

# 机器学习理论和实践

## 二分类模型评价指标



# 目录

---

- 混淆矩阵、正确率、召回率、精确率、F1
- ROC曲线和AUC
- 代码实践：`sklearn.metrics`

# 二分类评价指标

P或N都是针对于预测而言的

- T: True; F: False; P: positive; N: negative

混淆矩阵 (Confusion Matrix)		预测标签	
		0 ( 负例 )	1 ( 正例 )
真实标签	0 ( 负例 )	TN ( 真负样本数量 )	FP ( 假正样本数量 )
	1 ( 正例 )	FN ( 假负样本数量 )	TP ( 真正样本数量 )

正确率(*accuracy*) =  $\frac{TP + TN}{TN + FN + FP + TP}$

召回率(*recall*) =  $\frac{TP}{FN + TP}$

精确率(*precision*) =  $\frac{TP}{FP + TP}$

## 二分类评价指标

- 一位警察抓了很多犯人，但是這些犯人當中，只有少部分真正有罪，其他都是被冤枉的。「寧可錯抓一百，也不可放過一個」
  - recall高，因為該抓與不該抓的犯人都被抓到了
  - precision低，因為很多都是沒犯罪的人
- 一個警察非常嚴謹，只逮捕真正有犯罪的人，不抓實在是沒辦法肯定的犯人。「寧可錯放一百，也不可冤枉一個」
  - precision高，因為通常被抓到人的都是有罪的，
  - recall低，因為不小心放掉一大群犯人

$$F_1 = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}}$$

# 二分类评价指标

```
In [3]: y_true = [1] * 8 + [0] * 12
        executed in 5ms, finished 22:44:49 2021-07-11

In [4]: y_pred = [1] * 6 + [0] * 10 + [1] * 4
        executed in 4ms, finished 22:44:49 2021-07-11

In [5]: y_true
        executed in 11ms, finished 22:44:50 2021-07-11

Out[5]: [1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

In [6]: y_pred
        executed in 5ms, finished 22:44:50 2021-07-11

Out[6]: [1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1]
```

正确率(*accuracy*) =  $\frac{TP+TN}{TN+FN+FP+TP}=?$

精确率(*precision*) =  $\frac{TP}{FP + TP} = ?$

召回率(*recall*) =  $\frac{TP}{FN+TP}=?$

混淆矩阵 (Confusion Matrix)		预测标签	
		0 ( 负例 )	1 ( 正例 )
真实标签	0 ( 负例 )	TN= ?	FP= ?
	1 ( 正例 )	FN=?	TP=?

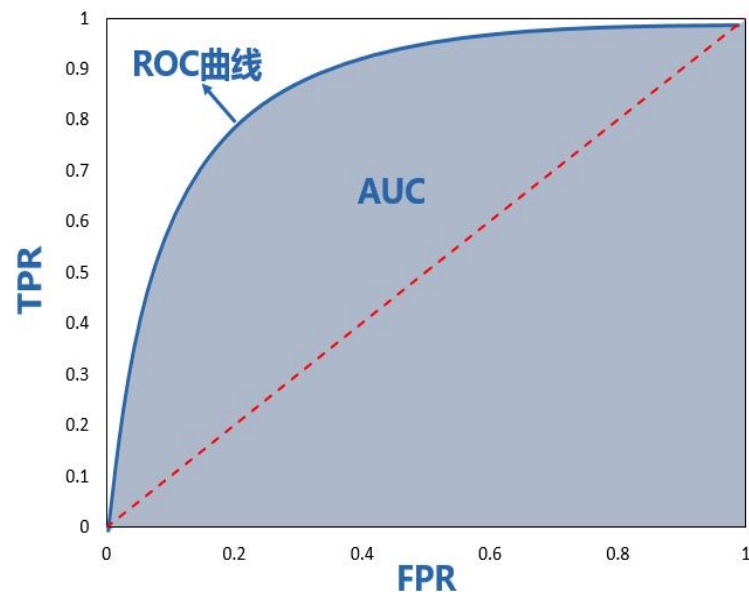
$F_1 = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} = ?$

# ROC曲线与AUC

- ROC曲线：
  - 横轴为 $FPR$ ，纵轴为 $TPR$
- AUC(The Area Under the Curve)：
  - ROC曲线与横轴 $FPR$ 之间的面积

$$\begin{aligned} \text{真正率}(TPR) \\ &= \frac{TP}{FN + TP} \end{aligned}$$

$$\begin{aligned} \text{假正率}(FPR) \\ &= \frac{FP}{TN + FP} \end{aligned}$$



# ROC曲线与AUC

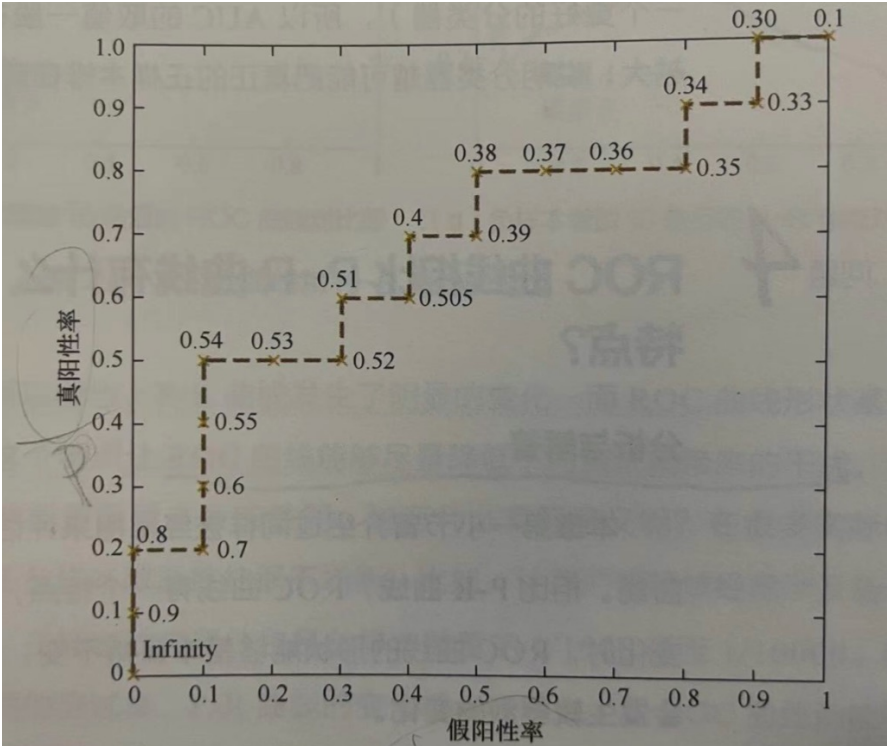
正负样本各10个

- 把全部样本预测为负例：FP和TP都为0，此时：FPR和TPR都为0，获得第一个点坐标(0,0)  
先对预测概率进行排序
- 把阈值调整为0.9时，一号样本被预测为正样本，且该样本确实为正样本，故  
 $TPR = TP / (TP + FN) = 1 / 10$ , 此时没有预测错的正样本，FP=0, 故FPR为0，对应ROC曲线为点(0,0.1)
- 阈值为0.7时：FPR=1/(9+1); TPR=2/(2+8)

假阳性率(FPR) =  $\frac{FP}{TN + FP}$

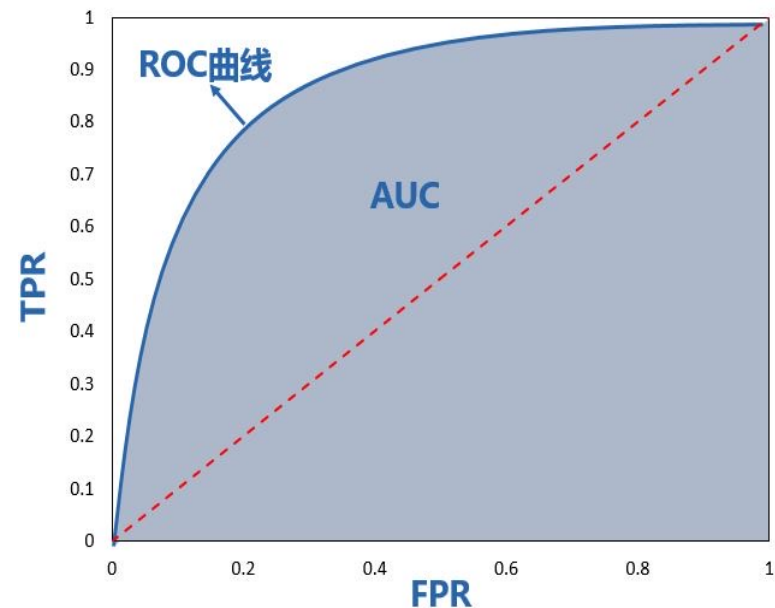
真阳性率(TPR) =  $\frac{TP}{TP + FN}$

样本序号	真实标签	模型输出 概率	样本序号	真实标签	模型输出 概率
1	p	0.9	11	p	0.4
2	p	0.8	12	n	0.39
3	n	0.7	13	p	0.38
4	p	0.6	14	n	0.37
5	p	0.55	15	n	0.36
6	p	0.54	16	n	0.35
7	n	0.53	17	p	0.34
8	n	0.52	18	n	0.33
9	p	0.51	19	p	0.30
10	n	0.505	20	n	0.1



# ROC曲线与AUC的特性

- ROC曲线：
  - 曲线越靠近(0, 1)证明模型整体预测能力越强
- AUC：
  - 数值越大证明模型整体预测能力越强





# 分类指标实践案例

---

- 分类模型评价指标.ipynb