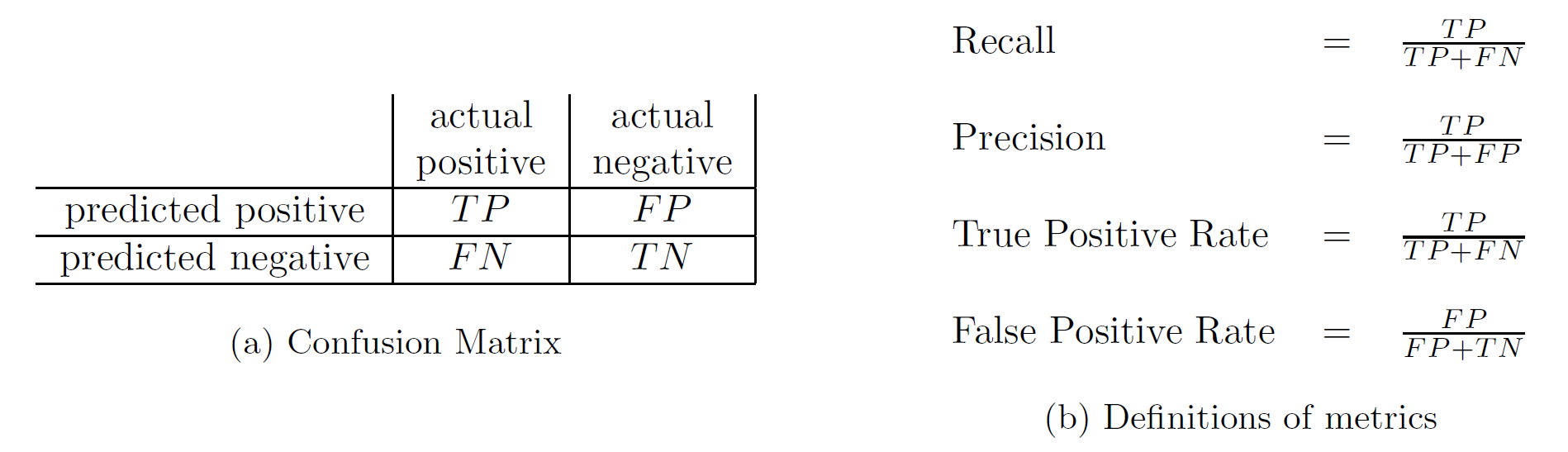
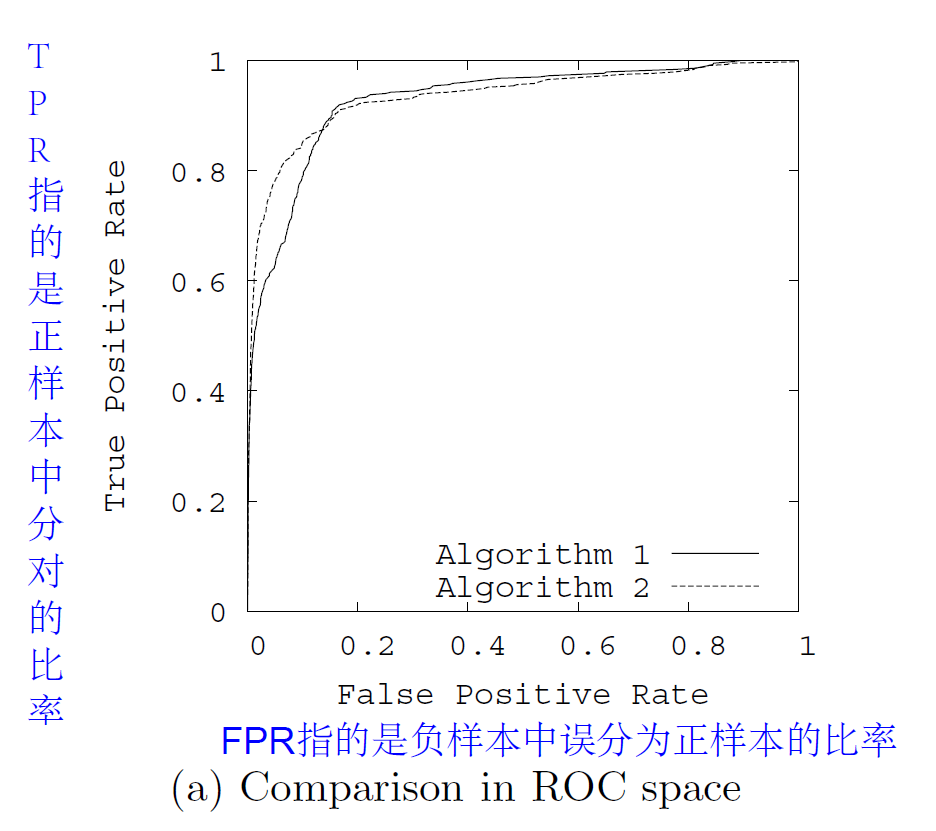
ROC PR Curve Notes

1. 混淆矩阵、计算公式



1. ROC曲线

* 示意图

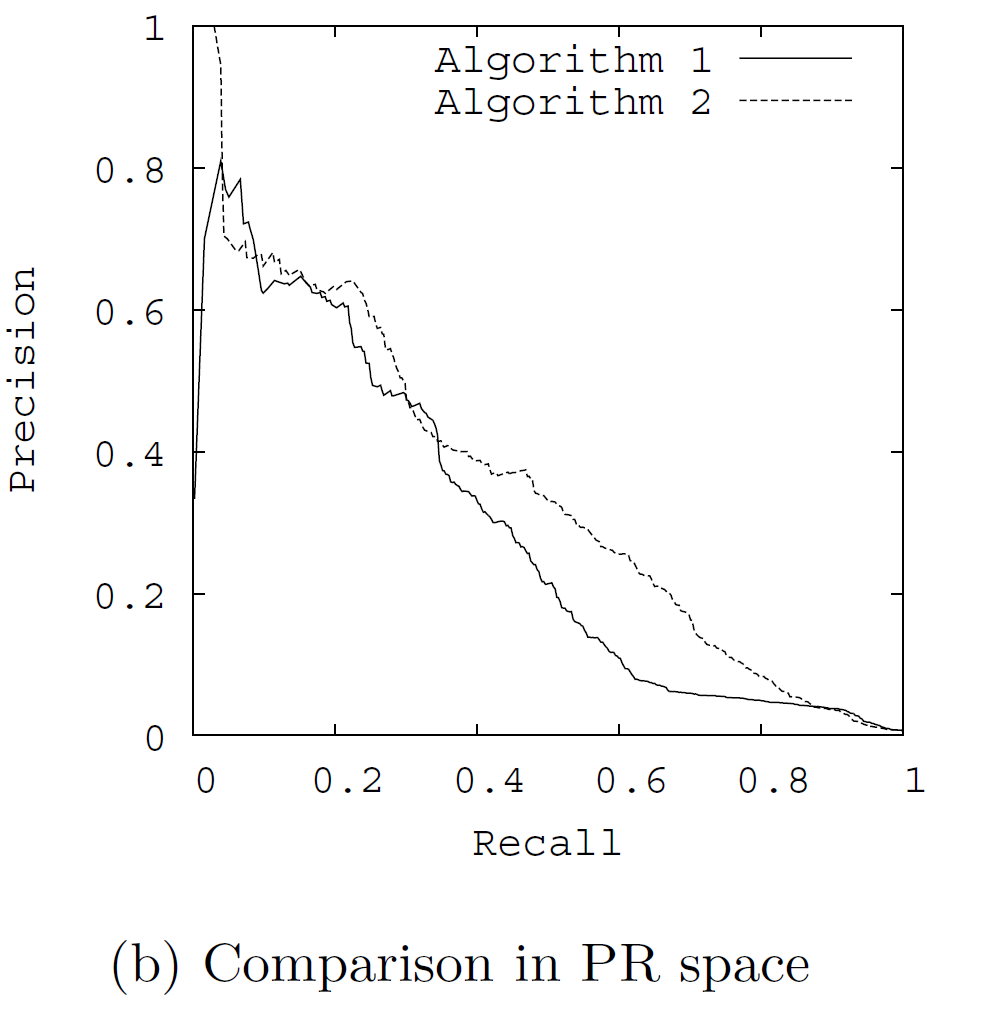


ROC曲线越靠近左上角结果越好

* 对于不均衡样本的数据集，不适合用ROC。样本不均衡一般指的是正样本太少，负样本很容易得到

1. PR曲线

* 示意图



PR曲线越靠近右上角结果越好

* 对于样本不平衡时，适合用这个

1. 绘图

选取不同的阈值进行分类，会得到不同的分类结果，对于每个阈值计算一次这些指标，会得到对应曲线上的一个点，用这些点绘图。

1. AUC

AUC是一个模型评价指标，只能用于二分类模型的评价，对于二分类模型，还有很多其他评价指标，比如logloss，accuracy，precision。如果你经常关注数据挖掘比赛，那你会发现AUC和logloss基本是最常见的模型评价指标。为什么AUC和logloss比accuracy更常用呢？因为很多机器学习的模型对分类问题的预测结果都是概率，如果要计算accuracy，需要先把概率转化成类别，这就需要手动设置一个阈值，如果对一个样本的预测概率高于这个预测，就把这个样本放进一个类别里面，低于这个阈值，放进另一个类别里面。所以这个阈值很大程度上影响了accuracy的计算。使用AUC或者logloss可以避免把预测概率转换成类别。

AUC是ROC曲线下的面积。