理论篇：

代价曲线

一．代价曲线的引入：

代价曲线是一种衡量分类器表现好坏的手段，作用如同AUC，但是代价曲线在传统的AUC上又做了改变。

在说明代价曲线之前先再明确几个观念：

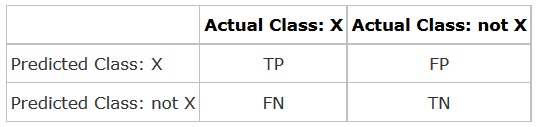
1. **TN:** True Negative （真负），被模型预测为负的样本，模型预测对了

2. **TP:** True Positive （真正），被模型预测为正的样本，模型预测对了

3. **FN:** False Negative （假负），被模型预测为负的样本，模型预测错了

4. **FP:** False Positive （假正），被模型预测为正的样本，模型预测错了

可以通过下图的矩阵来明确这几个的关系：



TPR = TP / (TP + FN)（刻画出的是分类器识别出的正实例占所有正实例的比例）

FPR = FP / (FP + TN)（刻画出的是分类器错以为是正类的负实例占所有负实例的比例）

机器学习后，我们总是可以得到以上的结果，如果将FPR作为横轴，TPR作为纵轴，我们即可得到其ROC曲线，如下图1所示：

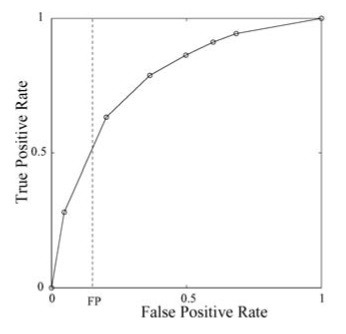
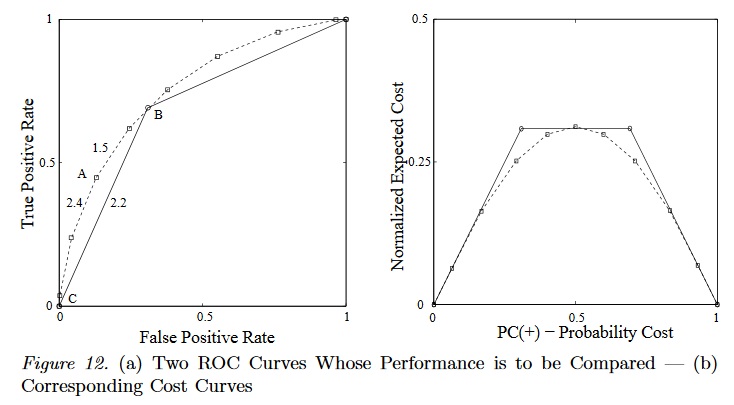
 

图1 图2

但是传统的ROC曲线存在着缺陷，因为对于二分类问题，假设是一个把人分成有病和无病的分类器，显然将有病的人分成无病的代价要远远比无病的人分成有病的代价要来得大，我们希望机器学习所得到的代价是最小的，对此我们可以对不同分类错误赋予不同的代价，如上图2所示。

对于ROC曲线和代价曲线，曲线下方面积的大小都反映该机器学习性能的好坏。**不过不相同的是在ROC曲线中，下方面积越大，说明分类效果越好，而在代价曲线中，下方面积越小，意味着代价越小，则分类效果越好。**

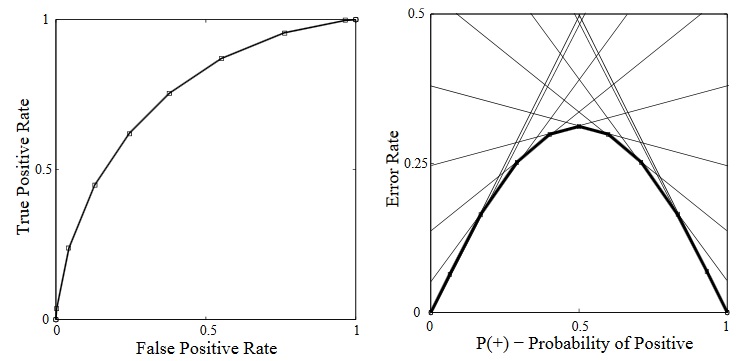


从该图中我们可以看出原本在下方的ROC曲线变为代价曲线后却在上方。

二．代价曲线的表示原理：

**不将代价考虑进去的时候：**

我们首先研究的是error rate,因为代价可能随着训练数据的不同而变化，所以先研究代价相同时的error rate 的情况。



在ROC中的一个点是代价空间中的一条线，ROC中的一条线代表代价空间中的一个点

若给定代价空间中的一点（，则其相应的在ROC中的线为：

相应若给定代价空间中的一条线,则其在ROC中的点为：

**ROC中的点在代价曲线图中将是连接 和 。**

这样子我们就可以通过ROC曲线上一系列点集得到代价空间中的一系列直线，对于直线集，总是取纵坐标最小的直线作为代价曲线中的线段，我们即可得到简单化的代价曲线。

**将代价考虑进去的时候：**

对于考虑代价的代价曲线，我们需要将横纵坐标作一些变化：

代价曲线的横坐标为，中文名称正例概率代价：

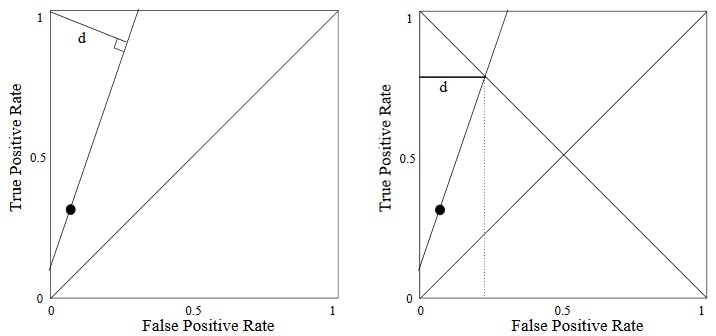
英文名为‘probability times cost’，一般情况下我们还会对PC作一定的归一化处理。

代价曲线的纵坐标即为归一化的代价，一般的表现形式为：

可以看到，当为0时，纵坐标值为FP，当为1时，纵坐标为FN。

所以ROC中的点在代价曲线图中依然将是连接 和 。至此，我们得到了代价曲线的生成方式。因为ROC曲线图中必存在，所以代价曲线中必存在一个类似于三角形的图案。

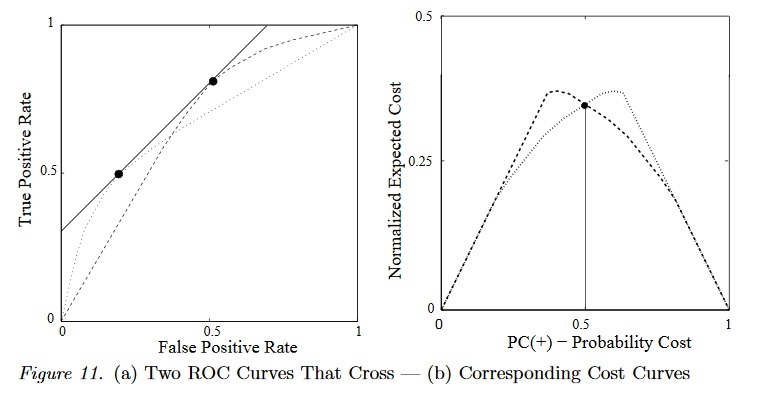
三．代价曲线的性质：



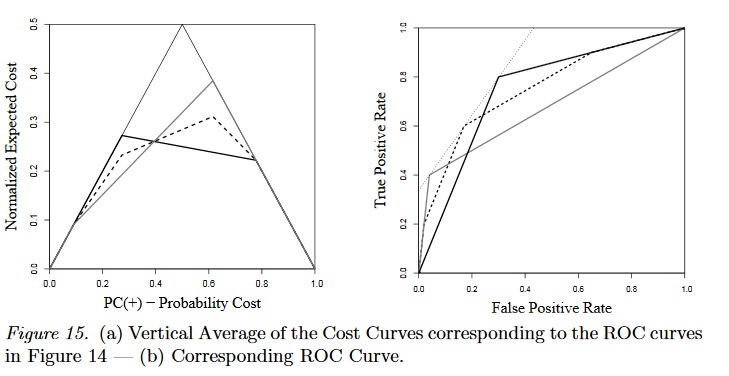
从ROC的一点作一条直线，直线的斜率k表示：

因此确定了ROC中的一点，我们相当于确定了代价空间中的一条直线，而确立了过ROC的点的斜率，我们相当于确定了代价空间中某条直线的横坐标。此刻的代价有两种表示方法，第一种即图1中的d,第二种即图二中的d.

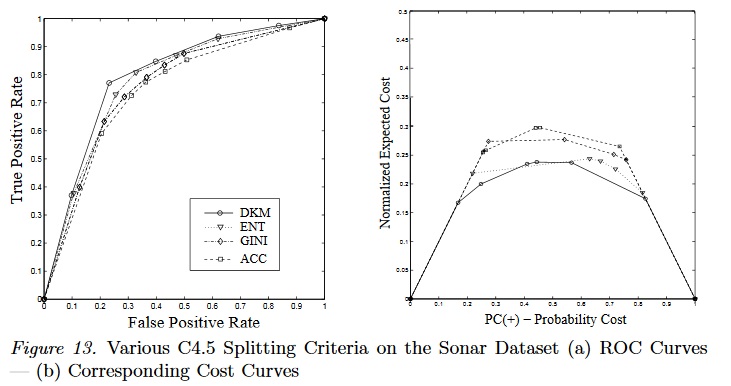
由上我们可知，对于交叉的ROC曲线，若两曲线分别存在一点，两点的连线在两条ROC曲线的上方，则两条ROC曲线对应的代价曲线的相交点横坐标由上述直线的斜率所决定，具体可表示为：



下图中我们可以看到每条ROC折线都由三个点所确定，而我们又可以看到其所对应的代价曲线由三条线段构成，因为存在ROC空间中的一条直线，三个点均在直线上且直线在三条ROC曲线的上方，所以我们可知三条ROC曲线所对应的代价曲线亦交于同一个点。



下图中我们可以看到不同指数的ROC曲线和代价曲线的不同：



代价曲线的缺点：只适用于二分类问题，高维问题需提前处理成二分类问题。