模型评估

- 机器学习定理
- 经验误差与过拟合
- 模型评估方法
- 模型性能度量

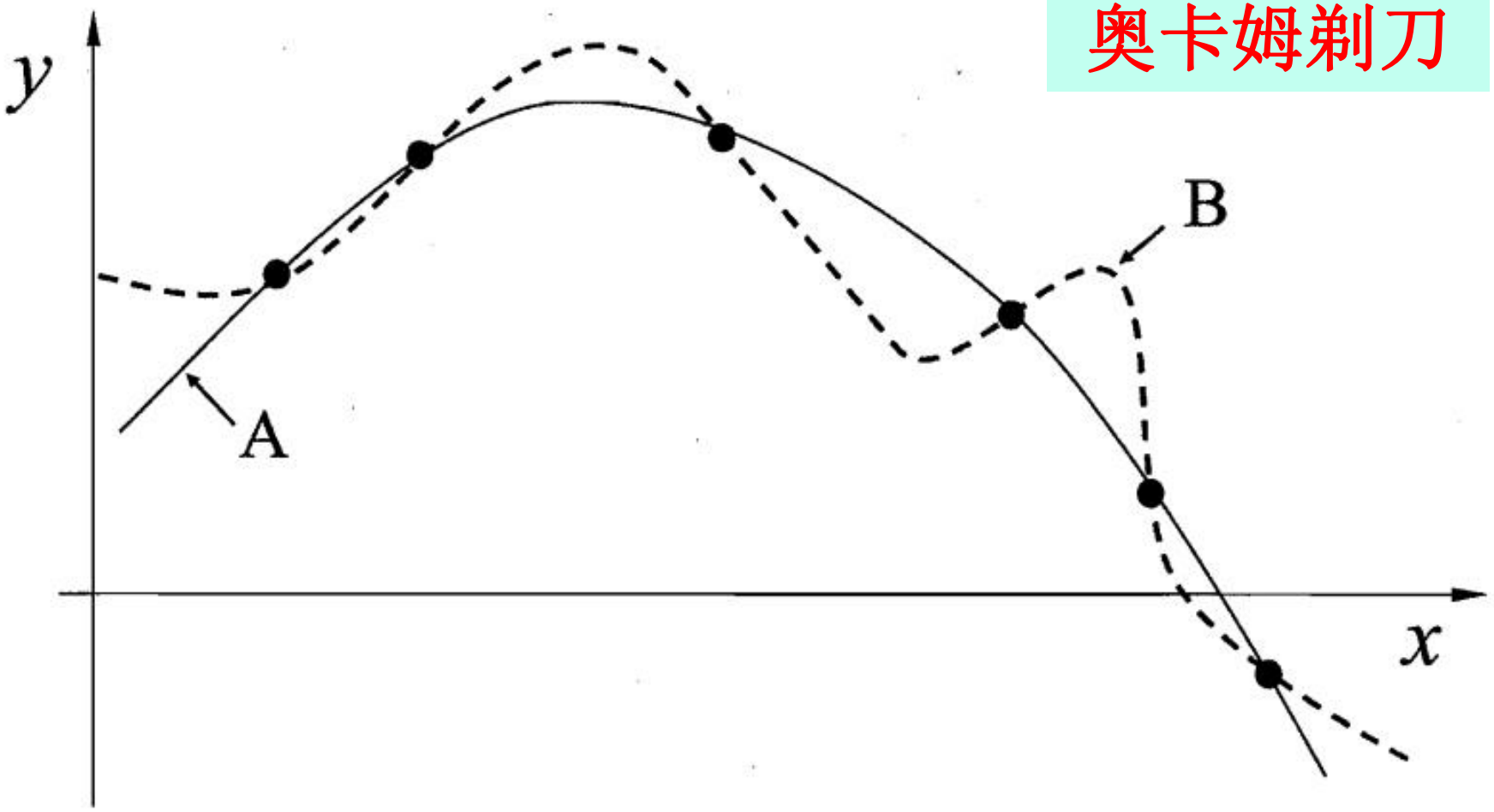
机器学习定理

- 没有免费的午餐
  - 没有天生优越的分类器
- 丑小鸭定理
  - 没有天生优越的特征
- 奥卡姆剃刀原理
  - 不要选用比“必要”更复杂的模型

经验误差与过拟合

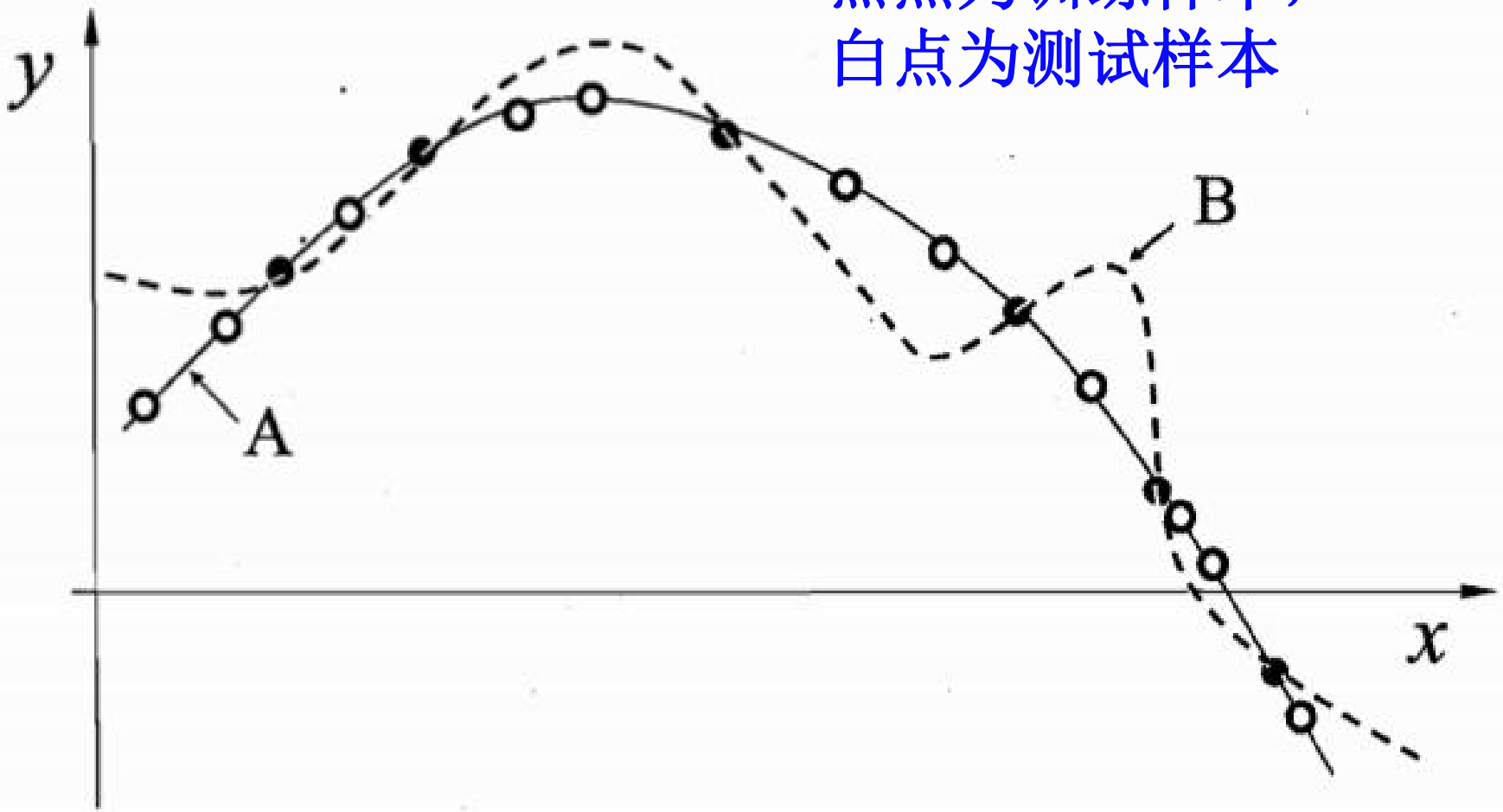
- 经验误差：模型在训练集上的误差
- 泛化误差：模型在测试集上的误差

奥卡姆剃刀



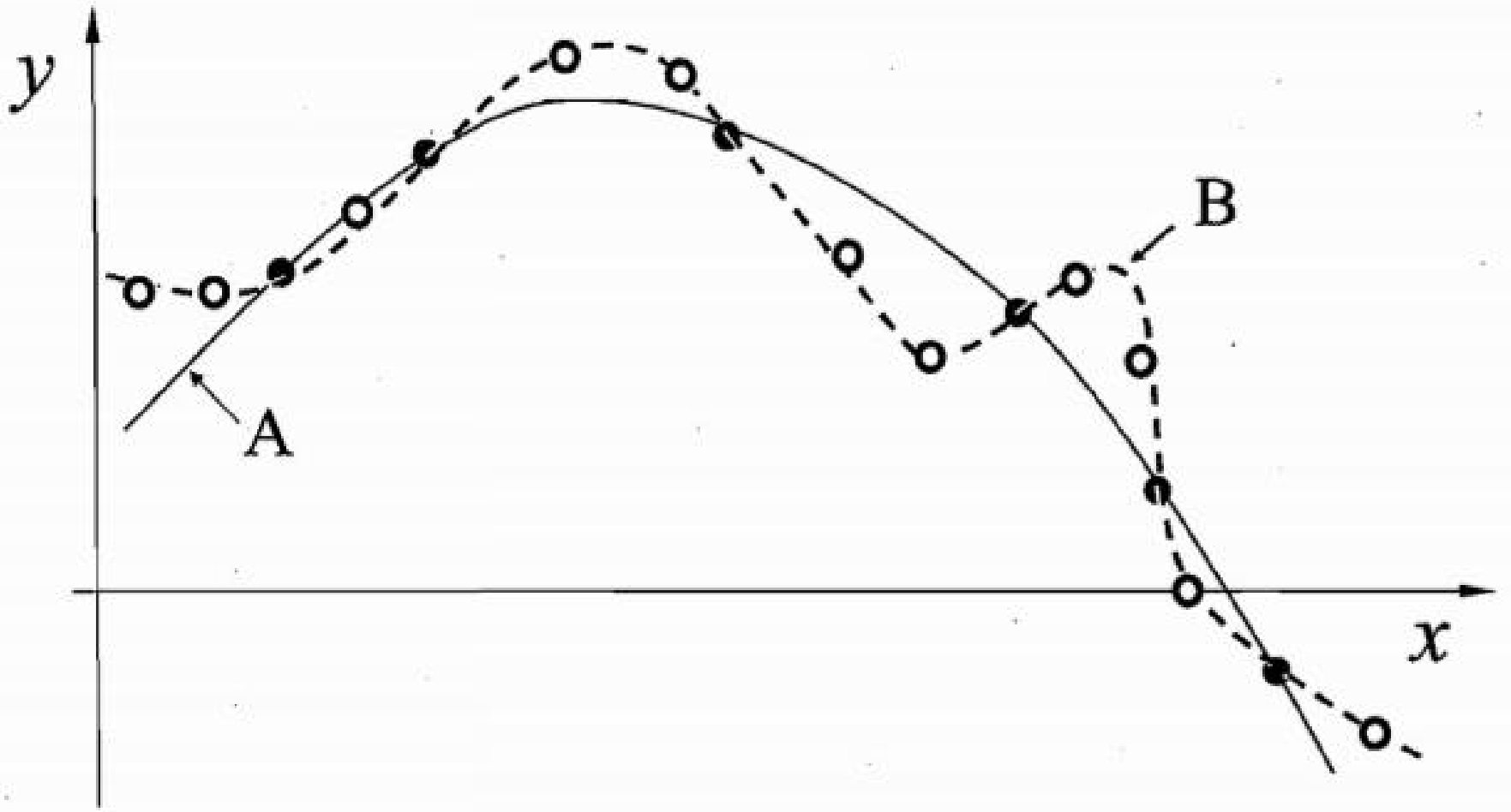
存在多条曲线与有限样本训练集一致

黑点为训练样本，
白点为测试样本



(a) A 优于 B

没有免费的午餐



(b) B 优于 A

模型评估方法

- 在学习得到的模型投放使用之前，通常需要对其进行性能评估。为此，需使用一个“测试集”（testing set）来测试模型对新样本的泛化能力，然后以测试集上的“测试误差”（testing error）作为泛化误差的近似。

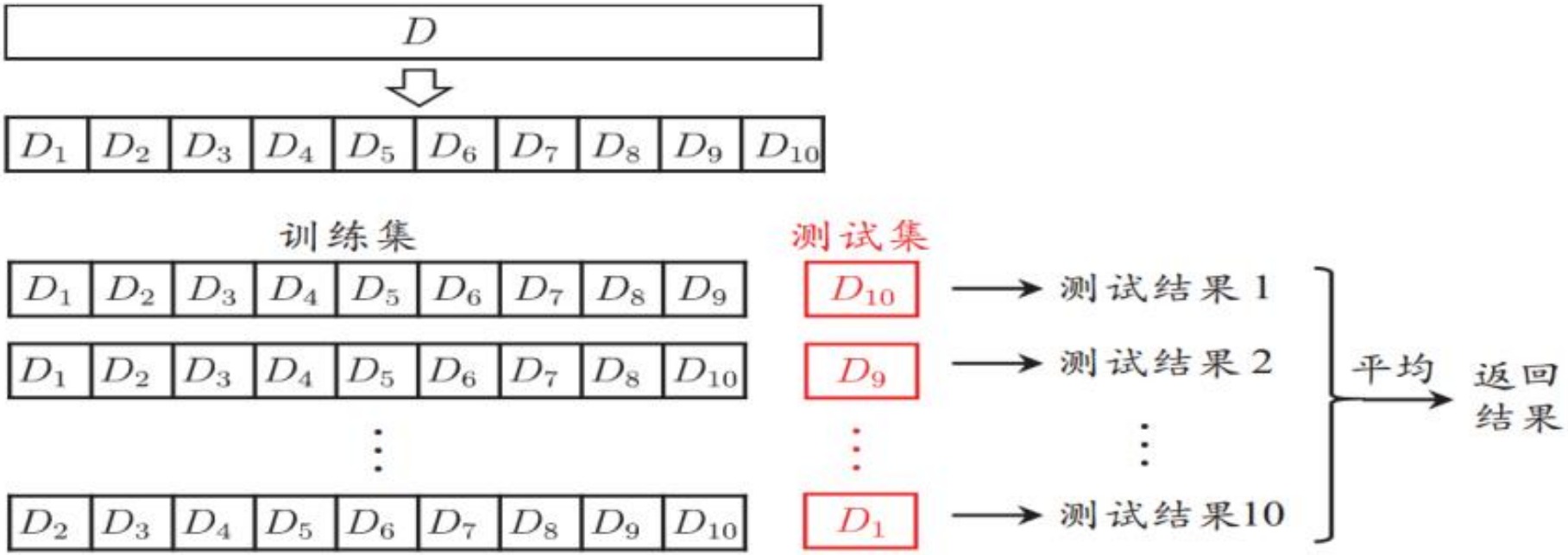
- 我们假设测试集是从样本真实分布中独立采样获得，所以测试集要和训练集中的样本尽量互斥。

- 给定一个已知的数据集，将数据集拆分成训练集S和测试集T，通常的做法包括**留出法、交叉验证法、自助法**。

- 留出法：
  - ✓ 直接将数据集划分为两个互斥集合
  - ✓ 训练/测试集划分要尽可能保持数据分布的一致性
  - ✓ 一般若干次随机划分、重复实验取平均值
  - ✓ 训练/测试样本比例通常为2:1~4:1

- **交叉验证法：**

将数据集分层采样划分为k个大小相似的互斥子集，每次用k-1个子集的并集作为训练集，余下的子集作为测试集，最终返回k个测试结果的均值，k最常用的取值是10。



10 折交叉验证示意图

- 与留出法类似，将数据集D划分为k个子集同样存在多种划分方式，为了减小因样本划分不同而引入的差别，k折交叉验证通常随机使用不同的划分重复p次，最终的评估结果是这p次k折交叉验证结果的均值，例如常见的“10次10折交叉验证”。
- 假设数据集D包含m个样本，若令k=m，则得到留一法：
  - ✓ 不受随机样本划分方式的影响
  - ✓ 结果往往比较准确
  - ✓ 当数据集比较大时，计算开销难以忍受

- **自助法：**

以自助采样法为基础，对数据集D有放回采样m次得到训练集 D', $D \setminus D'$ 用做测试集

- ✓ 实际模型与预期模型都使用m个训练样本
- ✓ 约有1/3的样本没在训练集中出现，用作测试集
- ✓ 从初始数据集中产生多个不同的训练集，对集成学习有很大的好处
- ✓ 自助法在数据集较小、难以有效划分训练/测试集时很有用；由于改变了数据集分布可能引入估计偏差，在数据量足够时，留出法和交叉验证法更常用。

模型评估三大原则

- 奥卡姆剃刀
  - 在性能得到满足的情况下，模型越简单越好
- 数据集划分时的样本采样原则
  - 训练集、测试集和验证集的分布应尽量一致
- 测试集使用原则
  - 训练阶段不要以任何理由偷看测试集
  - 对测试集的反复评估也是一种隐蔽地偷看行为

模型性能度量

准确率与错误率

- 准确率：
  - 分类结果正确的样本数量占总样本数量的比例
- 错误率：
  - 分类结果错误的样本数量占总样本数量的比例
- 特点
  - 准确率 + 错误率 = 1
  - 每个样本在统计时的权重相同
  - 适用于类样本平衡的数据集

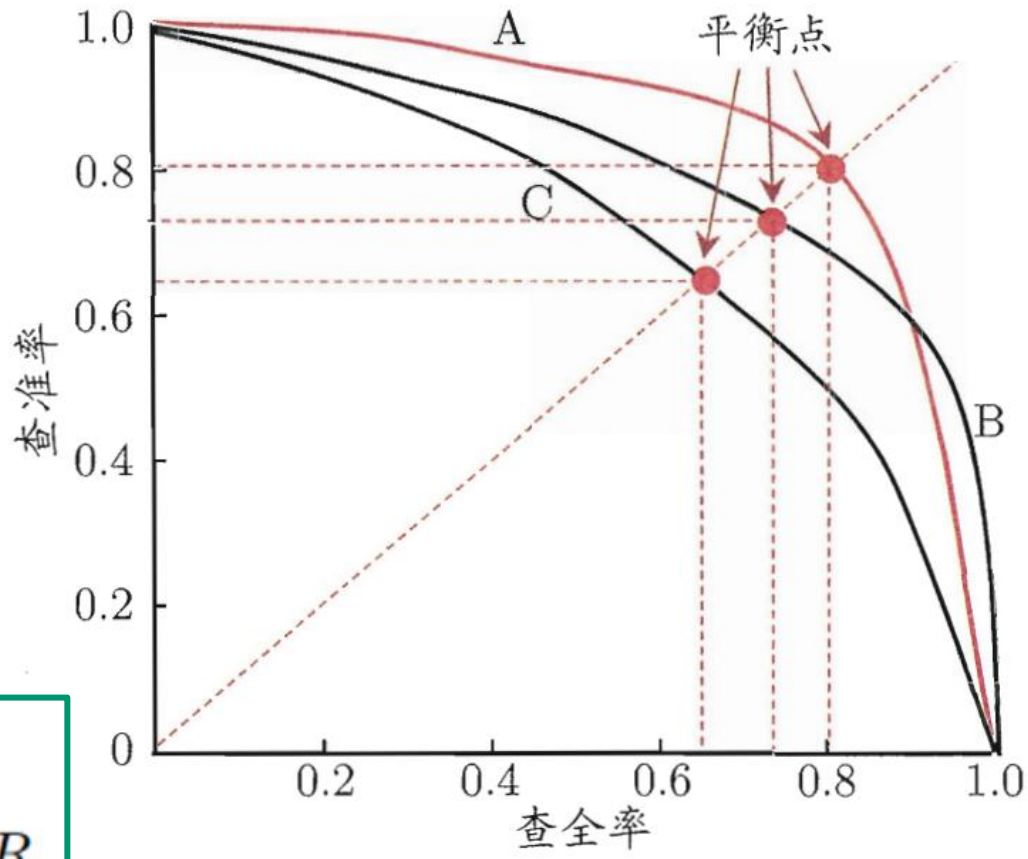
查准率与查全率

- 类样本不平衡
  - 稀有样本类别为正类，其余样本为负类。
  - 示例
    - 10个正类样本，990个负类样本。
    - 若全部分类为负类，则准确率高达99%，但无意义。
  - 应用：网页搜索、癌症筛查等

- **查准率：**
  - 分类结果为正类的样本中，实际结果为正类的比例。
  - **P=TP / (TP + FP)**
- **查全率：**
  - 实际结果为正类的样本中，分类结果为正类的比例。
  - **R=TP / (TP +FN)**

真实情况	预测结果	
	正例	负例
正例	TP（真正例）	FN（假负例）
负例	FP（假正例）	TN（真负例）

- **平衡点**
  - P与R往往是矛盾的
  - P = R
- **F1度量：**
  - 比平衡点更高效
  - F1=2*P*R / (P + R)



比F1更一般的形式 F_β，

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

$\beta = 1$：标准的F1
$\beta > 1$：偏重查全率
$\beta < 1$：偏重查准率