机器学习算法系列（19）：机器学习性能评价指标

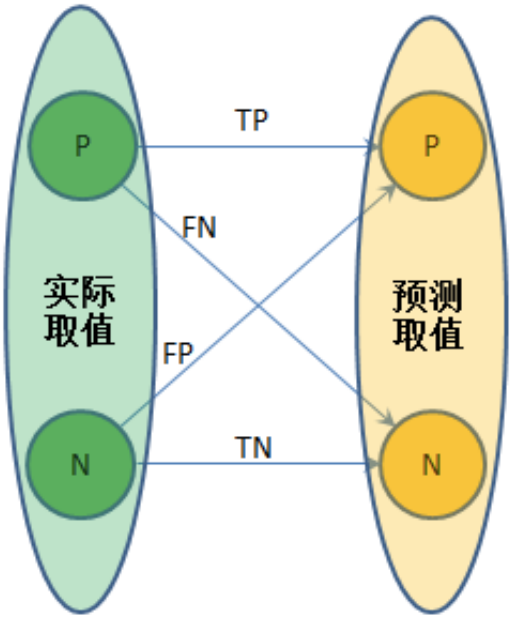
一、分类问题的评价指标

1.1 混淆矩阵

对一个二分类问题，将实例分成正类（postive）或者负类（negative），但在实际分类中，会出现以下四种情况：

- True Positive（真正，TP）：将正类预测为正类数
- True Negative（真负，TN）：将负类预测为负类数
- False Positive（假正，FP）：将负类预测为正类数
- False Negative（假负，FN）：将正类预测为负类数

从下图可以直观的看出四者的关系：



混淆矩阵（Confusion matrix）又被称为错误矩阵，它是一种特定的矩阵来呈现算法性能的可视化呈现。其每一列代表预测值，每一行代表的是实际的类别，这个名字来源于他是否可以非常容易的表明多个类别是否有混淆（也就是一个class被预测为另一个class）混淆矩阵的i行j列是列别i被分为类别j的样本个数。

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

1.2 精确率、召回率与F1值

精确率（precision rate）定义为：

$$P = \frac{TP}{TP + FP}$$

这里需要注意的是精确率（precision）和准确率（accuracy）是不一样的

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

在非平衡数据的情况下，准确率这个评价指标有很大的缺陷。比如在互联网广告里面，点击的数量是很少的，一般只有千分之几，如果用Accuracy，即使全部预测成负类（不点击），ACC也达到了99%以上，这就没有意义了。

召回率（Recall rate）定义为：

$$R = \frac{TP}{TP + FN}$$

此外，还有F1值，它是精确率和召回率的调和均值，即

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$$

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

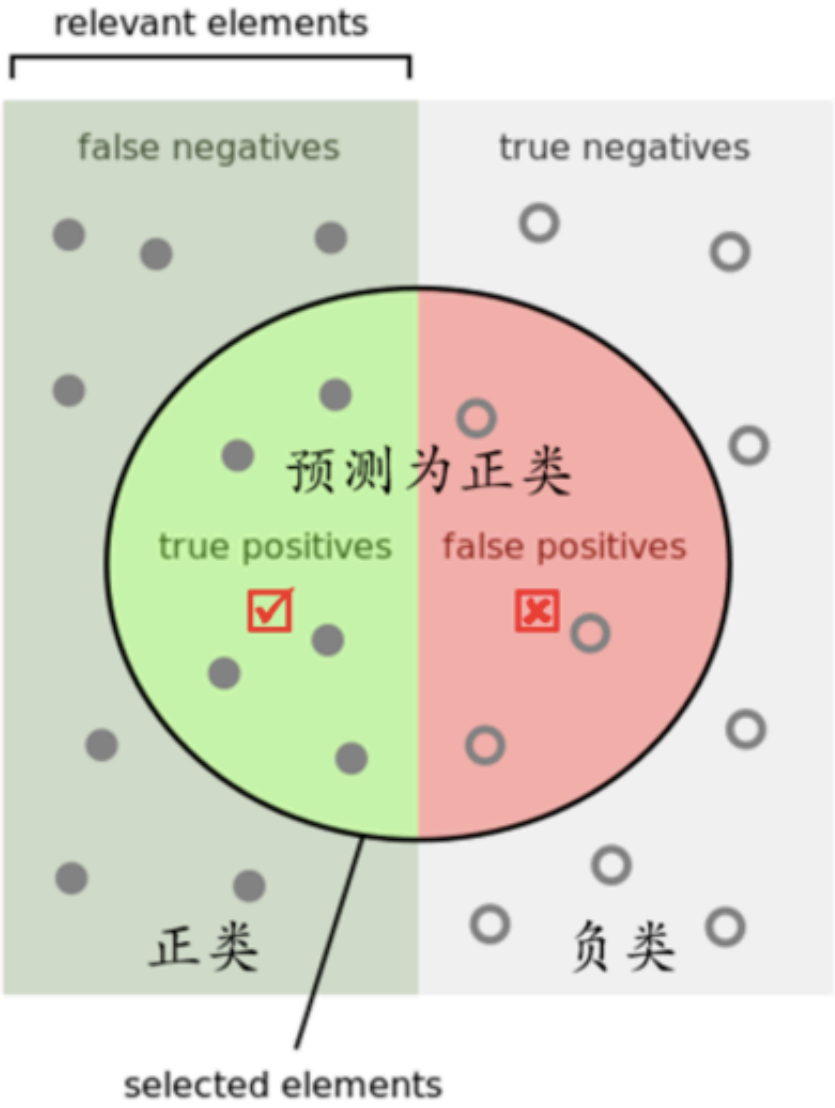
精确率与召回率都很高时，F_1值也会很高。

1.4 通俗理解

通俗来讲，精确率是针对我们的预测结果而言的，他表示的是预测为正的样本中有多少是对的，那么预测为正就有两种可能了，一种就是把正类预测为正类（TP），另一种就是把负类预测为正

类（FP）。

而召回率是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。那也有两种可能，一种是把原来的正类预测成正类（TP），另一种就是把原来的正类预测为负类



（FN）。

在信息搜索领域，精确率和召回率又被称为查准率和查全率

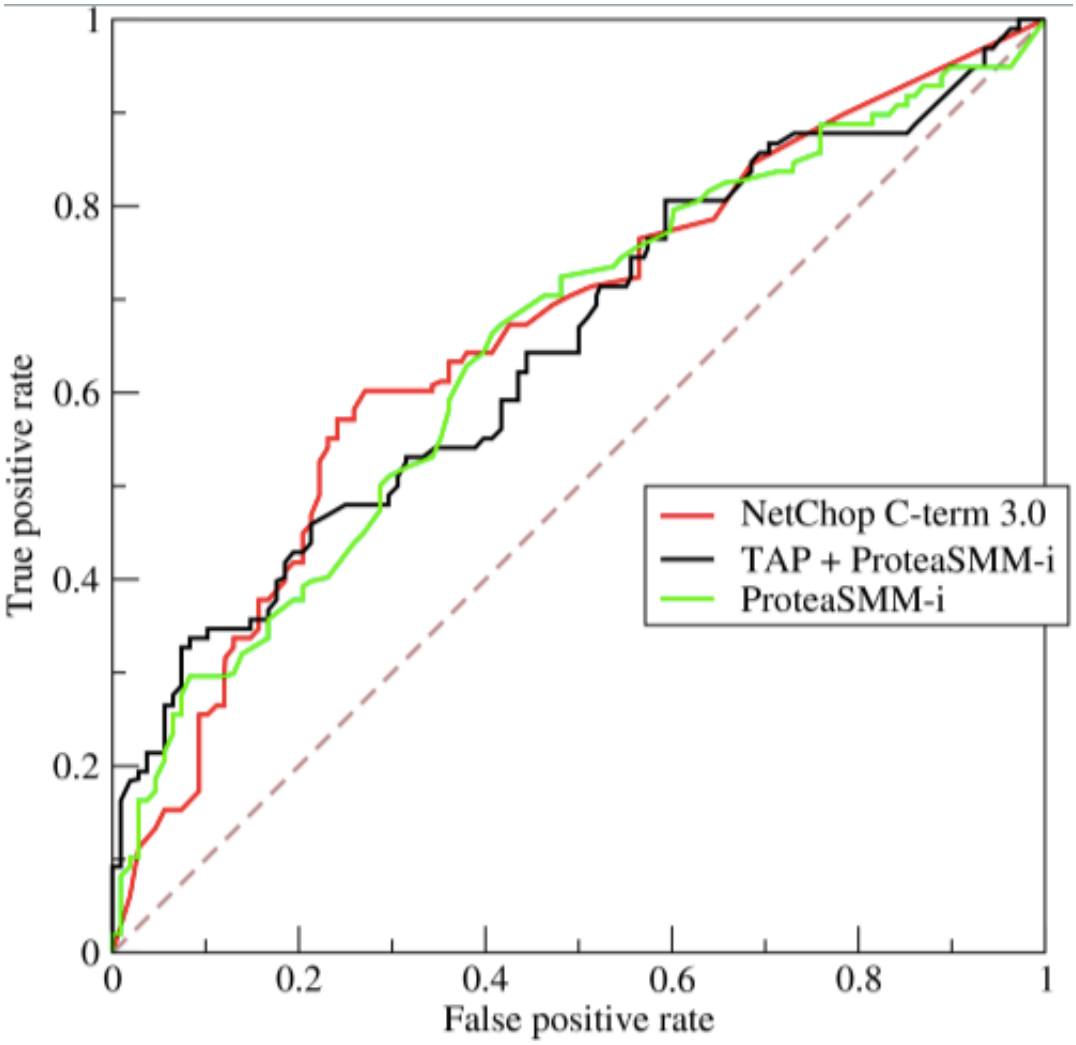
$$查准率 = \frac{检索出的相关信息量}{检索出的信息总量}$$

$$查\ 全\ 率 = \frac{检\ 索\ 出\ 的\ 相\ 关\ 信\ 息\ 量}{系\ 统\ 中\ 的\ 相\ 关\ 信\ 息\ 总\ 量}$$

1.5 ROC曲线

ROC曲线首先是由二战中的电子工程师和雷达工程师发明的，用来侦测战场上的敌军载具（飞机、船舰），也就是信号检测理论。之后很快就被引入了心理学来进行信号的知觉检测。数十年来，ROC分析被用于医学、无线电、生物学、犯罪心理学领域中，而且最近在机器学习（machine learning）和数据挖掘（data mining）领域也得到了很好的发展。

下图是一个ROC曲线的示例图。



在这个ROC曲线的示例图中，横坐标为false positive rate(FPR)，纵坐标为true positive rate（TPR）。由混淆矩阵可得到横纵轴的计算公式。

- 1)

$$TPR = \frac{TP}{TP + FN}$$

代表分类器预测的正类中实际正实例占所有正实例的比例。直观上代表能将正例分对的概

率。

- 2）

$$FPR = \frac{FP}{FP + TN}$$

代表分类器预测的正类中实际负实例占所有负实例的比例。直观上代表将负类错分为正例的概率。

假设采用逻辑回归分类器，其给出针对每个实例为正类的概率，那么通过设定一个阈值如0.6，概率大于等于0.6的为正类，小于0.6的为负类。对应的就可以算出一组(FPR,TPR)，随着阈值的逐渐减小，越来越多的实例被划分为正类，但是这些正类中同样也掺杂着更多的负实例，即TPR和FPR会同时增大。阈值最大时，对应坐标点（0，0），阈值最小时，对应坐标点（1，1）。

接下来我们考虑ROC曲线图中的四个点和一条线。第一个点，(0,1)，即FPR=0, TPR=1，这意味着FN（false negative）=0，并且FP（false positive）=0。这是一个完美的分类器，它将所有的样本都正确分类。第二个点，(1,0)，即FPR=1，TPR=0，类似地分析可以发现这是一个最糟糕的分类器，因为它成功避开了所有的正确答案。第三个点，(0,0)，即FPR=TPR=0，即FP（false positive）=TP（true positive）=0，可以发现该分类器预测所有的样本都为负样本（negative）。类似的，第四个点（1,1），分类器实际上预测所有的样本都为正样本。经过以上的分析，我们可以断言，ROC曲线越接近左上角，该分类器的性能越好。

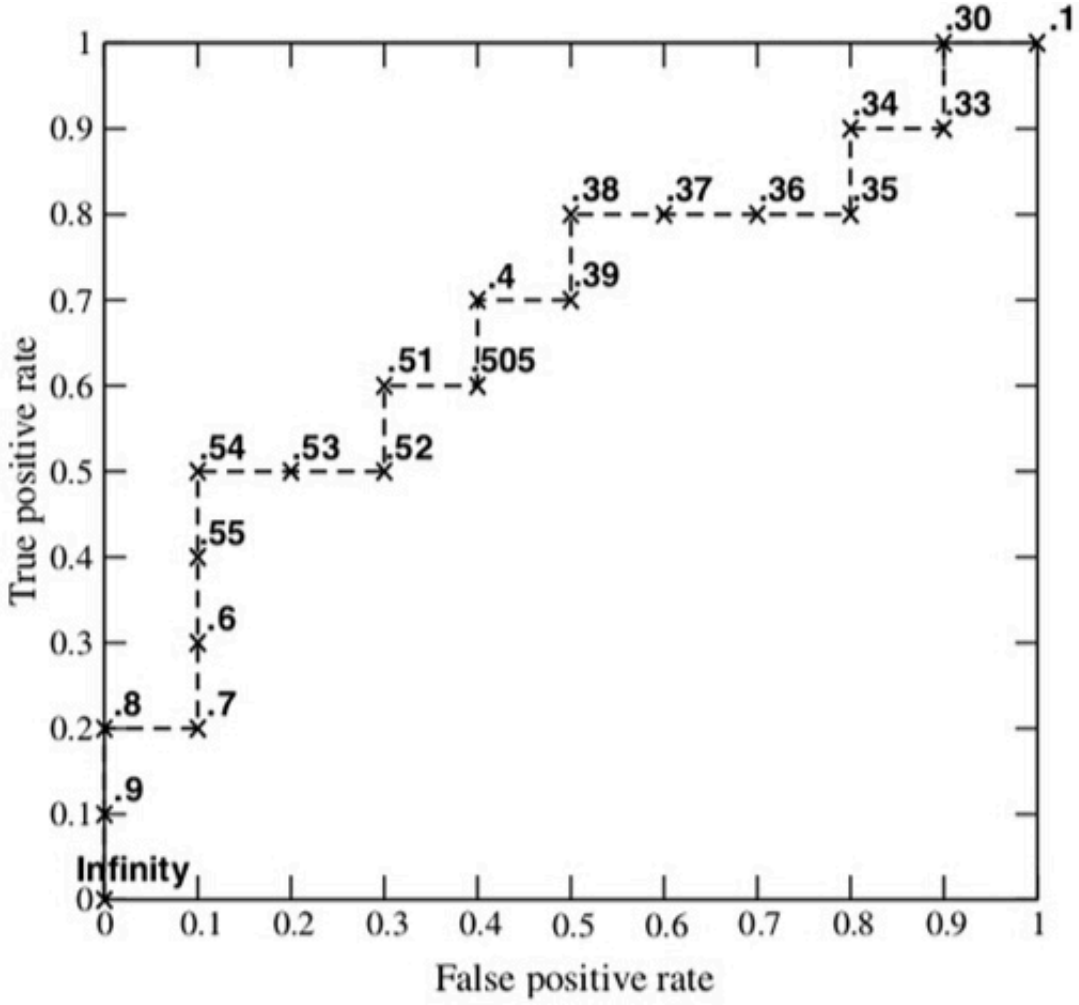
下面考虑ROC曲线图中的虚线y=x上的点。这条对角线上的点其实表示的是一个采用随机猜测策略的分类器的结果，例如(0.5,0.5)，表示该分类器随机对于一半的样本猜测其为正样本，另外一半的样本为负样本。

如何绘制ROC曲线呢?

假设已经得出一系列样本被划分为正类的概率，然后按照大小排序，下图是一个示例，图中共有20个测试样本，“class”一栏表示每个测试样本真正的标签（P表示正样本，n表示负样本），“Score”表示每个测试样本属于正样本的概率。

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

接下来，我们从高到低，依次将“Score”值作为阈值的threshold，当测试样本属于正样本的概率大于或等于这个threshold时，我们认为它为正样本，否则为负样本。举例来说，对于图中的第四个样本，其“Score”值为0.6，那么样本1，2，3，4都被认为是正样本，因为它们的“Score”值都大于等于0.6，而其他样本则都认为是负样本。每次选取一个不同的threshold，我们就可以得到一组FPR和TPR，即ROC曲线上的一点。这样一来，我们一共得到了20组FPR和TPR的值，将它们画在ROC曲线的结果如下图：



1.6 AUC

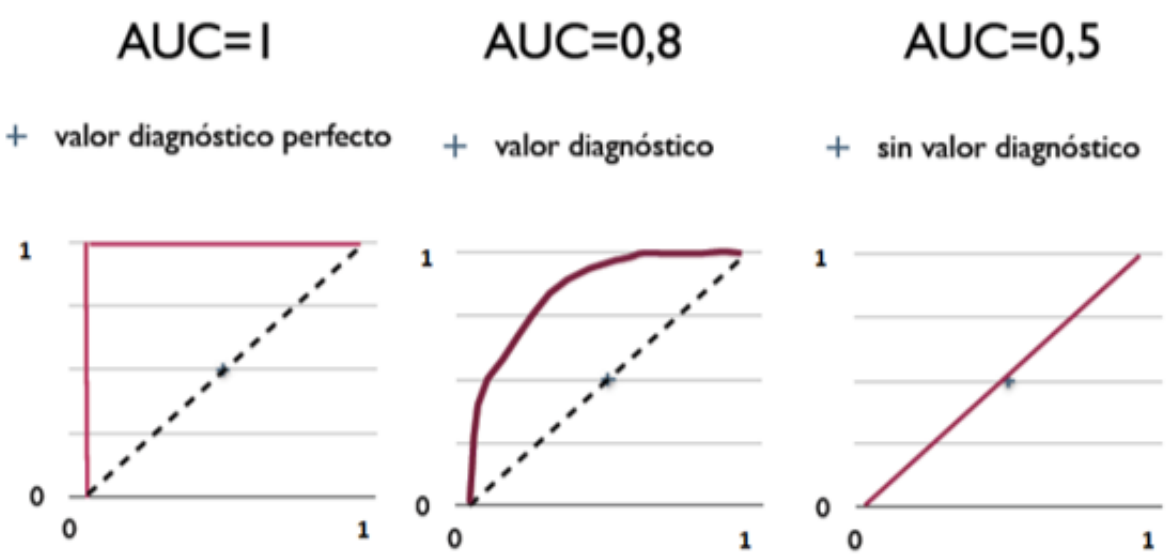
AUC（Area under Curve）指的是ROC曲线下的面积，介于0和1之间。AUC作为数值可以直观地评价分类器的好坏，值越大越好。

> The AUC value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example.

首先AUC是一个概率值，当你随机挑选一个正样本以及负样本，当前的分类算法根据计算得到的Score值将这个正样本排在负样本前面的概率就是AUC值，AUC值越大，当前分类算法越有可能将正样本排在负样本前面，从而能够更好地分类。

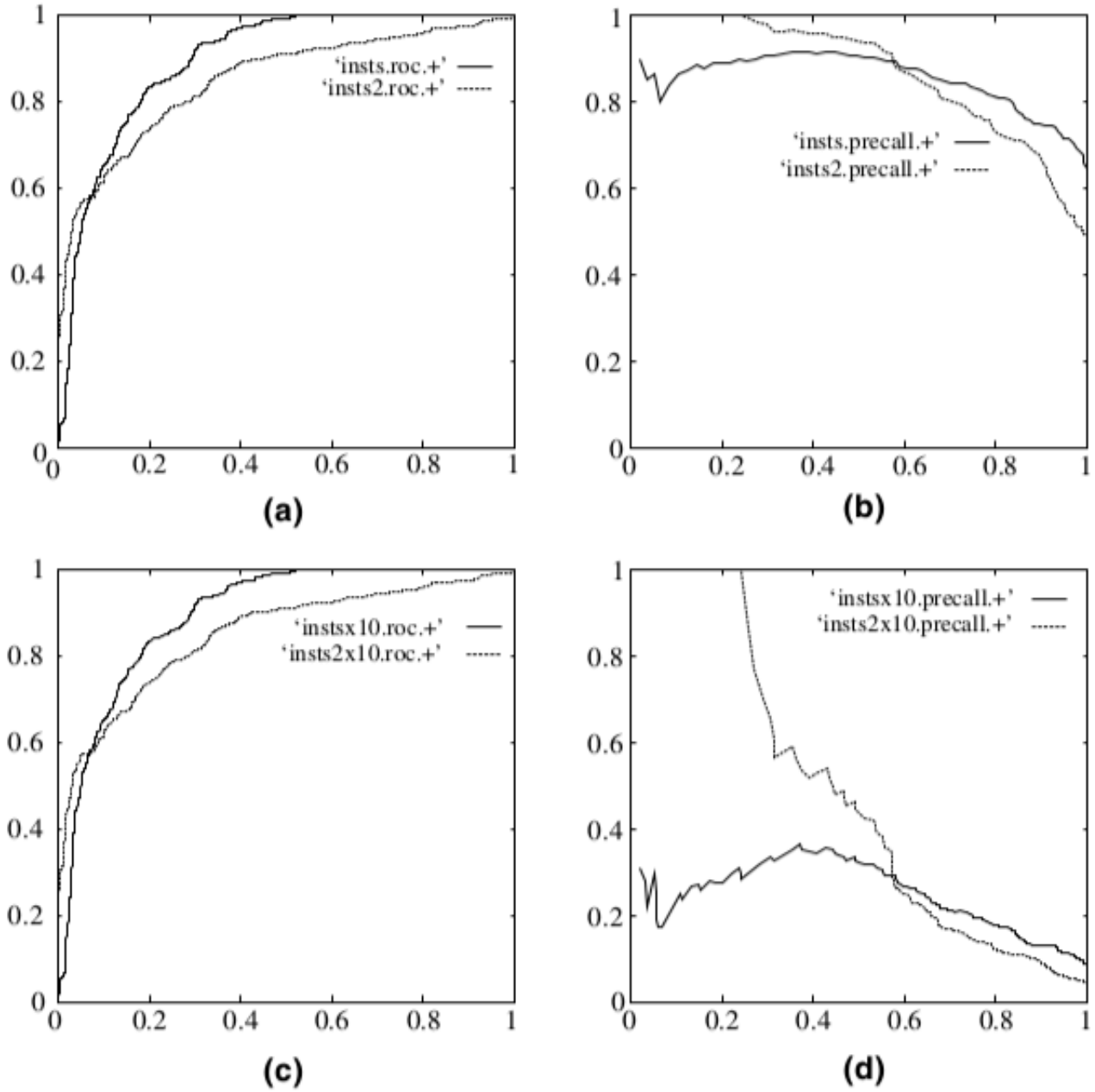
以下是根据AUC判断分类器优劣的标准：

- 1）AUC=1，是完美分类器，采用这个预测模型时，存在至少一个阈值能得出完美预测。绝大多数场合，不存在完美的分类器。
- 2）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。
- 3）AUC=0.5，跟随机猜测一样（如丢硬币），模型没有预测价值。
- 4）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。



那我们为什么使用ROC曲线呢?

既然已经有那么多的评价标准，为何还要使用ROC和AUC曲线呢？因为ROC曲线有个很好的特性：当测试集中的正负样本的分布变化的时候，ROC曲线能够保持不变。在实际的数据集中经常会出现非平衡数据的现象，即负样本比正样本多很多（或者相反），而且测试数据中的正负样本的分布也可能随着时间变化。下图是ROC曲线和Precision-Recall曲线的对比：



在上图中，a和c为ROC曲线，b和d为Precision-Recall曲线。a和b展示的是分类器在原始测试集（正负样本分布平衡）的结果，c和d是将测试集中负样本的数量增加到原来的10倍后，分类器的结果。可以明显的看出，ROC曲线基本保持原貌，而Precision-Recall则变化较大。

二、回归问题的评价指标

2.1 平均绝对误差

平均绝对误差MAE（Mean Absolute Reeor）又被称为L1范数损失（L1-norm loss）：

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} |y_i - \hat{y}_i|$$

2.2 平均平方误差

平均平方误差MSE（Mean Squared Error）又被称为L2范数损失（L2-norm loss）：

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} (y_i - \hat{y}_i)^2$$