ROC、AUC

混淆矩阵

混淆表(confusion table)		分类器预测的类别	
		y1	y2
实际的类别	y1	C11	C12
	y2	C21	C22

- 准确度Accuracy：(C11+C22)/(C11+C12+C21+C22)
- 正确率Precision（y1）：C11/(C11+C21)
- 召回率Recall（y1）：C11/(C11+C12)

评 测 指 标

混淆表(confusion table)		分类器预测的类别	
		军事	科技
实际的类别	军事(60)	50	10
	科技(40)	5	35

- 准确度Accuracy：(50+35)/(35+5+10+50)=85%
- 正确率Precision（y1）：50/(50+5)=90.9%
- 召回率Recall（y1）：50/(50+10)=83.3%

评 测 指 标

混淆表（confusion table）		分类器预测的类别	
		+1	-1
实际的类别	+1	真正例（TP）	伪反例（FN）
	-1	伪正例（FP）	真反例（TN）

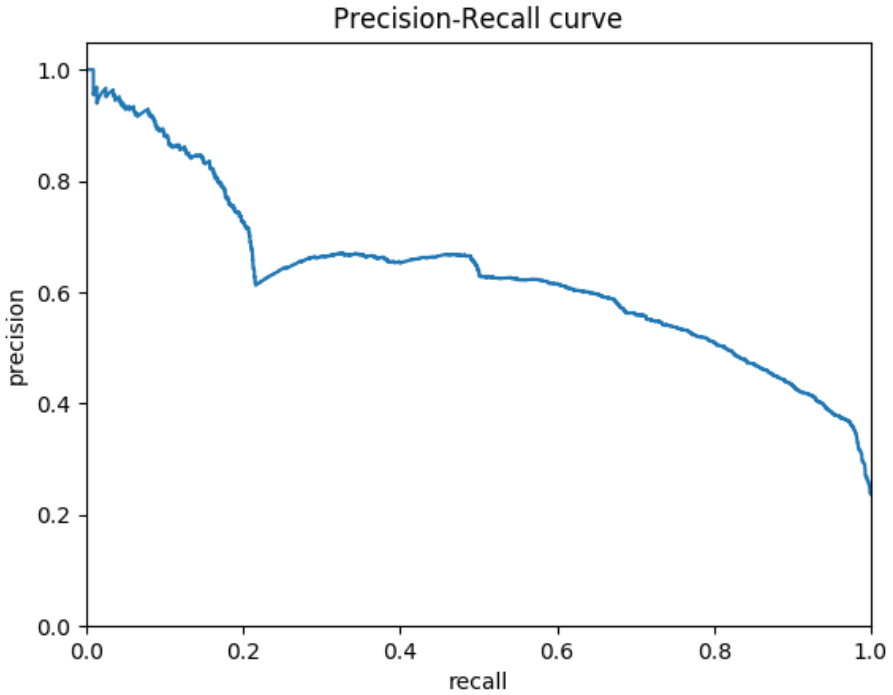
- 正确率Precision：TP/(TP+FP)
  - 预测为正例的样本中的真正正例的比例
- 召回率Recall：TP/(TP+FN)
  - 预测正例的真实正例占所有真实正例的比例

评 测 指 标

- 很容易构造一个高正确率或高召回率的分类器，但很难保证两全其美
- 一般情况下准确率高、召回率就低，召回率高、准确率低

- 搜索场景：
  - 保证召回为前提，提升准确
- 疾病监测、反垃圾场景：
  - 保证准确为前提，提升召回

正确率



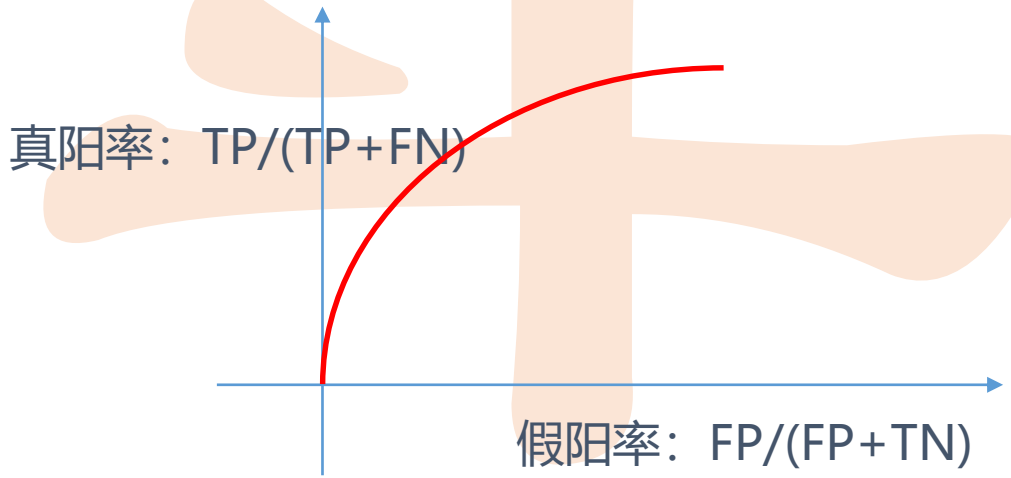
召回率

ROC

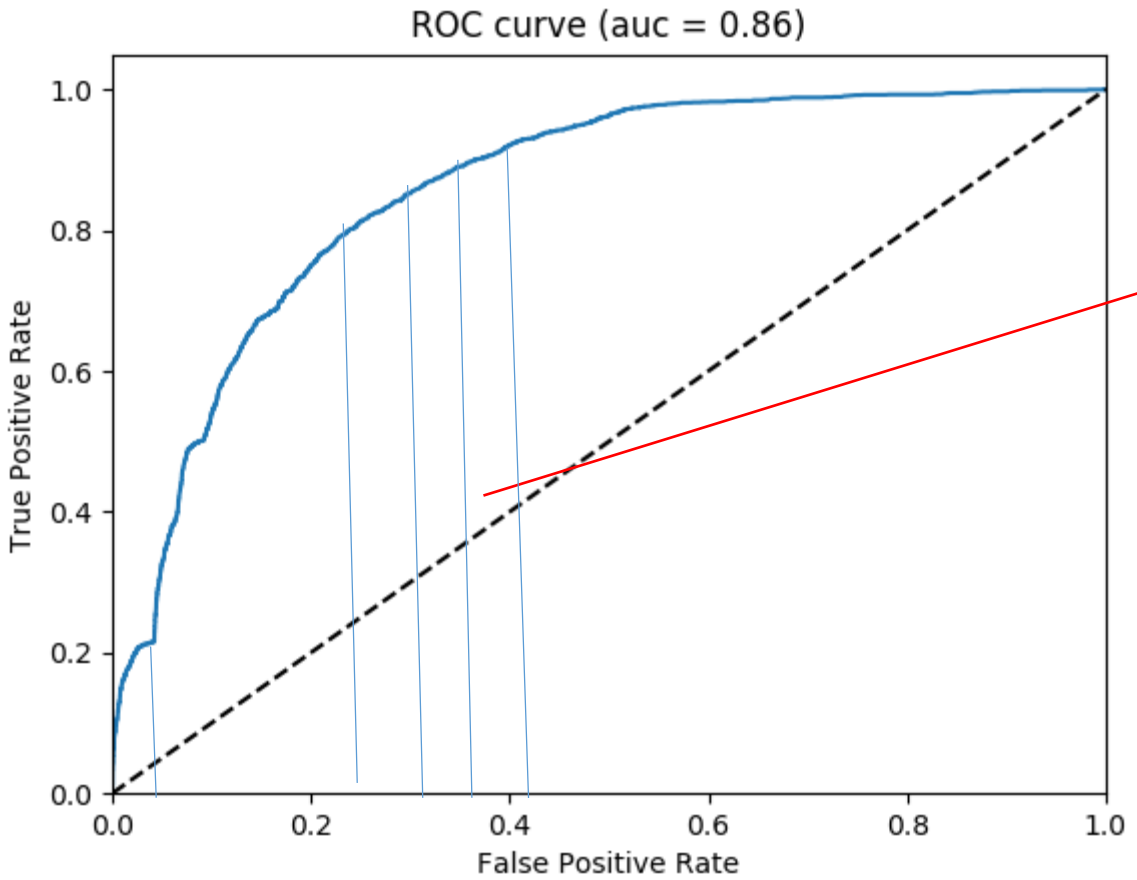
- ROC是个曲线

混淆表（confusion table）		分类器预测的类别	
		+1	-1
实际的类别	+1	真正例（TP）	伪反例（FN）
	-1	伪正例（FP）	真反例（TN）

- 纵轴：真阳率：TP/(TP+FN)
  - Recall
- 横轴：假阳率：FP/(FP+TN)



ROC、AUC



真阳率：TP/(TP+FN)

AUC
(area under curve)
ROC曲线下的面积

假阳率：FP/(FP+TN)

快 速 方 法

- 另一种理解AUC的方法：
  - 负样本排在正样本前面的概率

- cat auc.raw | sort -t$'\t' -k2g |awk -F'\t' '($1==-1){++x;a+=y;}($1==1){++y;}END{print 1.0-a/(x*y);} ‘

- x*y：正负样本pair对
- a：错误的pair对
- a/x*y：错误的概率
- 1-a/x*y：正确的概率

Q & A

@八斗学院