机器学习方法在车险定价中的应用

程建华

2021 年 5-6 月

目录

目录

1.1 保险

保险作为现代金融体系的组成部分，是人们管理纯粹风险最为有效的方法，在促进经济发展、安定人民生活和优化社会管理等方面发挥着重要作用.

- 社会保险；
- 政策保险；
- 商业保险.

1.2 精算

精算学是一门植根于实践活动的学科，它以数学、统计学的相关理论和方法为基础，并结合经济学的一般原理，以解决保险、金融、投资与财务等领域中的定量化问题为研究目的.

- 数学;
- 统计学;
- 经济学;
- 数据科学.

1.3 保费定价

保费定价也称为费率厘定，是保险公司运营的核心环节之一. 科学合理的费率厘定方法是保险公司稳健经营的前提，也是保险产品具有竞争力的关键因素.

保费定价过程就是根据保单（被保险人）的损失经验和其他相关信息建立模型，并对其未来的保险成本（赔款、代理人佣金、一般管理费用、理赔费用，以及支持该业务所需的资本金成本）进行预测的过程.

保险公司实际使用的费率还会受到市场供求关系和公司自身发展战略的影响.

1.4 汽车保险

汽车保险也称机动车辆保险，是以机动车辆本身及其第三者责任等为保险标的的一种运输工具保险，在非寿险精算领域内占有重要地位. 汽车保险是伴随着 19 世纪后期汽车在欧洲的普及而出现的，目前，已成为各国财产保险中最重要的业务险种. 在发达国家，汽车保险的保费收入一般要占财产险总保费的 50% 左右，在中国大陆地区实施交通事故强制保险制度后，汽车保险一度占到总财产险保费的 70% 以上.

1.5 车险定价

车险的精算定价是与车险同时诞生的，至今已经有一百多年的历史了. 由于车险已成为财产保险中名副其实的“龙头险种”，其经营效益的优劣直接影响到各财险公司财务盈亏，因此，各家保险公司对车险定价极其重视，车险定价的精算模型也成为非寿险精算领域的重要研究内容. 汽车保险的精算定价是保险公司承保风险之前最主要和最重要的风险管理工具, 精算师和学者对此进行了广泛而深入研究.

1.6 车险风险的度量

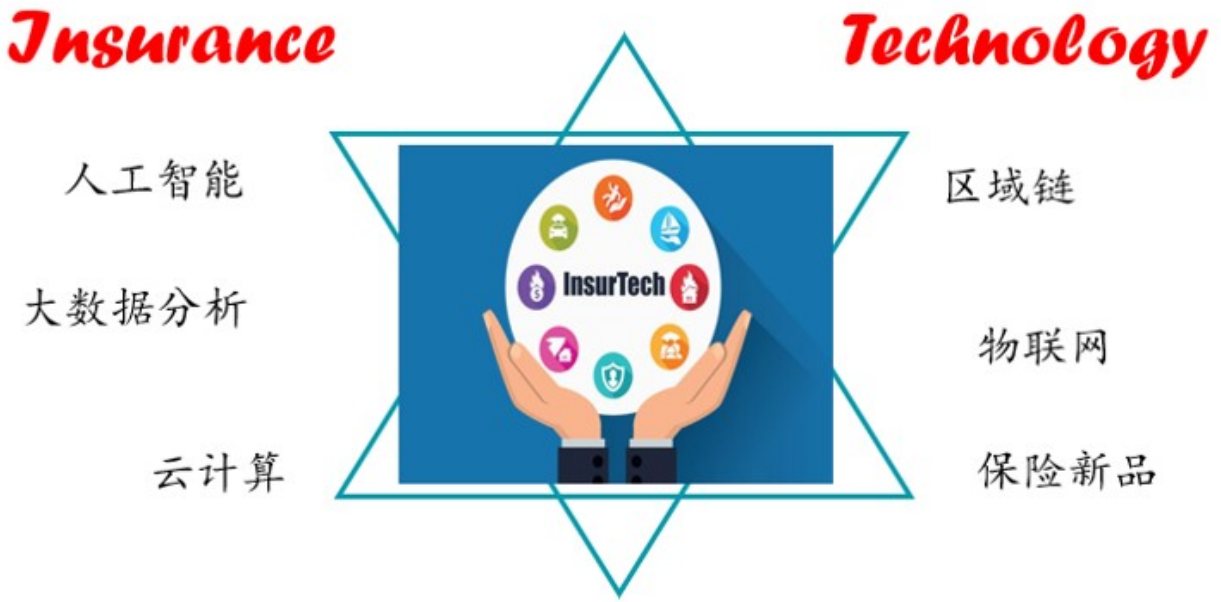
风险是未来的不确定性：

- 索赔概率：索赔发生的可能性；
- 索赔频率：给定时间区间内（平均每个风险单位）发生索赔的次数；
- 索赔强度：给定时间区间内平均每次索赔的额度.

车险的纯保费计算公式为

$$P = E(S) = E(X_1 + X_2 + \cdots + X_N) = E(N)E(X).$$

1.7 InsurTech 时代的保险与精算科学



目录

2.1 索赔概率研究

在监督学习中，存在大量关于“是与否”的二分类问题，在各行业有着广泛的应用. 比如，人脸识别 (是否为人脸、是否为某人的人脸)、自动驾驶 (是否应刹车)、银行贷款 (是否批准贷款申请)、邮件过滤 (是否为垃圾邮件) 等.

对于车险索赔风险，响应变量 Y 只有两种可能取值，即

$$Y = 1(\text{发生索赔}) \quad \text{或者} \, Y = 0(\text{未发生索赔}).$$

记被保险人的特征向量为 $\boldsymbol{X} = (X_1, X_2, \cdots, X_q)^{\mathrm{T}}$，考虑 X 和 Y 之间的关系.

2.2 线性回归模型

最简单的建模方法为“线性回归模型”，即

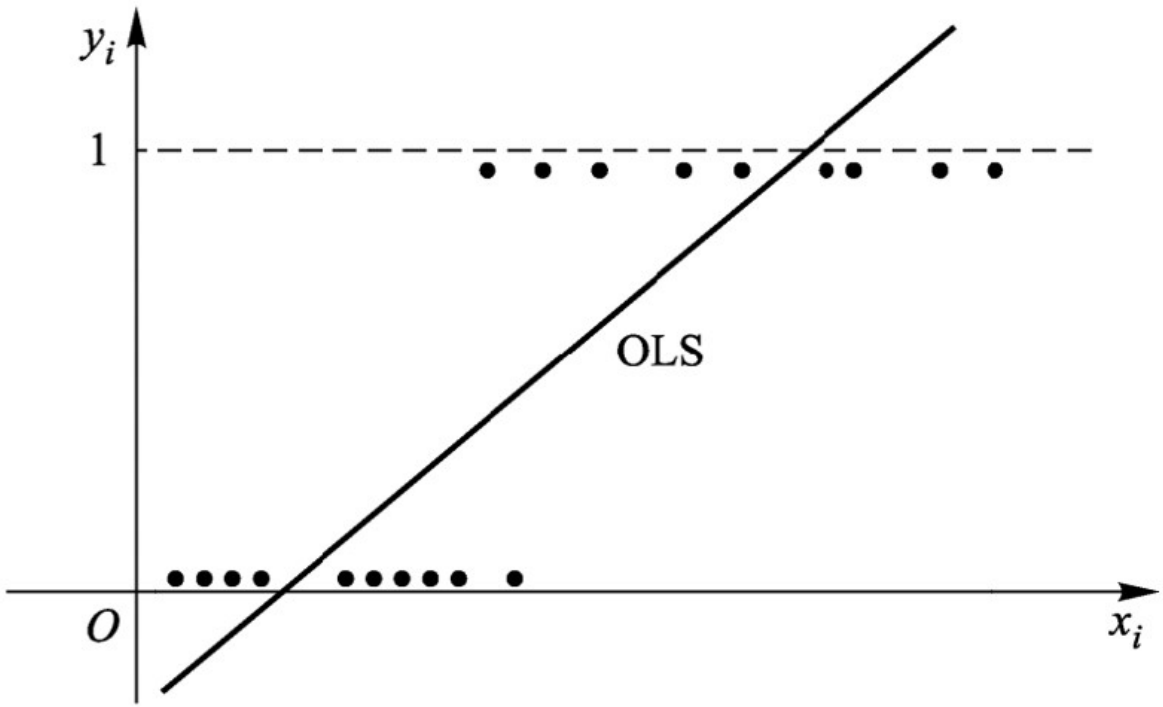
$$Y_i = \beta X_{i1} + \beta_2 X_{i2} + \cdots + \beta_q X_{iq} + \epsilon_i = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

其中

$$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{iq})^{\mathrm{T}}, \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_q)^{\mathrm{T}}.$$

线性回归模型的优点在于，计算方便 (就是 OLS 估计)，且容易得到边际效应 (即回归系数). 但线性回归模型并不适合作预测. 明知 Y 的取值非 0 即 1，但根据线性回归模型所作的预测值却可能出现 $\hat{Y} > 1$、$\hat{Y} < 0$ 或 $0 < \hat{Y} < 1$ 的不现实情形.

2.2 线性回归模型



2.3 逻辑回归模型 (Logistic Regression，简记 Logit)

为使 Y 的预测值总是介于 $[0,1]$ 之间，在给定 $\boldsymbol{X}$ 的情况下，考虑 Y 的两点分布概率：

$$P(Y=1|\boldsymbol{X})=F(\boldsymbol{X},\boldsymbol{\beta}),\quad P(Y=0|\boldsymbol{X})=1-F(\boldsymbol{X},\boldsymbol{\beta}),$$

其中 $F(\boldsymbol{X},\boldsymbol{\beta})$ 称为连接函数，因为它将特征向量 $\boldsymbol{X}$ 与响应变量 Y 连接起来.

连接函数的选择具有一定的灵活性. 通过选择合适的连接函数 $F(\boldsymbol{X},\boldsymbol{\beta})$(比如，某随机变量的累积分布函数)，可以保证 $0\leq Y\leq 1$.

如果选取 $F(\boldsymbol{X},\boldsymbol{\beta})$ 为逻辑分布的分布函数，即

$$P(Y=1|\boldsymbol{X})=F(\boldsymbol{X},\boldsymbol{\beta})=\frac{\exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}}=\frac{1}{1+\exp\{-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}},$$

此模型称为“逻辑回归” (Logistic Regression) 或“逻辑斯蒂回归”，简记 Logit.

2.4 逻辑回归模型的参数估计

给定样本 $\{\boldsymbol{X}_i, Y_i\}$，且假设样本中的个体相互独立，那么可以使用“最大似然估计”(Maximum Likelihood Estimation，简记 MLE) 对参数进行统计推断，即

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \max_{\boldsymbol{\beta}} \ln L(\boldsymbol{X}, \boldsymbol{Y}) \\ &= \max_{\boldsymbol{\beta}} \sum_{i=1}^{n} Y_i \ln\left(\frac{\exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}}\right) + \sum_{i=1}^{n}(1-Y_i)\ln\left(1-\frac{\exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}}\right).\end{aligned}$$

由于目标函数为非线性函数，故不存在解析解. 一般使用数值计算的方法，比如牛顿法，求解此非线性最大化问题.

注：参数的最大似然估计具有相合性和渐近正态性.

2.5 逻辑回归模型的解释

对于逻辑回归模型，因为

$$p = P(Y = 1|\boldsymbol{X}) = \frac{\exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}},$$

那么

$$1 - p = P(Y = 0|\boldsymbol{X}) = \frac{1}{1 + \exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}},$$

则

$$\frac{p}{1-p} = \exp\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\}, \quad \ln\left(\frac{p}{1-p}\right) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q,$$

其中 $\frac{p}{1-p}$ 称为几率 (odds)，$\ln\left(\frac{p}{1-p}\right)$ 称为对数几率 (log-odds).

2.5 逻辑回归模型的解释

若 X_k 为连续型变量，则

$$\beta_k = \frac{\partial \ln\left(\frac{p}{1-p}\right)}{\partial X_k} \approx \frac{\triangle\left(\frac{p}{1-p}\Big/\frac{p}{1-p}\right)}{\triangle X_k},$$

可将 β_k 解释为半弹性 (semi-elasticity)，即当 X_k 增加 1 单位，可引起几率 $\frac{p}{1-p}$ 变化的百分比：即当 $\triangle X_k = 1$ 时，

$$\frac{\triangle odds}{odds} = \frac{\triangle\left(\frac{p}{1-p}\right)}{\frac{p}{1-p}} \approx \beta_k.$$

2.5 逻辑回归模型的解释

若 X_k 为离散型变量 (比如，性别、子女数)，假设 X_k 增加 1 单位，从 X_k 变为 X_k+1，记概率 p 的新值为 p^*，则可根据新几率 $\frac{p^*}{1-p^*}$ 与原几率 $\frac{p}{1-p}$ 的比率定义几率比 (odds-ratio)：

$$OR = \frac{\frac{p^*}{1-p^*}}{\frac{p}{1-p}} = \frac{\exp\{\beta_1 X_1 + \cdots + \beta_k(X_k+1) + \cdots + \beta_q X_q\}}{\exp\{\beta_1 X_1 + \cdots + \beta_k X_k + \cdots + \beta_q X_q\}} = \exp\{\beta_k\}.$$

2.6 逻辑回归模型的拟合优度

对于使用 MLE 进行估计的非线性模型，可使用准 R^2(Pseudo R^2) 或伪 R^2 度量模型的拟合优度，其定义为：

$$Pseudo - R^2 = \frac{\ln L_0 - \ln L_1}{\ln L_0},$$

其中，$\ln L_1$ 为原模型的对数似然函数之最大值，而 $\ln L_0$ 为以常数项为唯一变量的对数似然函数之最大值.

在统计学中，还常使用偏离度 (deviance) 的概念. 偏离度也常常称为“残差偏离度”(residual deviance)，其定义为：

$$Residual\ deviance = -2 \ln L_1.$$

2.7 逻辑回归模型的预测

得到 Logit 模型的估计系数后，即可预测 $Y_i = 1$ 的条件概率：

$$\hat{p}_i = \hat{P}(Y_i = 1|\boldsymbol{X}_i) = \frac{\exp\{\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\}}{1 + \exp\{\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\}},$$

- 若 $\hat{p}_i > 0.5$，则预测 $\hat{Y}_i = 1$；
- 若 $\hat{p}_i < 0.5$，则预测 $\hat{Y}_i = 0$；
- 若 $\hat{p}_i = 0.5$，则预测 $\hat{Y}_i = 1$ 或 $Y_i = 0$.

2.7 逻辑回归模型的预测

对于 Logit 模型，也可使用对数几率来预测其响应变量的类别，即：

$$\frac{\hat{p}_i}{1-\hat{p}_i} = \boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}},$$

- 若 $\frac{\hat{p}_i}{1-\hat{p}_i} > 0$，则预测 $\hat{Y}_i = 1$；
- 若 $\frac{\hat{p}_i}{1-\hat{p}_i} < 0$，则预测 $\hat{Y}_i = 0$；
- 若 $\frac{\hat{p}_i}{1-\hat{p}_i} = 0$，则预测 $\hat{Y}_i = 1$ 或 $Y_i = 0$.

2.8 逻辑回归模型的评估

对于监督学习问题，一般用预测效果来评估其模型的性能. 具体到分类问题，一个常用指标为准确率 (accuracy)，也称为“正确预测的百分比”：

$$Accuracy = \frac{\sum_{i=1}^{n} I(\hat{Y}_i = Y_i)}{n},$$

其中，$I(\cdot)$ 为“示性函数”(indicator function)，如果其括号内的表达式为真，则取值为 1；反之，则取值为 0.

有时我们更专注于错误预测的百分比，即错误率 (error rate) 或错分率：

$$Errorrate = \frac{\sum_{i=1}^{n} I(\hat{Y}_i \neq Y_i)}{n}.$$

注：如果所考虑样本为训练集，则为“训练误差”(training error). 如果所考虑样本为测试集，则为“测试误差”(test error).

2.8 逻辑回归模型的评估

准确率或错分率并不适用于“类别不平衡”(class imbalance) 的数据. 例如，假设某种罕见病的发病率仅为百分之一. 此时，样本中的两个类别高度不平衡. 即使不用任何机器学习的算法，只要一直预测不发病，也能达到 99% 的准确率 (或 1% 的错分率).

我们更希望算法能够准确地预测那些发病的个体，即所谓“正例”(positive cases，简称 positives). 为此，根据模型预测的正例 (也称“阳性”) 与反例 (也称“阴性”)，以及实际观测的正例与反例，可将样本数据分为以下四类，并用一个矩阵来表示，即所谓混淆矩阵 (confusion matrix).

2.8 逻辑回归模型的评估

混淆矩阵

		实际观测值	
		正例	反例
预测值	正例	真阳性（TP） $\hat{Y}=1$，$Y=1$	假阳性（FP） $\hat{Y}=1$，$Y=0$
	反例	假阴性（FN） $\hat{Y}=0$，$Y=1$	真阴性（TN） $\hat{Y}=0$，$Y=0$

2.8 逻辑回归模型的评估

根据混淆矩阵的信息，可设计更为精细的模型评估指标：

(1) 从纵向角度考察混淆矩阵的第 1 列，在实际为正例的子样本中，定义其预测正确的比例为灵敏度 (sensitivity) 或真阳率 (true positive rate)：

$$Sensitivity = \frac{TP}{TP + FN}.$$

灵敏度也称为“查准率”(precision)，它反映了在实际为正例的子样本中，正确预测的比例；尤其适用于上文关于罕见病的案例.

(2) 考虑混淆矩阵的第 2 列，在实际为反例的子样本中，定义其预测正确的比例为特异度 (specificity)，也称为真阴率 (true negative rate)：

$$Specificity = \frac{TN}{FP + TN}.$$

2.8 逻辑回归模型的评估

(3) 进一步，“1-特异度”则为在实际为反例的子样本中，错误预测的比例，也称为假阳率 (false positive rate)：

$$False\ positive\ rate = 1 - Specificity = \frac{FP}{TP + FN}.$$

(4) 也可从横向角度考察混淆矩阵. 比如，考虑混淆矩阵的第 1 行，在预测为正例的子样本中，定义其预测正确的比例为查全率或召回率 (recall)：

$$Recall = \frac{TP}{TP + FP}.$$

2.8 逻辑回归模型的评估

(3) 进一步，“1-特异度”则为在实际为反例的子样本中，错误预测的比例，也称为假阳率 (false positive rate)：

$$False\ positive\ rate = 1 - Specificity = \frac{FP}{TP + FN}.$$

(4) 也可从横向角度考察混淆矩阵. 比如，考虑混淆矩阵的第 1 行，在预测为正例的子样本中，定义其预测正确的比例为查全率或召回率 (recall)：

$$Recall = \frac{TP}{TP + FP}.$$

2.9 ROC 与 AUC

迄今为止，我们默认用于分类的“门槛值”(threshold) 为 $\hat{p} = 0.5$. 事实上，从决策理论的角度，这未必是最佳选择. 从混淆矩阵可知，在作预测时，可能犯两类不同的错误，即“假阳性”与“假阴性”. 在具体的业务中，这两类错误的成本可能差别很大. 例如：银行在审批贷款申请时 (断供 = 正例)，“假阳性”将正常客户视为劣质客户而拒绝贷款，其成本只是少赚些利润；而“假阴性”将劣质客户视为正常客户而放贷，则会面临因断供而损失本金的巨大成本

作预测的两类错误，其成本可能并不对称. 此时，应根据具体的业务需要，考虑使用合适的门槛值 $\hat{p} = c$ 进行分类. 比如，为了降低错误放贷的损失，银行可将分类为劣质客户的门槛值降低到 $\hat{p}$. 这意味着，如有 20% 或以上的概率客户会断供，则判断为断供，并拒绝贷款.

2.9 ROC 与 AUC

使用更低的门槛值，将预测更多的正例，而预测更少的反例。此时，在实际为正例的子样本中，预测准确率将上升，即灵敏度上升. 但在实际为反例的子样本中，预测准确率将下降，即特异度下降，故“1-特异度”上升.

灵敏度与“1-特异度”均为门槛值 $\hat{p} = c$ 的函数，将其分别记为 Sensitivity(c) 和 1-Specificity(c). 把 1-Specificity(c) 放于坐标横轴，把 Sensitivity(c) 放于纵轴，然后让门槛值 $\hat{p} = c$ 的取值从 0 连续地变为 1，则可得到一条曲线，即所谓接收器工作特征曲线 (Receiver Operating Characteristic Curve，简记 ROC 曲线).

2.9 ROC 与 AUC

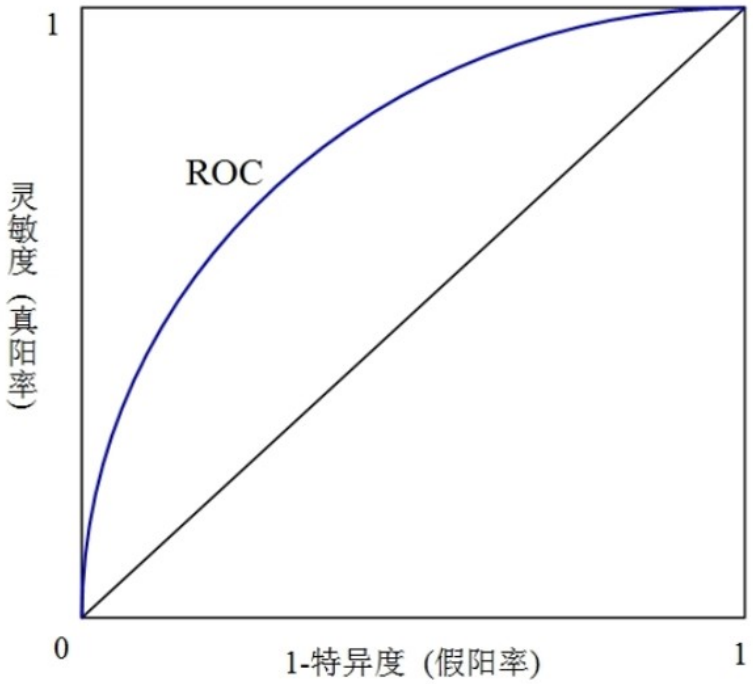


图 5.5　ROC 曲线的示意图

2.9 ROC 与 AUC

当 $0 \leq c \leq 1$ 时，则可得到整条 ROC 曲线. 由于纵轴为实际正例中的准确率 (灵敏度)，而横轴为实际反例中的错误率 (1-特异度)，故我们希望模型的 ROC 曲线越靠近左上角越好. 为衡量 ROC 曲线的优良程度，可使用 ROC 曲线下面积 (Area Under the Curve，简记 AUC) 来度量.

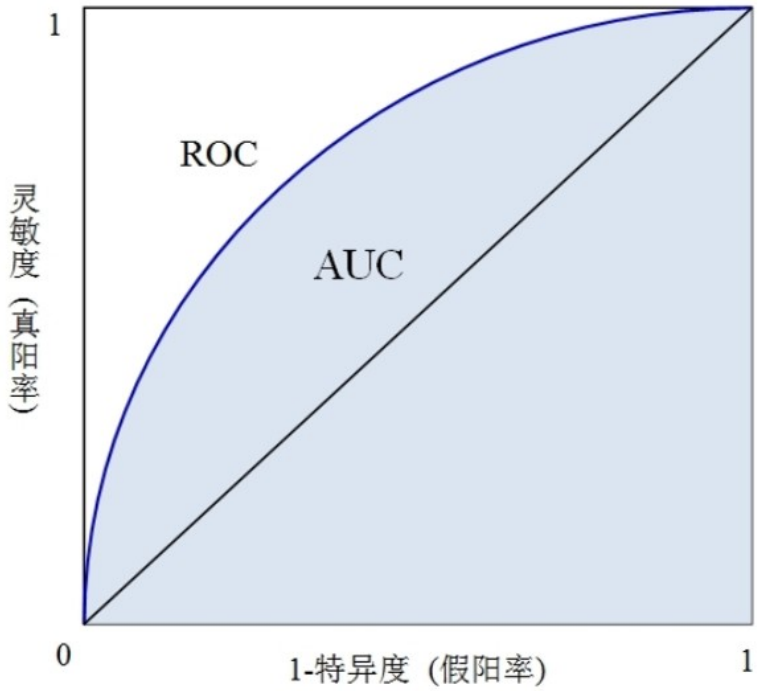


图 5.6　AUC 的示意图

2.10 ROC 与 AUC

如果 AUC 为 1，则意味着模型对于所有正例与反例的预测都是正确的，这一般是无法达到的理想状态.

如果 ROC 曲线与 45 度线 (从原点到 $(1,1)$ 的对角线) 重合，则意味着该模型的预测结果无异于随机猜测. 比如，样本中正例与负例各占一半，而通过从 $[0,1]$ 区间的均匀分布随机抽样来预测概率. 此时，AUC 为 0.5.

AUC 小于 0.5 的情形十分罕见，这意味着模型的预测结果还不如随机猜测.

对于二分类问题，在比较不同模型的预测效果时，常使用 AUC. 由于 AUC 为衡量预测效果的综合性指标，可使用此单一指标比较不同的算法.

2.11 科恩 Kappa(Cohen's Kappa)

假设预测值与实际值分别来自两个不同的“评分者”(raters). 我们希望评估这两个评分者之评分结果 (rating) 的一致性 (agreement). 与仅适用于二分类的 AUC 指标相比，科恩 Kappa 可用于多分类问题.

kappa 的取值	kappa 的解释
$\text{kappa} \leq 0.2$	一致性很差 (poor agreement)
$0.2 < \text{kappa} \leq 0.4$	一致性较差 (fair agreement)
$0.4 < \text{kappa} \leq 0.6$	一致性中等 (moderate agreement)
$0.6 < \text{kappa} \leq 0.8$	一致性较好 (good agreement)
$0.8 < \text{kappa} \leq 1$	一致性很好 (great agreement)