The ‘P’

性能度量与评估

The Performance measure and evaluation of machine learning approaches

陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

Experiences（经验）

- Experience = Data we have for training the machine learning model.

- 对于特定机器学习任务，已存在的可利用数据即是解决该机器学习任务的经验。

- 数据为王：大数据=丰富经验=训练更好的机器学习模型

数据划分

- 训练集（Training Set）
  - 用来训练模型或确定模型参数。
- 测试集（Testing Set）
  - 测试已经训练好的模型的推广能力。
- 验证集（Validation set）可选
  - 用来做模型选择（model selection），即做模型的最终优化及确定的。

数据集划分策略

- 利用测试集或验证集评估学习器的泛化误差，进而进行模型优化与选择，避免过拟合。

- 常见划分策略：
  - 留出法
  - 交叉验证法
  - 自助法

- 数据集划分各子集之间不能有重合。

留出法（Hold-out）

- 留出法（hold-out）直接将数据集D划分为两个互斥的集合，分别为训练集S与测试集T，即 $D = S \cup T, S \cap T = \varnothing$ 。
- 训练/测试集划分尽量保持数据一致性
- 采用合理的采样，合理控制训练集与测试集比例。
- 多次使用留出法，重复进行实验评估，求均值，减少数据分布差异造成的偏差。

交叉验证法（Cross Validation）

- n-折交叉验证法（n-fold Cross Validation）:把数据集等分为n份相互不重叠的子集，每次以其中1份子集作为测试集，其余n-1份子集作为训练集，充分n次，直至所有子集都作为测试集进行过一次实验评估，最后返回n次实验评估的平均结果。常见n取值2、5、10、20。

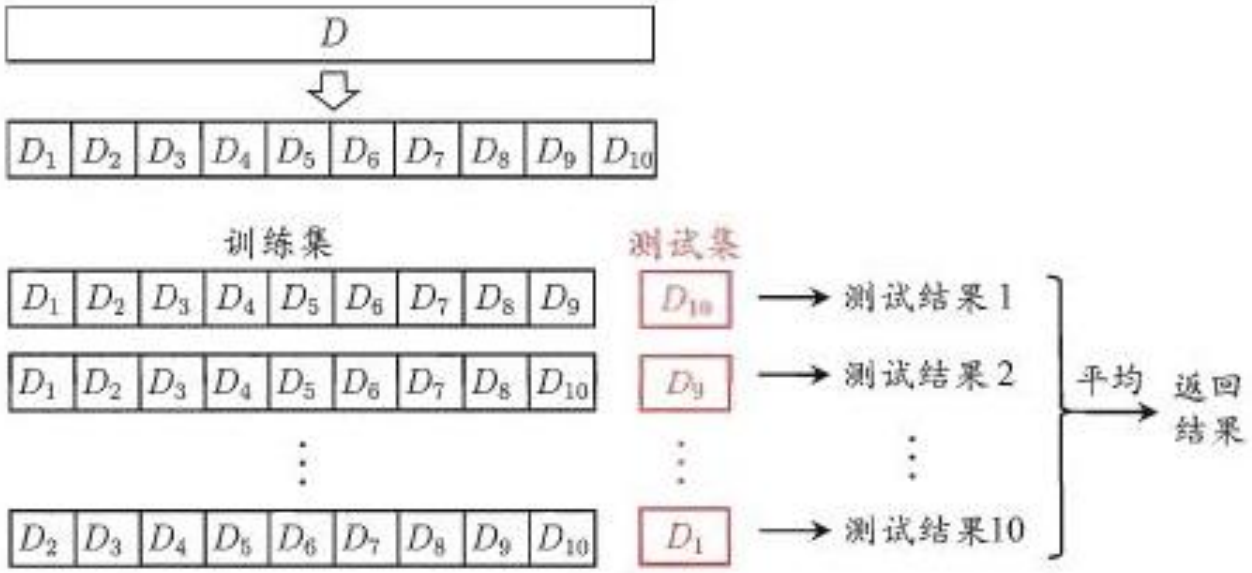


图 2.2 10 折交叉验证示意图

交叉验证法（ Cross Validation ）

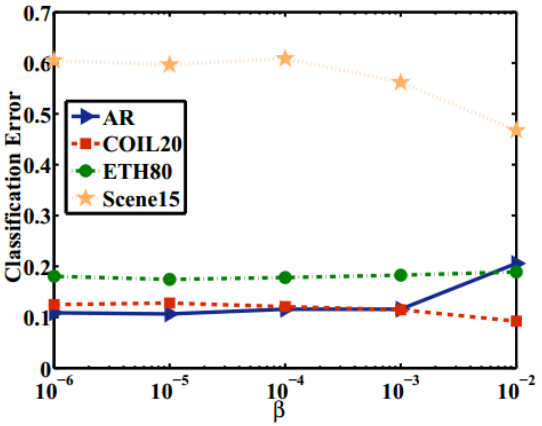
- 交叉验证是最常见数据集划分方法。

- 留一法（Leave-One-Out，简称LOO）：特殊的交叉验证法，每个被划分的子集只有一个样本。
  - 优点：
    - 训练集比例高，训练出来模型与用所有数据进行训练的模型相似度高。
  - 缺点：
    - 评估开销大
    - 测试集比例太低，模型调参不便。

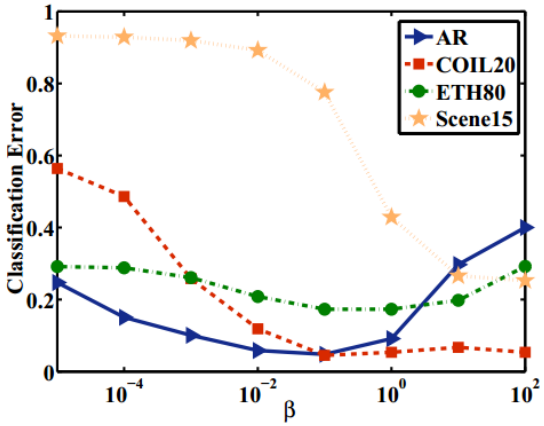
自助法 （Bootstrapping）

- 自助法（Bootstrapping）:假设一个由m个样本组成数据集D，对其进行m次随机采样构造一个由m个样本组成新数据集D’, 由于m次随机采样可能会对D中部分样本重复采样，所以D’中有部分样本是完全相同，而D中有部分样本是没有被采样到数据集D’中。因此我们可以把D中这部分没有被采样到样本D\D’构造测试集，而D’作为训练集。
- 这种没被采样到样本在数据集D中比例一般占25%~36.8%之间。
- 自助法通常用于数据集较小或难以有效划分训练/测试集情况。

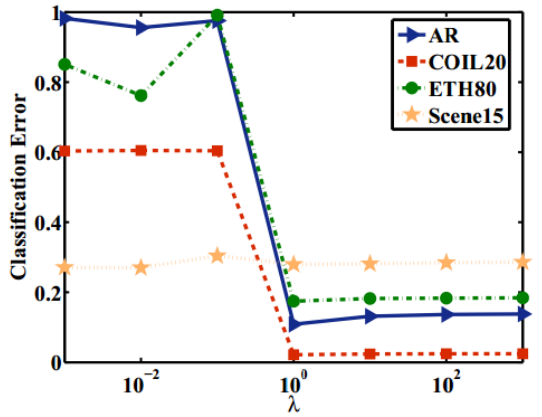
调参与最终模型



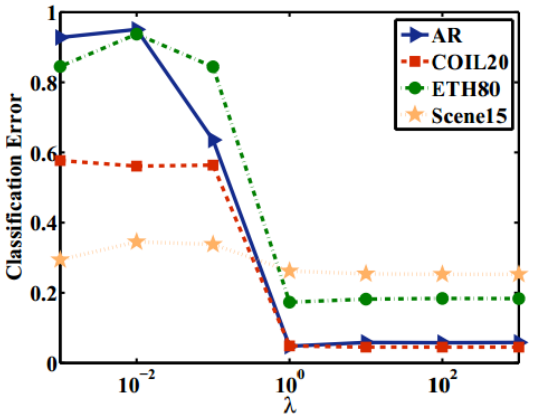
(a) The impact of β to L_1HT



(b) The impact of β to L_2HT

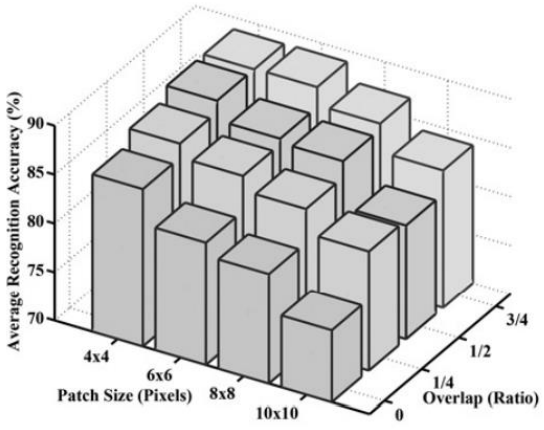


(c) The impact of λ to L_1HT



(d) The impact of λ to L_2HT

- 参数测试选择步长，均值或指数。如[0.1, 0.2, 0.3] 或 [0.1, 1, 10]
- 选取最优参数组合，网格法，假设两个参数的候选值均为5个，则其最优参数候选集合为5x5=25.



Performance Measure

- 常见性能度量
  - 均方误差
  - 误差与精度
  - 查准率、查全率、F1
  - ROC与AUC

均方误差（mean squared error）

多用度量回归任务的性能。

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2$$

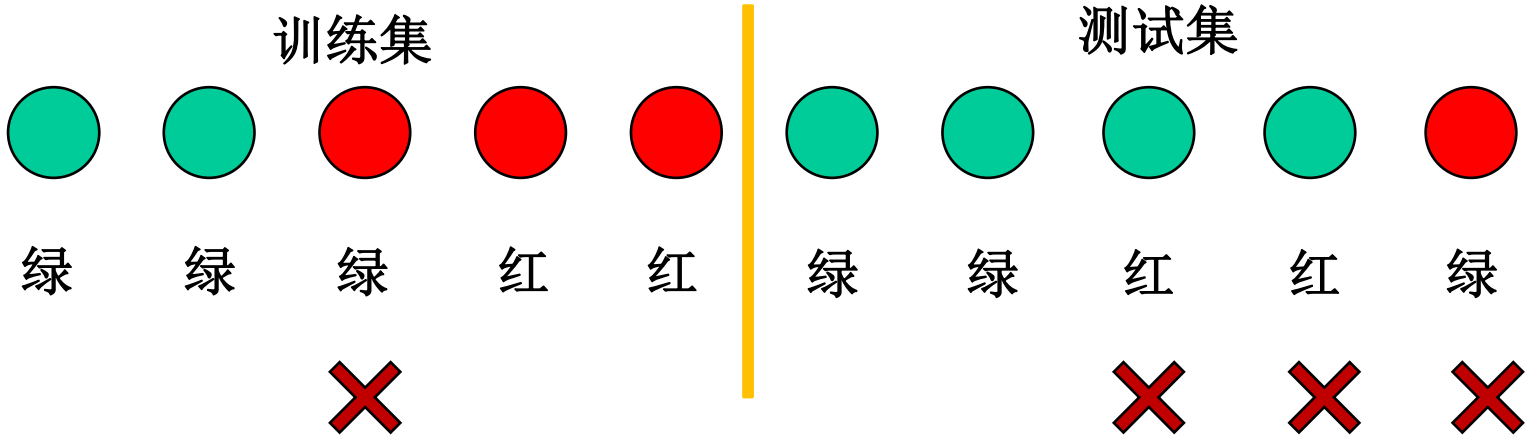
$D=\{(x_1,y_1),(x_2,y_2),\dots,(x_m,y_m)\}$　　$f(\cdot)$: 学习器

$$E(f;\mathcal{D})=\int_{x\sim\mathcal{D}}(f(x)-y)^2\,\rho(x)dx$$

误差与精度

- **误差（error）**：学习器（Learner）的实际预测输出与样本的真实输出之间的差异。
- **错误率（error rate）**: 被错误分类的样本在总样本中的比例。
- **精度（accuracy）**:被正确分类的样本在总样本中的比例，即1 – error rate。
- **训练误差（training error）**:学习器在训练集上的误差，也称作经验误差（empirical error）。
- **测试误差（testing error）**：学习器在测试集上的误差，用来近似泛化误差。
- **泛化误差（generalization error):** 在新样本的误差，实际误差！

Example



$$\text{Error rate} = \frac{4}{10} = \frac{2}{5} = 0.4$$

$$\text{empirical } error = \frac{1}{5} = 0.2 \qquad testing\ error = \frac{3}{5} = 0.6$$

What about generalization error?

Don’t know, but that should be not good!

查准率、查全率、F1

- 对各类别重视程度不一样情况（选瓜，疾病筛查）

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

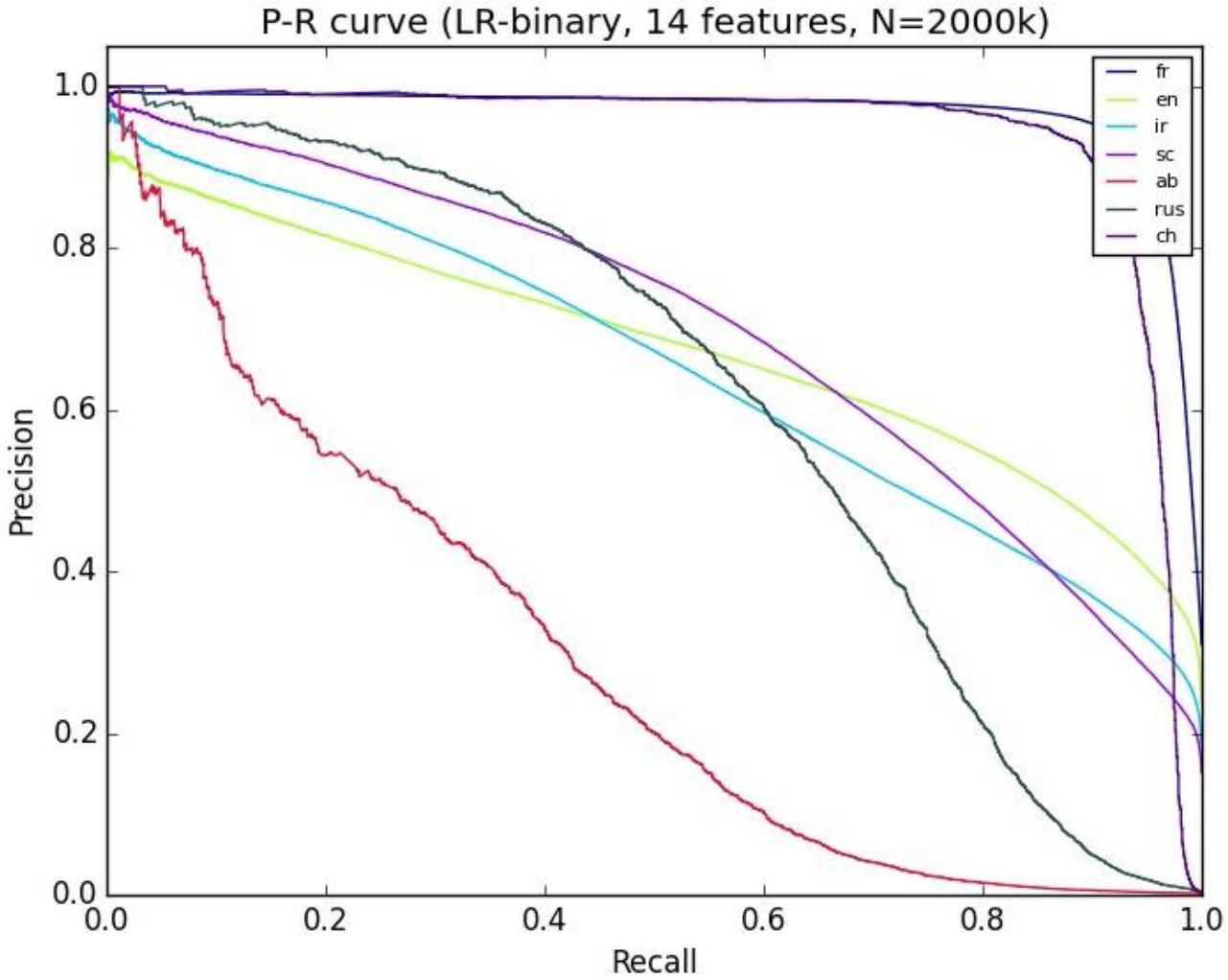
  - 二分类中真实类别与预测类别的四种情况。
    - 真正例（True Positive）TP
    - 真反例（True Negative）TN
    - 假正例（False Positive）FP
    - 假反例（False Negative）FN
    - (TP+FN)+(TN+FP)=P+N=样本总数

查准率、查全率、F1

- 查准率、查全率与F1
  - 查准率(Precision):
    - 被正确分类的正例样本在被学习器分类成为正例样本中所占的比例。
    - 查准率 $P = \frac{TP}{TP+FP}$
  - 查全率(Recall):
    - 被正确分类的正例样本在正例样本中的比例。
    - 查全率$R = \frac{TP}{TP+FN}$
  - 查全率与查准率是一对相互矛盾的度量。

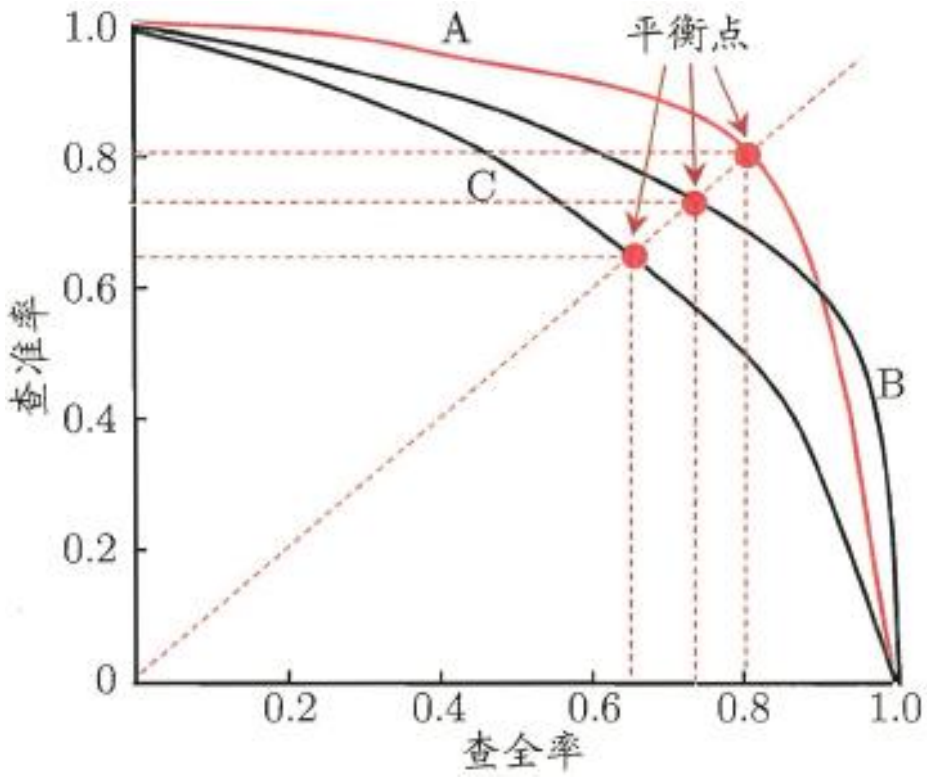
查准率、查全率、F1

- 查准率-查全率曲线：P-R曲线或P-R图



查准率、查全率、F1

- 基于查准率-查全率的学习器性能度量：
  - 平衡点（Break-Even Point, 简称BEP）
    - “查准率=查全率”时取值

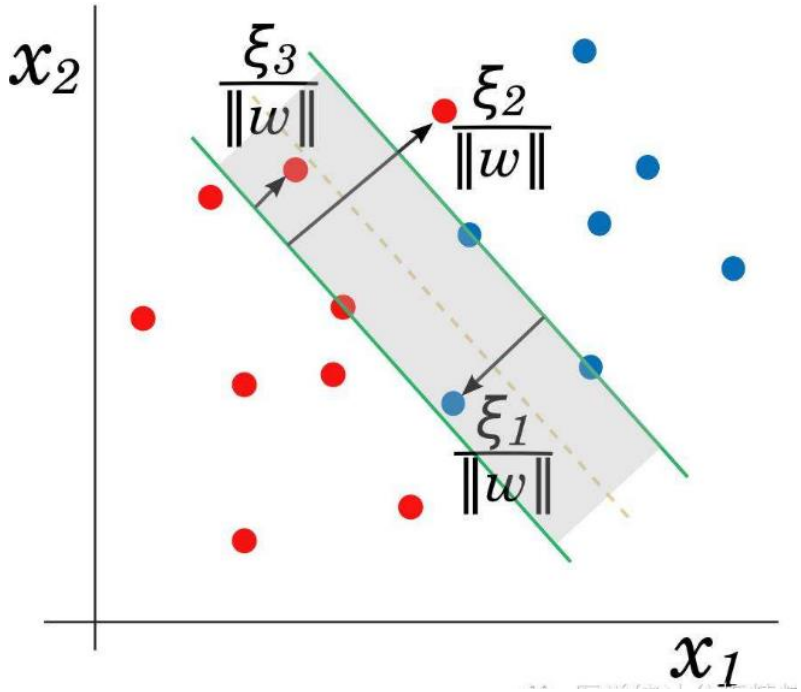


查准率、查全率、F1

- 基于查准率-查全率的学习器性能度量：
  - $F_1 - Score$：查准率与查全率的调和平均。
    - $F_1 = \frac{2 \times P \times R}{P+R} = \frac{2 \times TP}{样本总数+TP-TN}$
    - $\frac{1}{F_1} = \frac{1}{2} \times \left(\frac{1}{P} + \frac{1}{R}\right)$
  - $F_\beta - Score$: $F_1 - Score$的推广
  - 查全率、查准率、PR图、$F_1 - Score$多用于评估检索与检测任务的性能。

ROC与AUC

- 二分类问题的性能度量探讨：
  - 学习器并不是直接输出类别标签，输出是一个概率预测或置信值。
  - 分类阈值（Threshold）与截断点（Cut Point）



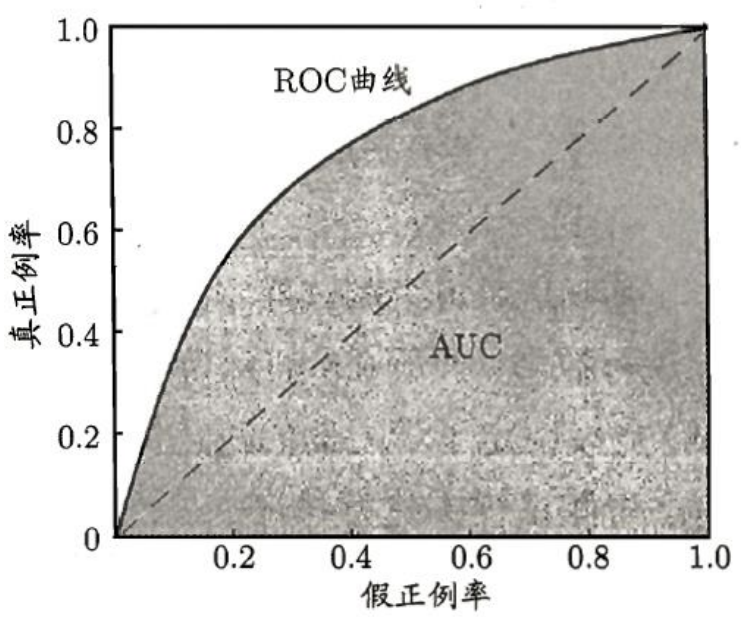
支持向量机的例子：分类器输出的是点到分类面的距离！参见教材第五章。

ROC与AUC

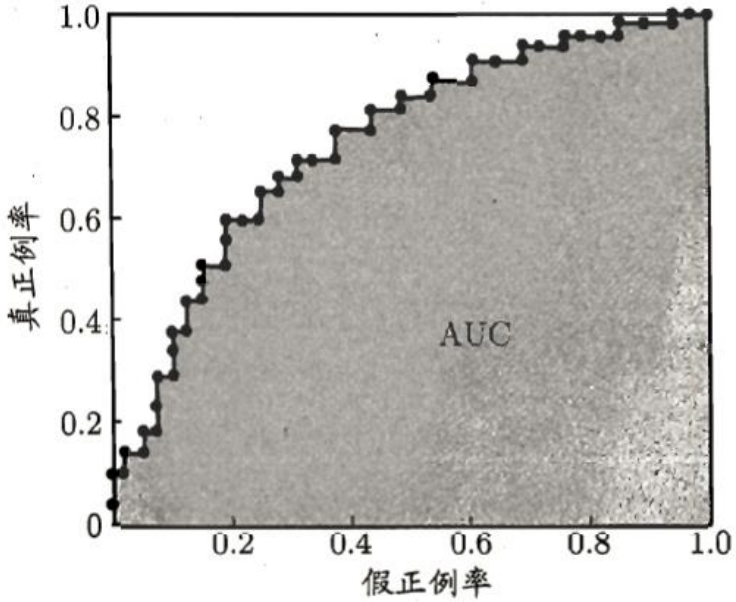
- 二分类问题的性能度量探讨：
  - 根据置信度对样本进行降序排序，一个泛化性能较强的学习器应该满足以下特性：正例拥有更高置信度，因此排序的排名比较靠前，负例拥有置信度较低，排名应靠后。
- 受试者工作特征（Receiver Operating Characteristic，简称ROC）曲线：
  - 真正例率（True Positive Rate，简称TPR）
    - $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$
  - 假正例率（False Positive Rate，简称FPR）
    - $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}}$

ROC与AUC

- AUC（Area Under ROC Curve）：
  - 即ROC曲线下面积。
  - $\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)\cdot(y_i + y_{i+1})$
  - ROC曲线坐标: $\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$



(a) ROC 曲线与 AUC



(b) 基于有限样例绘制的 ROC 曲线与 AUC