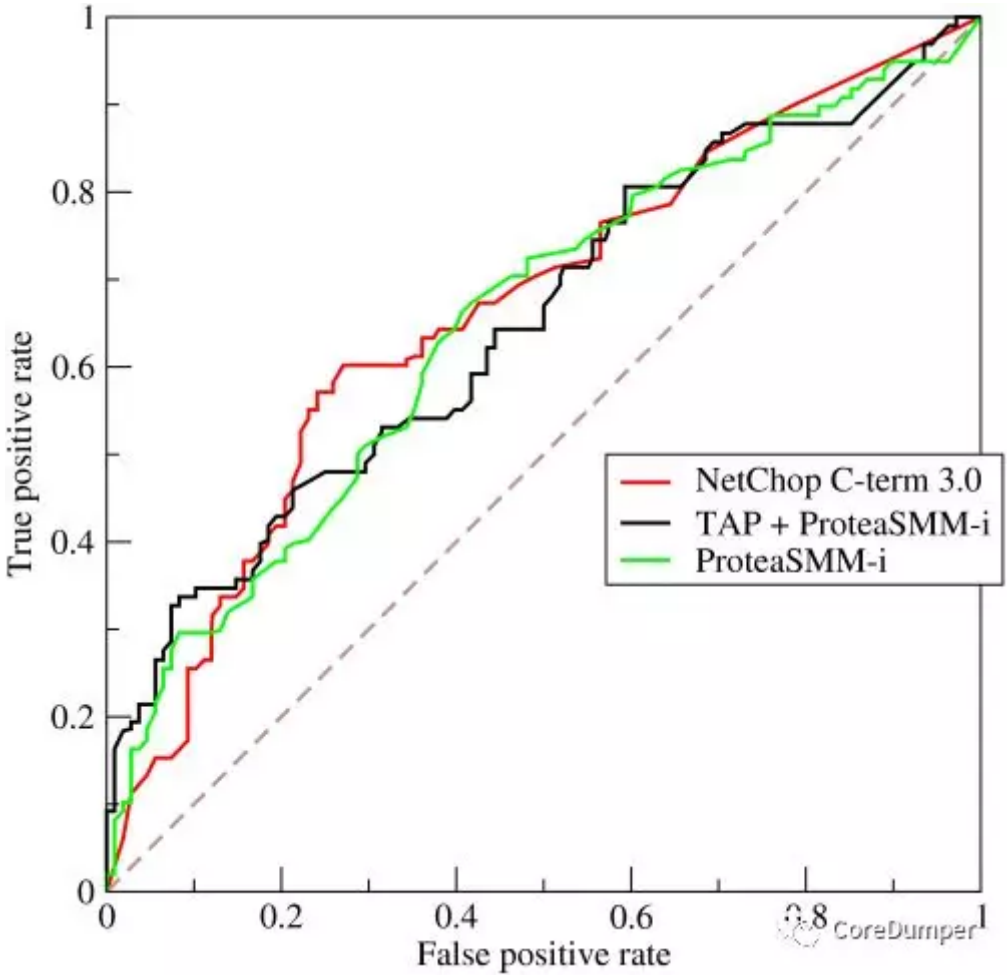


根据 TP、FP、TN、FN 这几个指标可以计算得到如下两个指标：

- True Postive Rate（TPR）：TP/(TP+FN)
- False Postive Rate（FPR）：FP/(FP+TN)

显然，TPR 越大，FPR 越小，表示分类器越优，理想情况下，TPR 应该接近 1，FPR 应该接近 0。

下图是 ROC 曲线的一个实例，以 FPR 为横轴，TPR 为纵轴，三种颜色的曲线分别代表三种分类器的 ROC 曲线。



对于一个特定的分类器和样本集合，显然只能计算得到一对 FPR 和 TPR，也就是曲线上的一个点，而要得到一个曲线，我们需要一系列 FPR 和 TPR 的值，这又是如何得到的呢？

通常分类器在预测某个样本的类型时，会给出这个样本具有多大的概率属于正类，然后根据设定的某个阈值，预测其为正类还是负类。根据某个阈值我们可以计算出相应的一对 FPR 和 TPR，通过改变阈值的大小，就可以计算出一系列的 FPR 和 TPR 了。随着阈值的逐渐减小，越来越多的样本被划分为正类，但是这些正类中同样也掺杂着真正的负类，即 TPR 和 FPR 会同时增大。当阈值取最大值 1 时，对应坐标点为 (0,0)，当阈值取最小值 0 时，对应坐标点为 (1,1)。

AUC 值的计算

AUC 被定义为 ROC 曲线下的面积（从其英文名便可看出），显然这个面积的数值不会大于 1。

由于 ROC 曲线越靠近坐标点 (0,1) 分类器越优，所以从 AUC 判断分类器优劣的标准如下：

- AUC = 1：完美分类器，采用这个预测模型时，存在至少一个阈值能得出完美预测。
- 0.5 < AUC < 1：优于随机猜测。这个分类器妥善设定阈值的话，能有预测价值。
- AUC = 0.5：跟随机猜测效果一样，分类器没有预测价值。
- AUC < 0.5：比随机猜测还差，但只要总是反预测而行，就优于随机猜测。

简单来说就是 AUC 值越大，则分类器越优。

为什么使用 AUC

其实对于分类器有很多种评价标准，为什么工业界会普遍使用 AUC 呢？这是因为 AUC 有个很好的特性：当样本集中的正负样本的分布发生变化的时候，AUC 值能够保持基本稳定。在实际的样本集中经常会出现样本分布不平衡的现象，即负类比正类多很多(或者相反)，而且样本集中的正负样本的分布也可能随时间而发生变化。

相关文章：

机器学习书籍推荐

如果想阅读到完整的文章内容，请点击阅读原文。

阅读原文