关于查准率(Precision)与召回率(Recall)关系

实际 \ 预测	正例	反例
正例	TP	FN
反例	FP	TN

(Confusion Matrix)
(混淆矩阵)

TP: true positive (将正样本识别为正样本的数量)
(或概率)

FN: False Negative (将正样本识别为负样本的数量)
(或概率)

FP: False Positive (将负样本识别为正样本的数量)
(或概率)

TN: True Negative (将负样本识别为负样本的数量)
(或概率)

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

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

为了方便起见，下面的讨论都是以概率为基准。

首先明白三个关系：

① $TP+FN=1$

② $FP+TN=1$

③ 对同一个系统，若TP增加，则FP也增加。

①是因为将正例识别为正例的概率加上将正例识别为反例的概率相加正好是1。②是同样道理。

③的道理是这样：一个系统在自身性能不变的前提下，如果将更多的正例识别为正例（TP↑），那么一定也会将更多的反例识别为正例（FP↑）。小平同志的[redacted]来。”说的也是这个道理。

因此：

TP增加 ←③→ FP增加 ←②→ TN减少

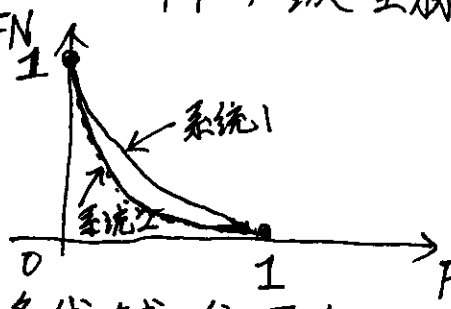
↕①

FN减少

下面介绍常见的表征系统性能的指标。

① ROC曲线（Receiver Operator Characteristic Curve）

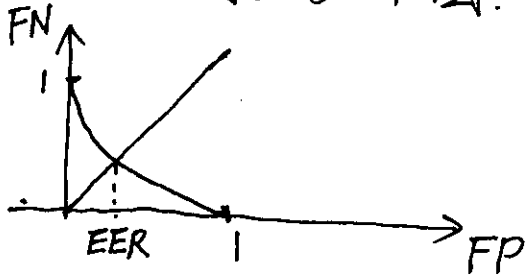
横坐标FP，纵坐标FN



ROC曲线

这条线越往原点处凹陷，系统性能越好。例如上图中，系统2性能就比系统1好。

有时，我们用唯一的数来表示系统性能，这个数叫Equal Error Rate（等错误率），就是FP=FN时的FP（或FN）值。用一条斜率为1的线交ROC曲线就可求得。如下图：



个人认为周志华老师<<机器学习>> P31 图2.3 错了，画的应该是ROC曲线，而不是PR曲线。