人工智能与深度学习

2-6 二分类模型的衡量指标

王中雷

经济学院、王亚南经济研究院、邹至庄经济研究院

厦门大学

（第一版）

内容摘要

1. 混淆矩阵（Confusion matrix）

2. ROC 曲线

回顾

1．二分类模型

- 逻辑回归模型
- 支持向量机
- 随机森林
- （全连接）神经网络

2．对于大多数二分类模型而言，

- 我们首先估计一个“得分”
- 然后，用一个阈值决定对应的类别

回顾

1．我们利用训练集学习模型

2．我们利用测试集，来衡量所学习模型的泛化性能

3．在本节中，我们考虑模型已经为测试集中中的每个样本给出预测类别

混淆矩阵

1．对于测试集中的真实类别而言，我们有如下两种可能

- 真实正例数（True positive number）：真实正例的个数
- 真实负例数（True negative number）：真实负例的个数

2．一个特定的模型将预测测试集中的标签，我们有如下两种可能

- 预测正例数（Predicted positive number）：预测为正例的个数
- 预测负例数（Predicted negative number）：预测为负例的个数

混淆矩阵

1．对于一个预测模型而言，共有四种可能性

- TP （真正例个数）：正确预测为正例的个数
- FN （假负例个数）：错误地被预测为负例的个数

  (真实标签为正例，但被错误地预测为负例)
- TN （真负例个数）：正确预测为负例的个数
- FP （假正例个数）：错误地被预测为正例的个数

  （真实标签为负例，但被错误地预测为正例）

混淆矩阵

1．混淆矩阵

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

度量指标--准确率

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

1．准确率（Accuracy）

$$\begin{aligned} \text{准确率} &= \frac{\text{正确预测个数}}{\text{全部预测个数}} \\ &= \frac{TP+TN}{TP+TN+FP+FN} \end{aligned}$$

2．备注

- 准确率是一种整体性能指标——它不区分标签的类型
- 当正负类别大致平衡时，准确率是一个不错的衡量指标
- 当类别不平衡时，准确率可能严重高估模型表现，甚至具有误导性

度量指标--精确率

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

1．精确率（Precision）

$$\text{精确率} = \frac{\text{真正例个数}}{\text{真正例个数} + \text{假正例个数}} = \frac{TP}{TP + FP}$$

2．备注

- 关注预测正类结果的可靠性，关注误报（即第一类错误）
- 当“假阳性”的代价很高时，精确率是核心指标
- 在不平衡数据集中，是比准确率更敏感的指标
- 当“真正例个数 + 假正例个数 =0”时无定义（下同）

度量指标--精确率

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

3．缺点

- 忽略了“假阴性”，对漏检不敏感
- 在类别极度不平衡且关注负类时，可能不适用
- 单独使用时信息不全面，精确率通常应结合召回率等指标使用

度量指标--召回率

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

1．召回率（Recall）

$$\text{召回率} = \frac{\text{真正例个数}}{\text{真正例个数} + \text{假负例个数}} = \frac{TP}{TP + FN}$$

2．备注

- 召回率有时也被称为：真阳率（True Positive Rate, TPR）或者敏感度（sensitivity）
- 对于高漏报成本场景而言，召回率是核心衡量指标
- 在不平衡数据集中，是比准确率更敏感的指标
- 衡量模型发现所有正例的能力，即关注漏报（即第二类错误）
- 与精确率结合，提供更全面的模型评价

度量指标--召回率

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

1．缺点

- 可能忽视误报（即假阳性）成本
- 可能牺牲模型效率
- 单独使用时信息不全面，召回率必须与精确率等其他指标一起使用

度量指标--假阳率

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

1. 假阳率（False Positive Rate，FPR）

$$\text{假阳率} = \frac{\text{假正例个数}}{\text{假正例个数} + \text{真负例个数}} = \frac{FP}{FP + TN}$$

2. 备注

- 假阳率有时也被称为：误报率（false alarm rate）、虚警率或者误检率
- 在一些场景下，可直接衡量“社会成本”或“资源浪费”
- 在不平衡数据集中，是比准确率更敏感的指标
- 关注第一类错误
- 与召回率结合，提供更全面的模型评价

度量指标--假阳率

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

1．缺点

- 在分类不平衡场景下的误导性
- 忽略实际成本和后果
- 单独使用时信息不全面，假阳率必须与召回率等其他指标一起使用
- 容易与精确率相混淆（其实，并不是缺点）

假阳率与精确率在误报上的异同

	假阳率	1-精确率
定义	真实负例中， 被错误标记为正例的比例	预测正例中， 实际为负例的比例
公式	FP / (FP + TN)	FP / (FP+TP)
视角	从真实情况出发， 审视模型对真实负例的误报率	从预测结果出发， 审视模型预警的可靠性
意义	误伤好人的概率有多大? 关系到筛查的广泛性代价	虚假警报的可能性是多大? 关系到警报的可信度
备注	真实负例占绝对多数时， FPR可能看起来很低， 但模型可能完全没用	真实负例占绝对多数时， 即使FP只增加几个， 也可能导致精确率``雪崩式"下降， 反映了``狼来了"的严重性

度量指标--F1分数

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

1. F1 分数

$$\text{F1 分数} = \frac{1}{\frac{1}{2}\left(\frac{1}{\text{精确率}} + \frac{1}{\text{召回率}}\right)} = 2 \times \frac{\text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}}$$

2. 备注

- F1 分数直接依赖于 TP、FP 和 FN
- 取值范围为 $[0, 1]$
  - F1 分数等于 1，对应于完美的精确率和召回率
  - F1 分数等于 0，模型在正类识别方面表现极差，或没有产生真正例
- F1 分数适当权衡了精确率与召回率
- 更加关注正例预测的表现平衡

度量指标--F1分数

	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

1．缺点

- 忽略了真负例（TN）
- 调和平均数对极低值非常敏感
- 单一数值可能掩盖模型的具体行为

备注

1．以上五个指标，均是基于一组预测好的标签，进行计算的

2．对于某特定模型而言，

- 如果该模型仅预测类别，则只对应一组衡量指标
- 如果该模型先预测“得分”，则对应多组衡量指标
  - 不同阈值对应不同预测类别

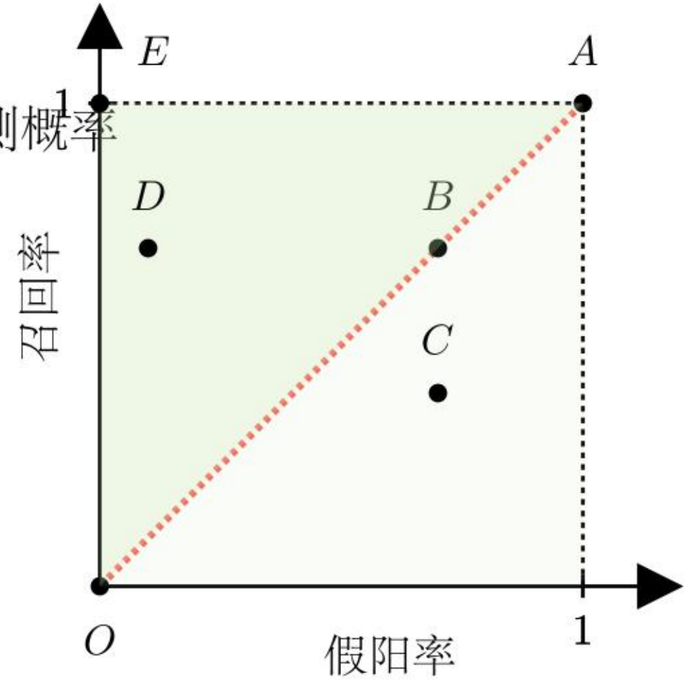
ROC (Receiver Operating Characteristic) 曲线

1. ROC 曲线绘制召回率-假阳率
2. 假设我们知道测试集上模型的预测结果后，
   - 纵轴：召回率 (Recall, 敏感度，Sensitivity)
   - 横轴：假阳率 (FPR，1-特异度 (Specificity))
3. ROC 曲线对应的两种情况
   - 对于只预测标签正负的模型，其对应于一个点
   - 对于预测“得分”的模型，其对应一个曲线

ROC (Receiver Operating Characteristic) 曲线

1. 不同的预测结果，对应着不同的点

- 对角线上的点：无区分能力的随机分类器；不同位置对应不同的正类预测概率
  - $O\,(0,0)$：将所有测试集标签预测成负类
  - $B\,(0.7,0.7)$：以0.7为概率，随机将测试集标签预测成正类
  - $A\,(1,1)$：将所有测试集标签预测成正类
- 对角线上面的点：好模型
  - $D\,(0.1,0.7)$：模型的假阳率低，且能够将 70% 的正类预测正确
  - $E\,(0,1)$：完美模型
  - 越好的模型，对应的点越偏左上角
- 对角线下面的点：实际上，没有那么糟糕
  - $C\,(0.7,0.4)$: 尽管召回率低且假阳率高，原因在于模型把指标搞反了



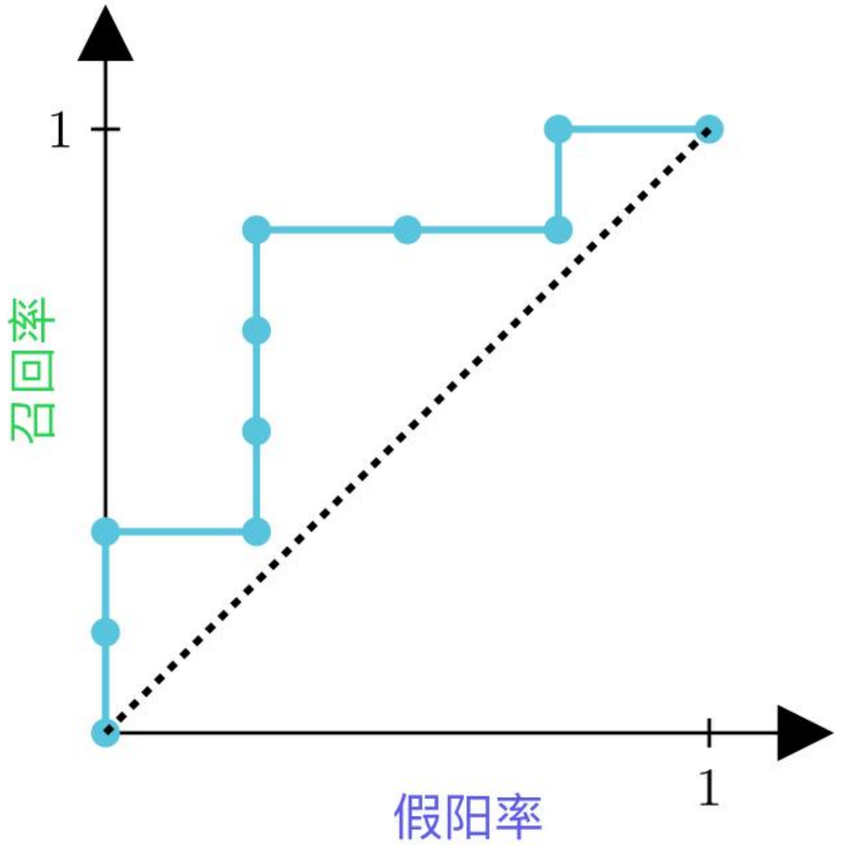
绘制一条ROC 曲线

1．考虑一个预测“得分”的二分类模型，

- 不同的阈值对应着 ROC 曲线上的一个点
- 将这些点连起来，就得到了一条 ROC 曲线

2．我们用一个简单的例子，展示如何绘制一条 ROC 曲线

假阳率	召回率
0	0
0	1/6
0	2/6
1/4	2/6
1/4	3/6
1/4	4/6
1/4	5/6
2/4	5/6
3/4	5/6
3/4	1
1	1



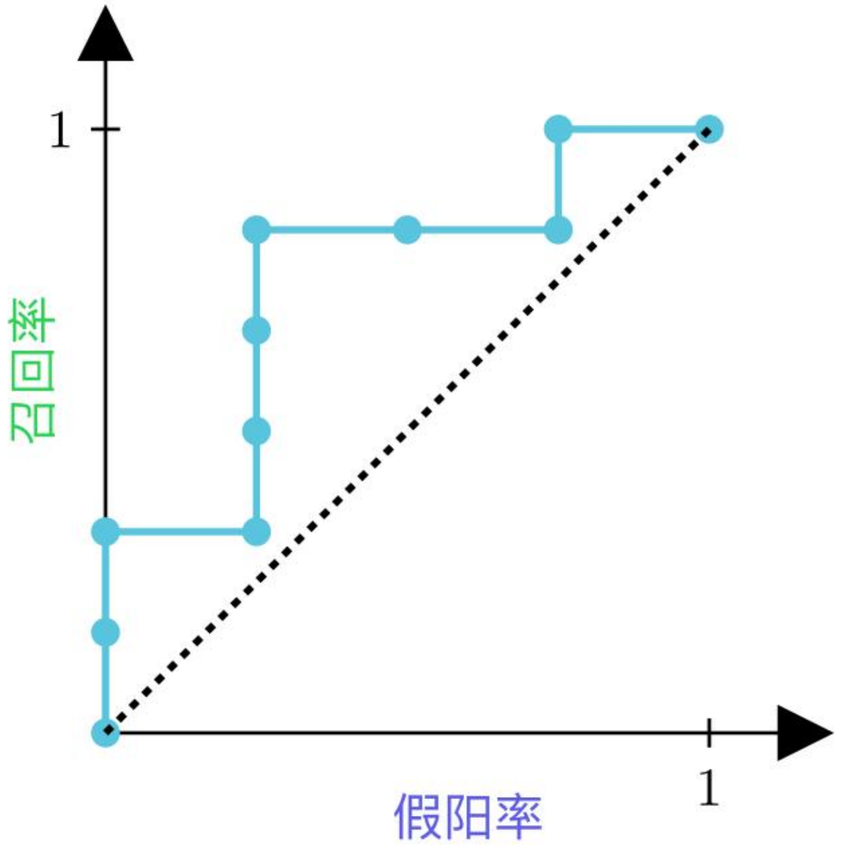
样本编号	真实标签	预测标签	预测得分
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

问题 1：纵向和横向移动的步长分别是多少?

答案：纵向移动步长是 1 除真实正例个数，在本例中为 1/6

横向移动步长是 1 除真实负例个数，在本例中为 1/4

假阳率	召回率
0	0
0	1/6
0	2/6
1/4	2/6
1/4	3/6
1/4	4/6
1/4	5/6
2/4	5/6
3/4	5/6
3/4	1
1	1

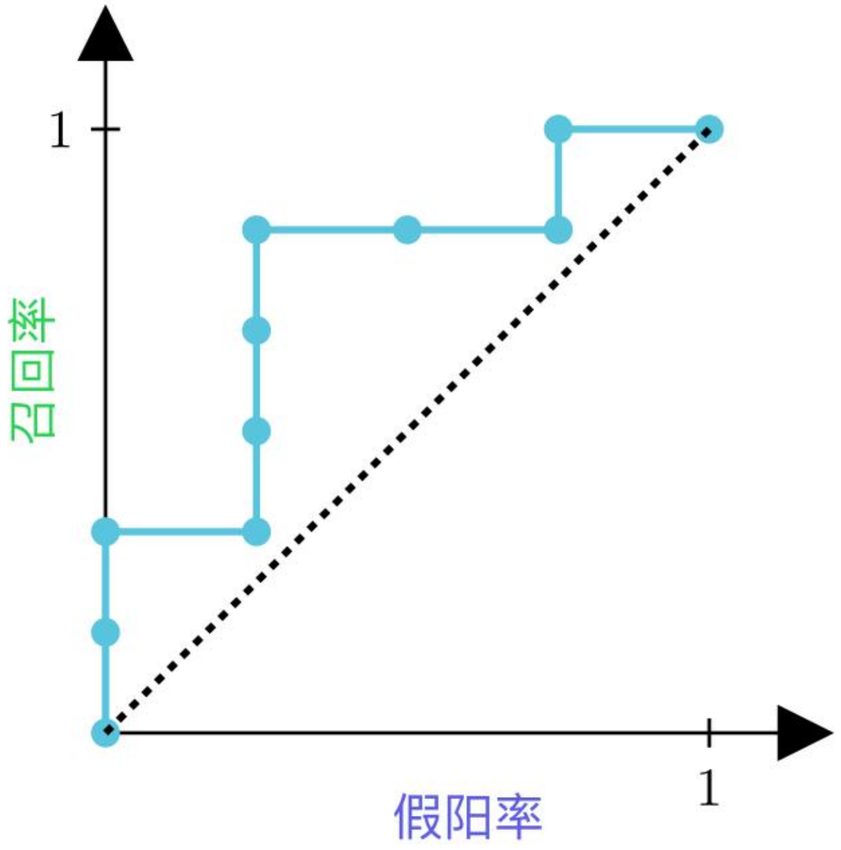


样本编号	真实标签	预测标签	预测得分
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

问题 2：在什么情况下，我们向上绘制下一个点?

答案：基于排序后的结果，如果下一个真实标签是正例，我们向上绘制下一个点

假阳率	召回率
0	0
0	1/6
0	2/6
1/4	2/6
1/4	3/6
1/4	4/6
1/4	5/6
2/4	5/6
3/4	5/6
3/4	1
1	1

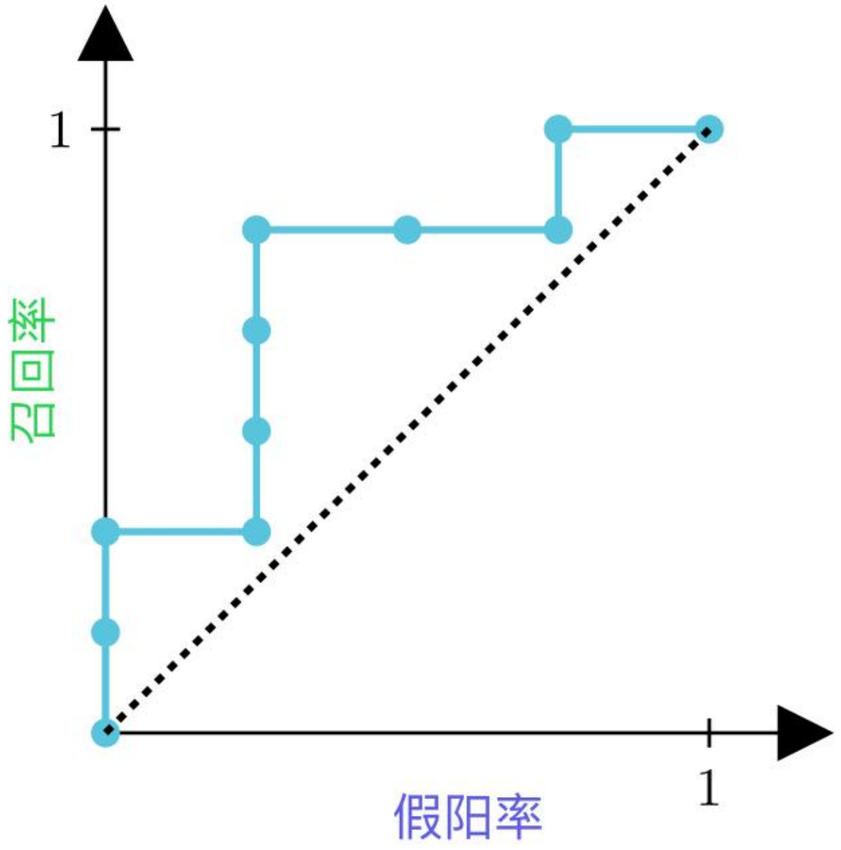


样本编号	真实标签	预测标签	预测得分
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

问题 3：在什么情况下，我们向右绘制下一个点?

答案：基于排序后的结果，如果下一个真实标签是负例，我们向右绘制下一个点

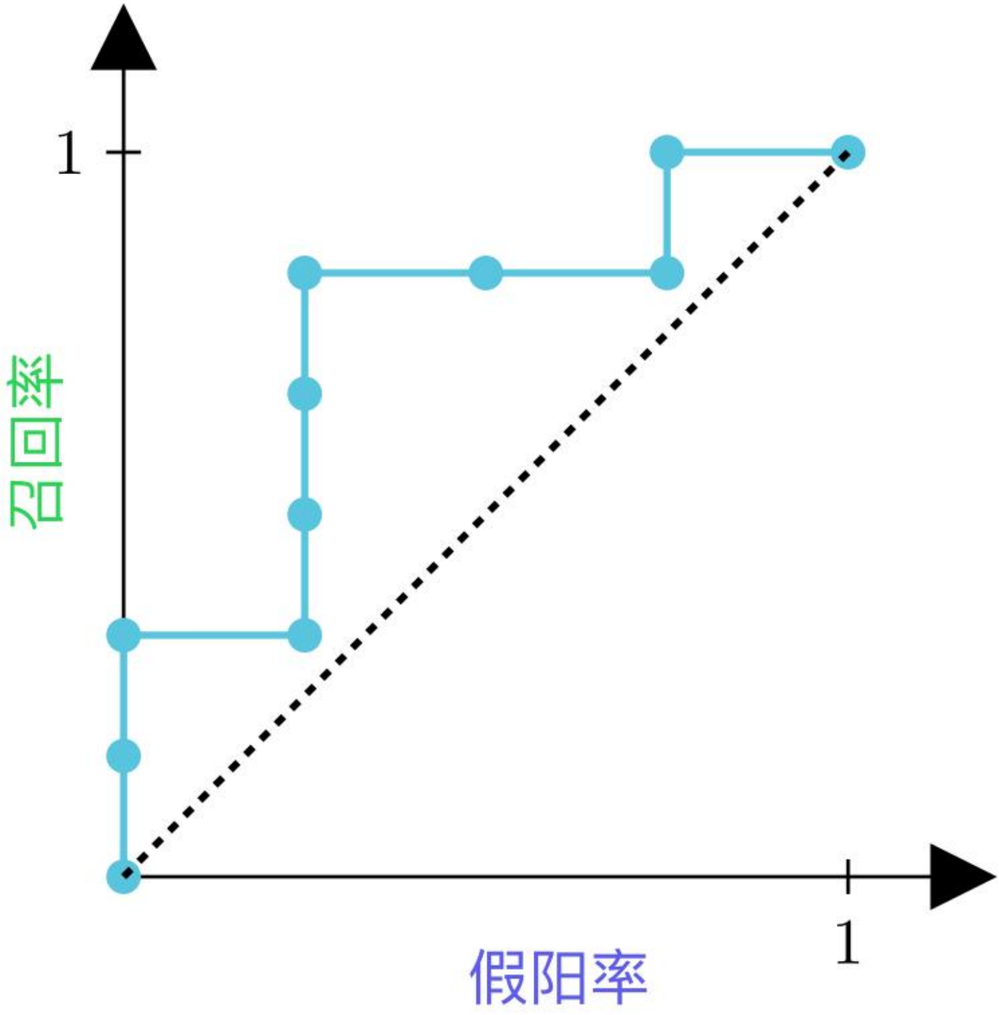
假阳率	召回率
0	0
0	1/6
0	2/6
1/4	2/6
1/4	3/6
1/4	4/6
1/4	5/6
2/4	5/6
3/4	5/6
3/4	1
1	1



样本编号	真实标签	预测标签	预测得分
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

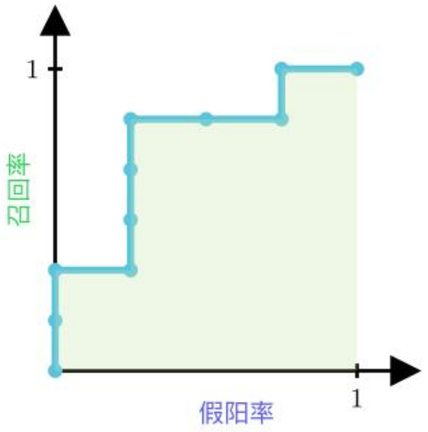
问题 4：完美模型的 ROC 曲线是什么样子的?

答案：一个完美的模型，能够将所有的真实正例，排在真实负例之前。因此，其 ROC 曲线的绘制从原点出发，首先垂直移动到 $(0,1)$ 点，然后水平移动至 $(1,1)$ 点。



样本编号	真实标签	预测标签	预测得分
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

AUC



1. AUC 是 “area under the ROC curve” 的缩写
   - AUC 计算的是，ROC 曲线下方的面积
2. 一个好的二分类模型，往往对应一个比较大的 AUC 值
   - 如果一个二分类模型，能够利用 “得分”，将真实正例样本排在真实负例样本之前，其 AUC 值往往较高
   - 完美模型的 AUC 等于 1，其能够将所有真实正例排在面
   - AUC $= 0.5$ 表示模型的整体排序能力与随机排序相当

备注

1. AUC 提供单一的、可排序的量化指标
2. 平衡不同阈值下的整体性能
3. 对类别不平衡的情况，相对稳健
4. 取值范围为 $[0, 1]$，但通常是 $[0.5, 1]$
5. 如果一个模型的 AUC 很低，其可能没有那么差，我们也许可以颠倒其预测类别，来可得到一个更好的模型
6. 当正例稀少而我们非常在意精确率时，我们可以使用 Precision-Recall 曲线，及其下方面积进行模型比较

课程总结

1．不同指标的作用是不一样的

- 准确率：总体评价
- 精确率：聚焦预测正类的可靠性
- 召回率：聚焦模型漏掉的正类
- 假阳率：聚焦模型误伤负类的比例
- F1 分数：平衡精确率和召回率
- ROC 和 AUC：模型在不同阈值下的综合表现