



中國農業大學
China Agricultural University

人工智能——机器学习概述

胡标





目录

Contents

1. 机器学习简介

1.1 概念与原理

1.2 机器学习算法分类

1.3 常用定理

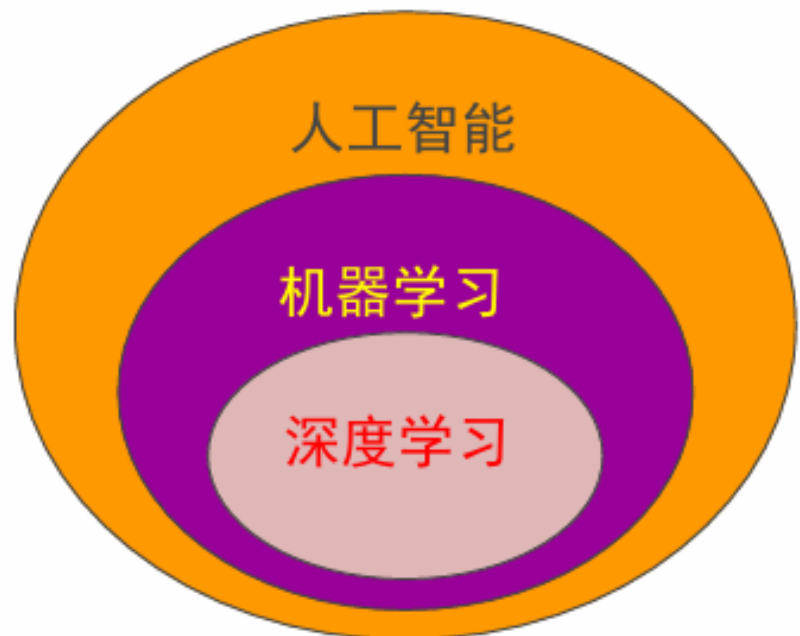
2. 模型评估与选择

2.1 评估方法

2.2 性能度量

2.3 比较检验

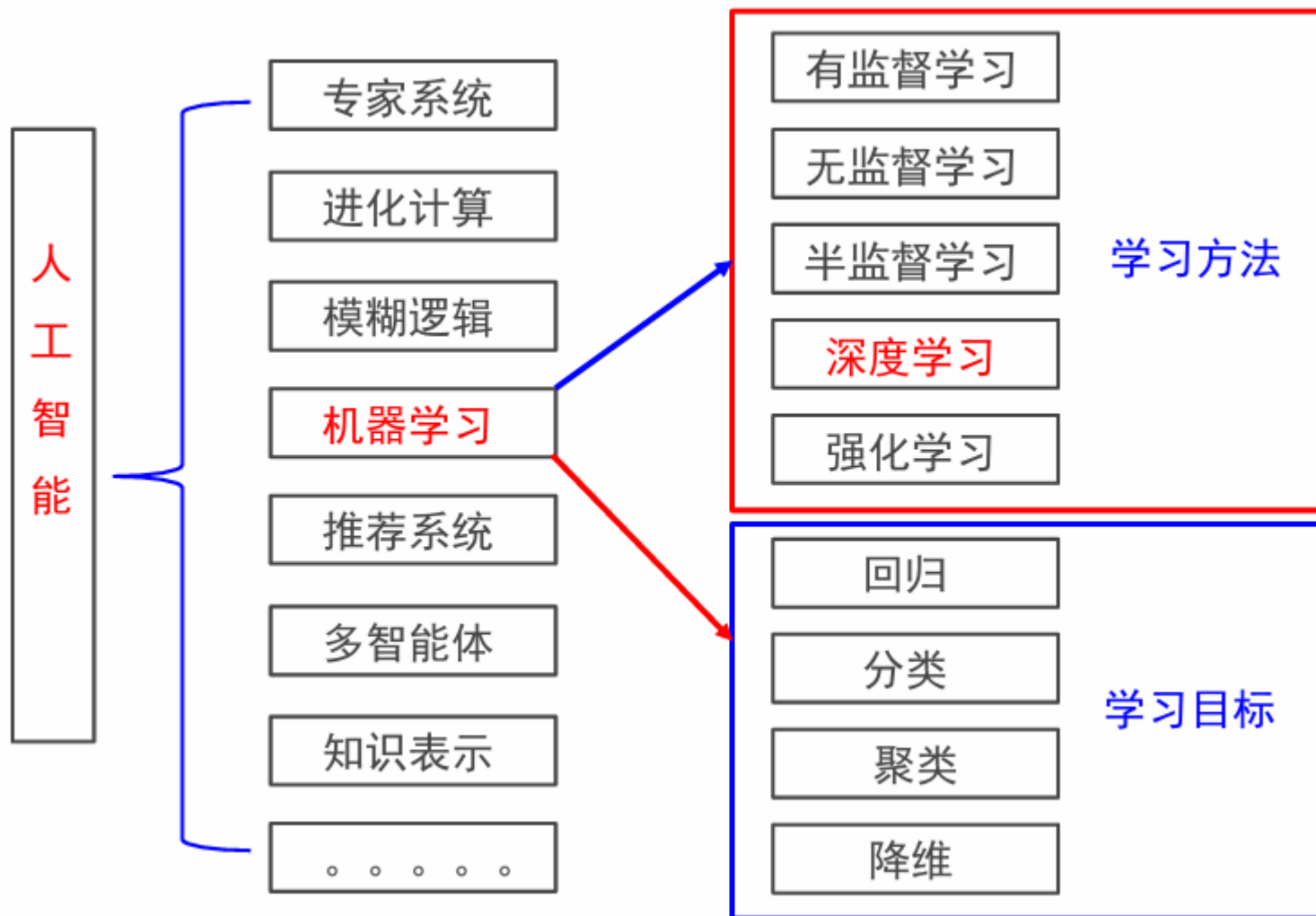
- 人工智能、机器学习和深度学习三者之间的关系？



三者覆盖的技术范畴是逐层递减的，人工智能是最宽泛的概念。机器学习是人工智能的一个子集，有些是可以不用机器学习的算法进行智能化的运算

机器学习：专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能

深度学习是机器学习算法中最热的一个分支，在近些年取得了显著的进展，并代替了多数传统机器学习算法



- 关于机器学习，引用卡内基梅隆大学（CMU）Tom Michael Mitchell教授在其1997年出版的书籍 Machine Learning 中的定义：
 - 如果一个程序可以在任务 **T** 上，随着经验 **E** 的增加，效果 **P** 也可以随之增加，则称这个程序可以从经验中学习
- 简而言之，机器的“学习”，是通过以往的经验，即历史数据（经验），学习数据内部的逻辑模型，并将学到的模型应用在新数据上，进行预测

- 举例：预测校车到达时间问题描述：每天早上校车从东区 7:10 发往西校区，到达东校区地的时间如何准确预测？
- 如果第一次乘坐班车，预测通常不太准。一周之后大概能预测出 7:50 左右到达西校区
- 一个月之后，随着经验的增加你还会知道周一常堵车会晚 10 分钟，下雨常堵车会晚 20 分钟。于是可以总结出一张树状图：

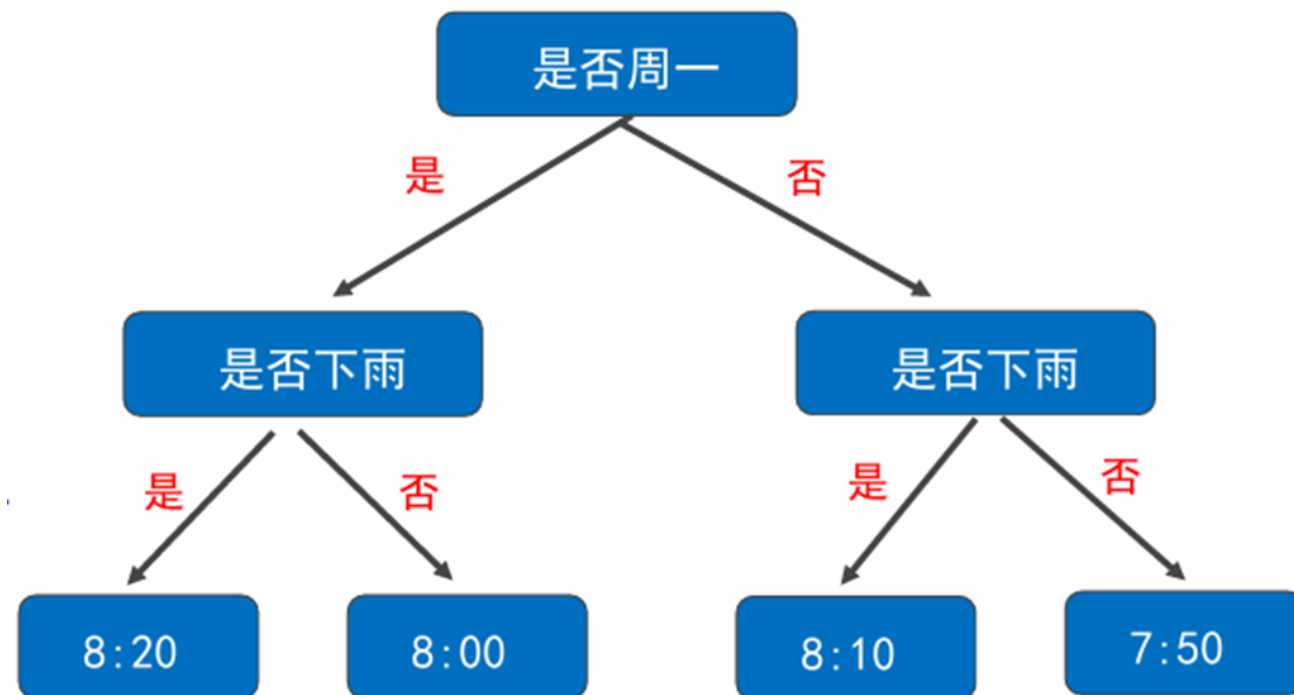
□ 这张图就是一张简单的机器是否周一学习模型，称为决策树模型

– 如果一个程序可以在任务是否 **T** 上，随着经验 **E** 的增加，效果 **P** 也可以随之增加，则称这个程序可以从经验中学习

– **T**: 预测班车时间

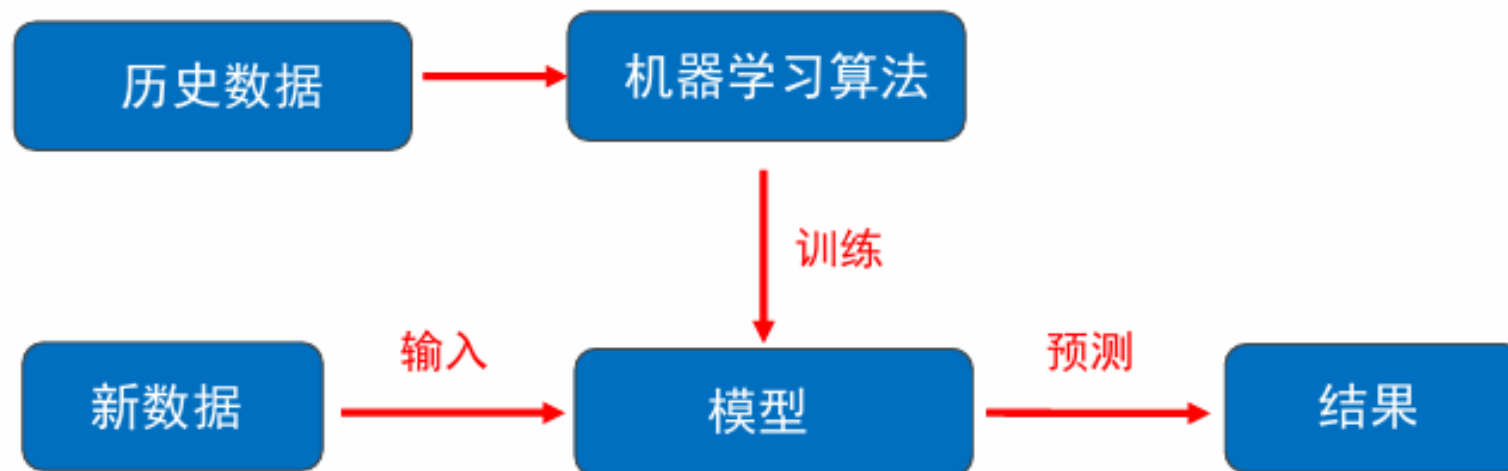
– **E**: 坐班车次数的增加

– **P**: 预测越来越准



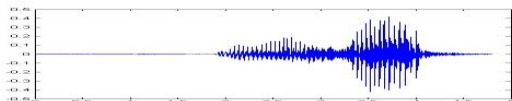
机器学习是一种**统计方法**，计算机利用已有**数据**得出某**模型**，再利用此模型预测结果

- 特点是先用已有**数据**训练**模型**，再用模型**预测新数据**的结果



- 机器学习通常是从已知数据中去学习数据中蕴含的规律或者判断规则，从数据中训练出模型，然后通过**迭代**学习去不断优化模型，把学到的模型应用到未来的新**数据**上并作出判断或预测
- 机器学习中的“训练”与“预测”过程可以对应到人类的“归纳”和“推测”过程

- 机器学习中的每一种算法，究其根本，都是一种**数学**表达。无论是机器学习，还是深度学习，都是试图找到一个函数
- 这个函数可以简单，可以复杂，函数的表达并不重要，只是一个工具。重要的是这个函数能够尽可能准确的拟合出输入数据和输出结果间的关系

– 语音识别 $f(\text{  }) = \text{“你好”}$

– 图像识别 $f(\text{  }) = \text{“猫”}$

– 围棋 $f(\text{  }) = \text{“5-5” (落子位置)}$

□ 机器学习 \approx 构建一个映射函数



目录

Contents

1. 机器学习简介

1.1 概念与原理

1.2 机器学习算法分类

1.3 常用定理

2. 模型评估与选择

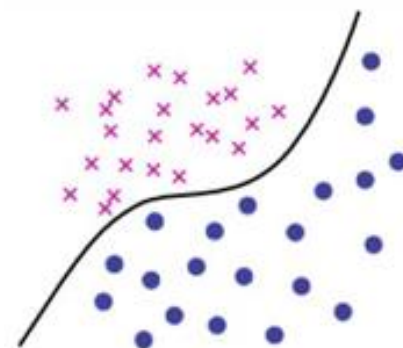
2.1 评估方法

2.2 性能度量

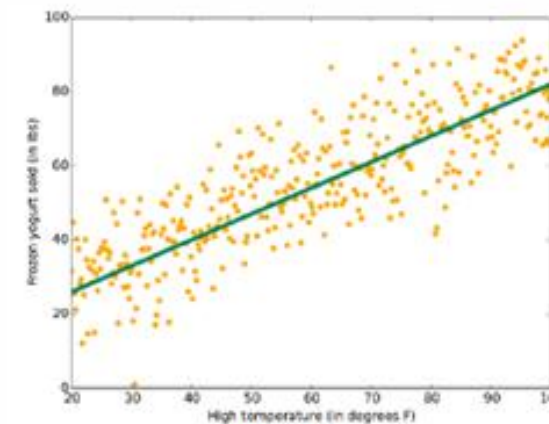
2.3 比较检验

□ 机器学习算法可以分为四类：

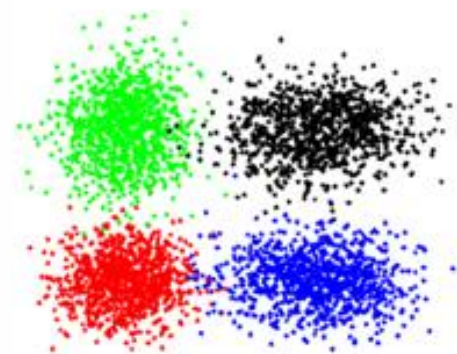
- 分类
- 回归
- 聚类
- 降维



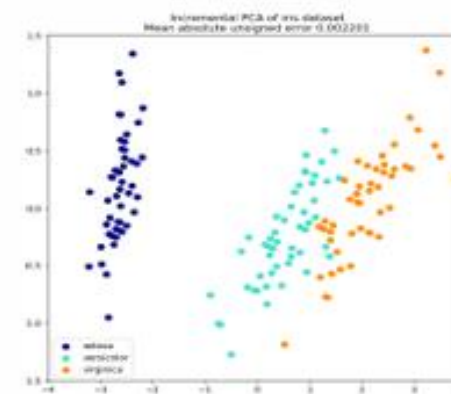
分类



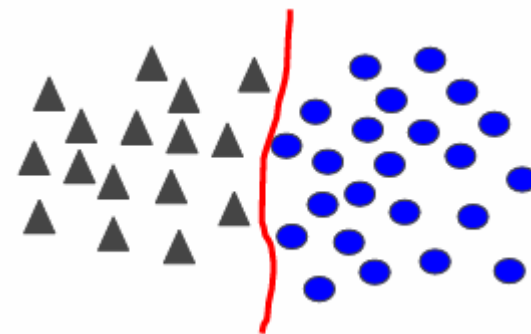
回归



聚类



降维



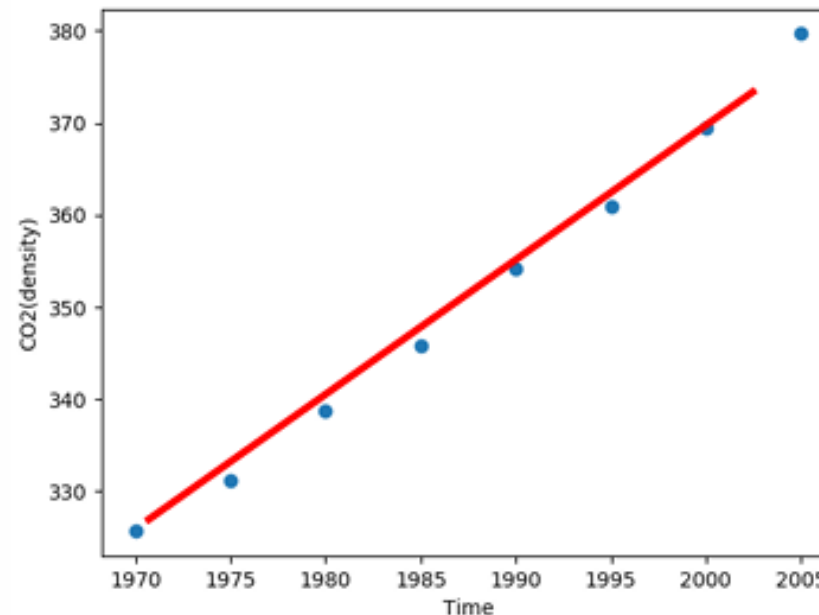
- 图中的每一个点代表一个样本，而一个样本由一组数据构成
- 这些数据 (X) 实际上都有一个标签 (Y)。 Y 的值要么是蓝色圆，要么是黑三角
- 分类算法就是根据已知这些数据 (X) 的分类 (Y)，找到这个边界
- 一旦进来一批新的没有标签的数据 (X)，那么可以根据这个算法把这些数据 (X) 要么归类到蓝色圆，要么归类到黑三角

□ 通过这些样本点拟合出来一条直线（曲线），称为回归曲线。这条线就是通过机器学习得到

□ 分类和回归有什么区别？

□ 当标签是离散的，就是分类。当标签是连续的，就是回归

□ 分类是解决离散数据的归类，回归是连续数据

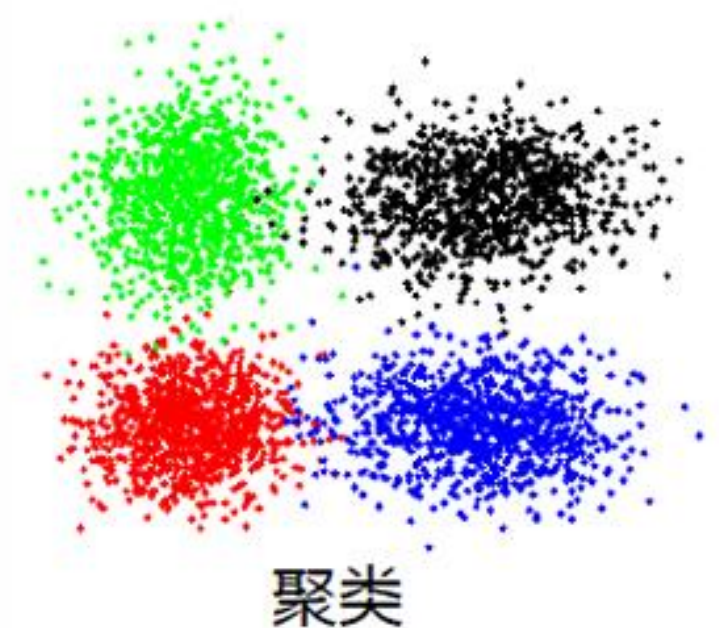


□ 不管**分类**还是**回归**，都要有一个标签（label）。

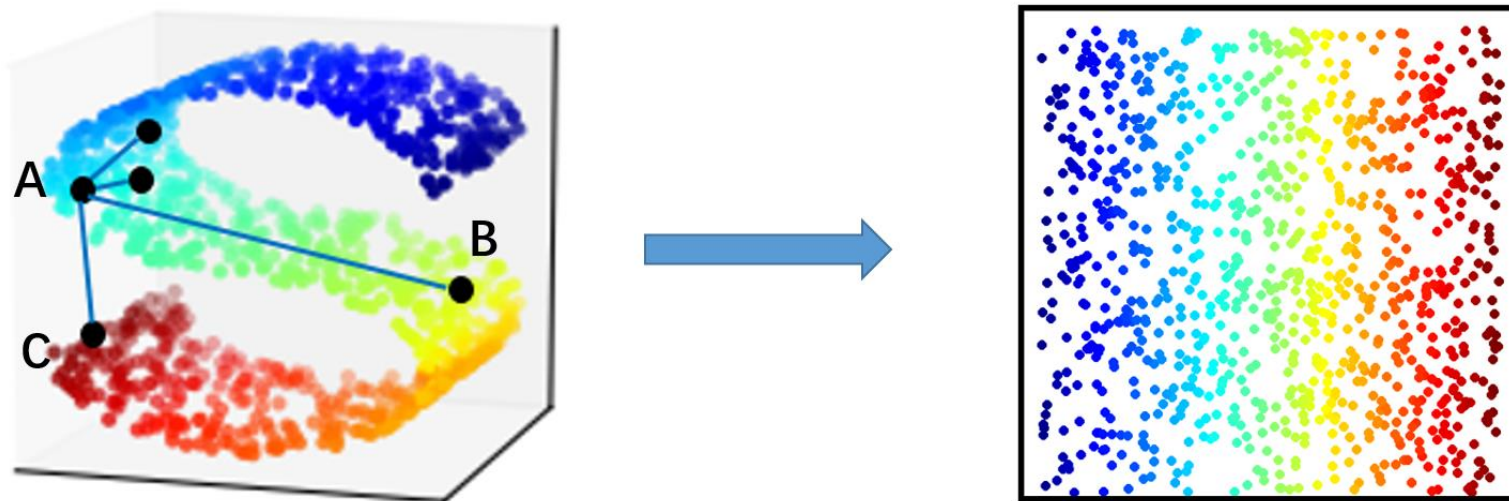
而**聚类**是没有标签的数据

□ 输入被划分为若干个事先**未知**的组。通过一个算法学习，把它们归纳出几类

□ 但并不知道这一类代表**哪一类**，因为事先数据没有**标签**



- ❑ 降维是机器学习另一个重要的领域。特征的维数过高, 会增加训练的负担与存储空间, 降维就是希望去除特征的冗余, 用更加少的维数来表示特征输入被划分为若干个事先~~未知~~的组。通过一个算法学习, 把它们归纳出几类
- ❑ 降维是试图压缩维度, 并尽可能地保留分布信息。降维在图像处理中叫图像压缩、特征提取



□ 训练样本

- 训练样本（**历史数据**）相当于平时的各种学习资料和练习，目的是帮助提升学习能力

□ 测试样本

- **测试**样本类似于考试试卷，检测学习成绩

□ 标签

- **标签** (label) 可以认为是习题和考试的 **标准答案**

□ 特征

- 每个样本包含的多个属性（多维数据）被称作 **“特征”**

□ 样本的属性和特征

- 每个样本可以包含多个属性。如 excel 表，**每一行数据是一个样本，每一列就是属性**

1	序号	课程号	课程名	课序号	校区	原授课教师(*负责)	所属单位
2	1	58308010	计算机网络与实验(双学位)	90	东校区	黄燕*	双学位
3	2	58308012	项目实训III(双学位)	90	东校区	雷宏洲*	双学位
4	3	58308016	Linux应用基础与实验(双学位)	91	东校区	李振波*	双学位
5	4	58308016	Linux应用基础与实验(双学位)	90	东校区	李振波*	双学位
6	5	58308017	Web技术应用与实验(双学位)	91	东校区	吕春利*	双学位
7	6	58308017	Web技术应用与实验(双学位)	90	东校区	吕春利*	双学位
8	7	58308018	程序设计与实验(双学位)	90	东校区	吕春利*	双学位
9	8	58308018	程序设计与实验(双学位)	91	东校区	李辉*	双学位
10	9	58308019	数据库原理与实验(双学位)	90	东校区	陈雷*	双学位

- 也可以简单地把属性看成**特征**，但属性和特征稍微有些区别，比如通过几个属性的组合可以构造出特征，当然这个构造出的特征是不在原来的属性里

□ 监督学习

- 学习的是带有标记的数据

□ 非监督学习

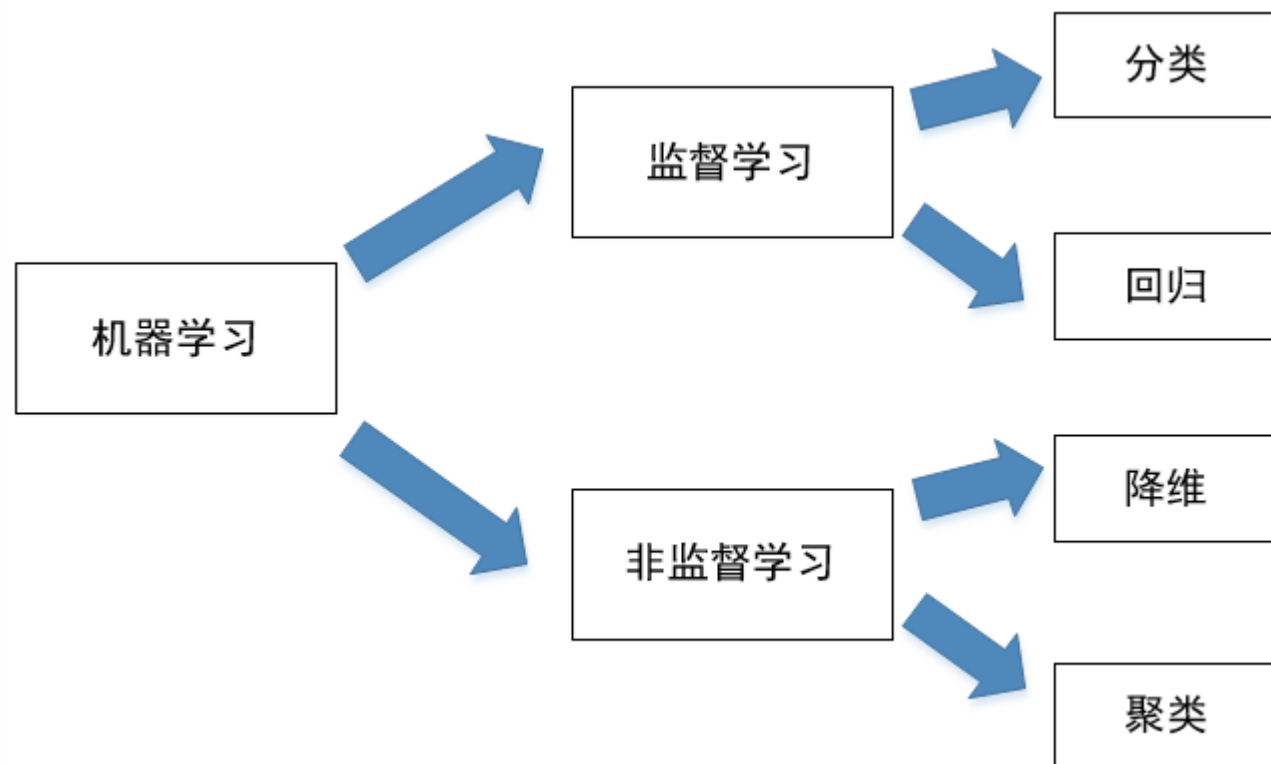
- 学习的是未被标记的数据

□ 半监督学习

- 部分数据有标签，部分数据无标签

- 要让计算机具备学习能力，首先要告诉它哪些是对的，哪些是错的？哪些是想要的，哪些是不想要的
- 监督学习最大的特点是样本数据有**标签**。比如哪些数据属于第一类，哪些数据是第二类，算法采用标注好的数据进行训练半监督学习
- 然后机器学习算法从带标签的样本数据中发现规律得出模型，再给一组新的数据，用该模型做预测和判断：是属于第一类还是第二类
- 监督学习主要解决两类问题：**分类** 和 **回归**

- 输入的数据没有标签，即没有事先给数据打好标签
- 一般解决的问题是**聚类**，物以类聚人以群分。直接给应用数据，让计算机自动在数据中发现规律并且进行分类。在没有给定划分类的情况下，根据信息相似度进行信息聚类
- 聚类的输入是一组未被标记的数据，根据样本特征的**距离**或相似度进行划分。算法仅接收未标注的数据进行学习并得出模型，再对所有的未知数据做出预测



- 总之，机器学习是有一个基本的模型，有大量训练的数据，通过机器学习算法训练出预测的模型。有了模型以后，再把新的数据输入进来，就能获得相应的实际预测结果



目录

Contents

1. 机器学习简介

1.1 概念与原理

1.2 机器学习算法分类

1.3 常用定理

2. 模型评估与选择

2.1 评估方法

2.2 性能度量

2.3 比较检验

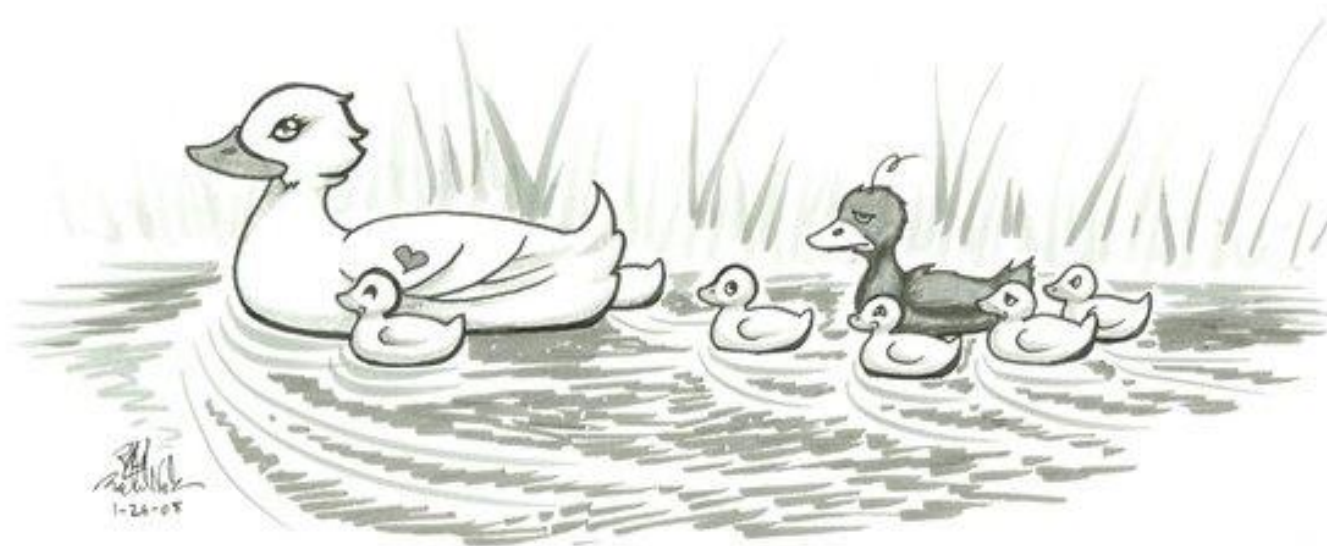
□ 没有免费午餐定理 (No Free Lunch Theorem, NFL)

- 对于基于迭代的最优化算法，不存在某种算法对所有问题（有限的搜索空间内）都有效。
如果一个算法对某些问题有效，那么它一定在另外一些问题上比纯随机搜索算法更差。



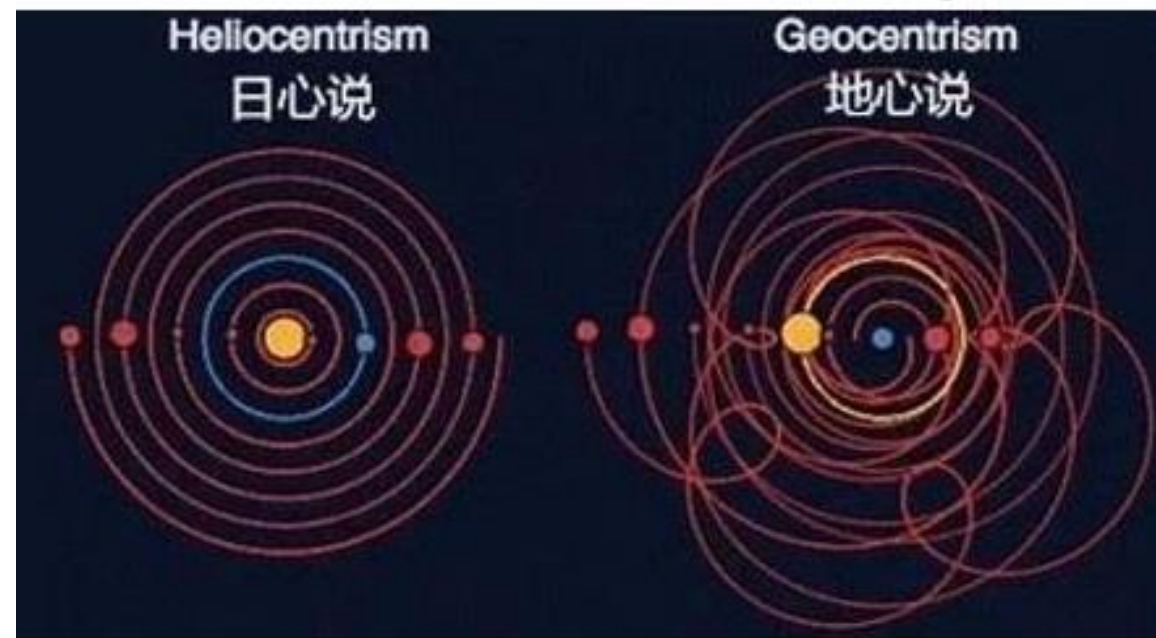
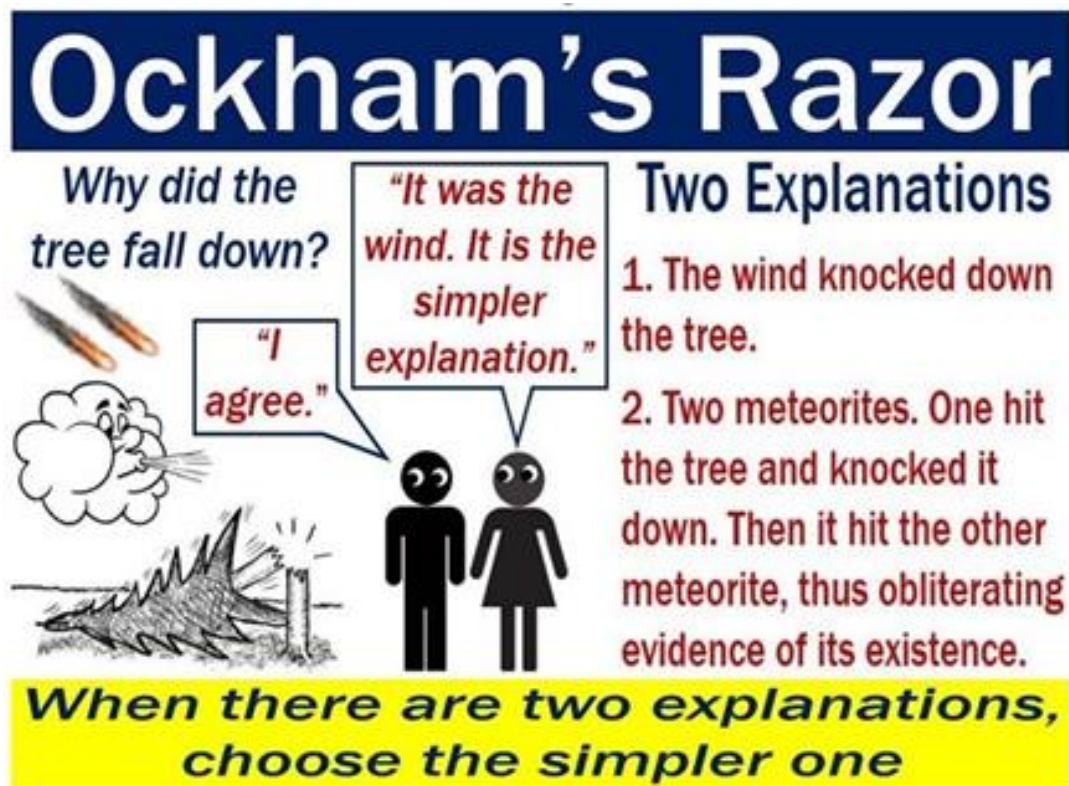
□ 丑小鸭定理(Ugly Duckling Theorem)

□ 丑小鸭与白天鹅之间的区别和两只白天鹅之间的区别一样大



□ 奥卡姆剃刀原理(Occam's Razor)

□ 如无必要，勿增实体 (Entities should not be multiplied unnecessarily)





目录

Contents

1. 机器学习简介

1.1 概念与原理

1.2 机器学习算法分类

1.3 常用定理

2. 模型评估与选择

2.1 评估方法

2.2 性能度量

2.3 比较检验

- 泛化误差：在“未来”样本上的误差，**越小越好？**
- 经验误差：在训练集上的误差，亦称“训练误差”，**越小越好？**



□ 如何获得测试结果？



评估方法

□ 如何评估性能优劣？



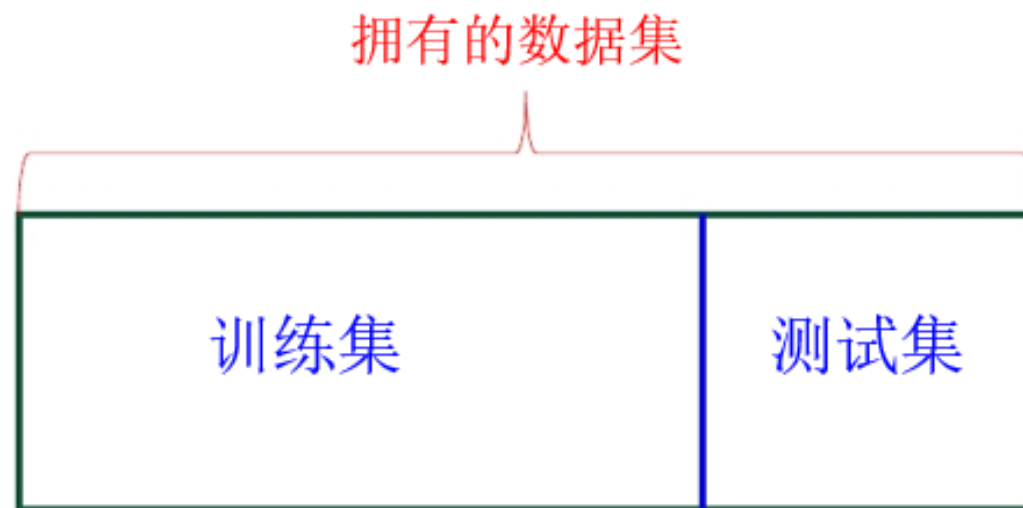
性能度量

□ 如何判断实质差别？



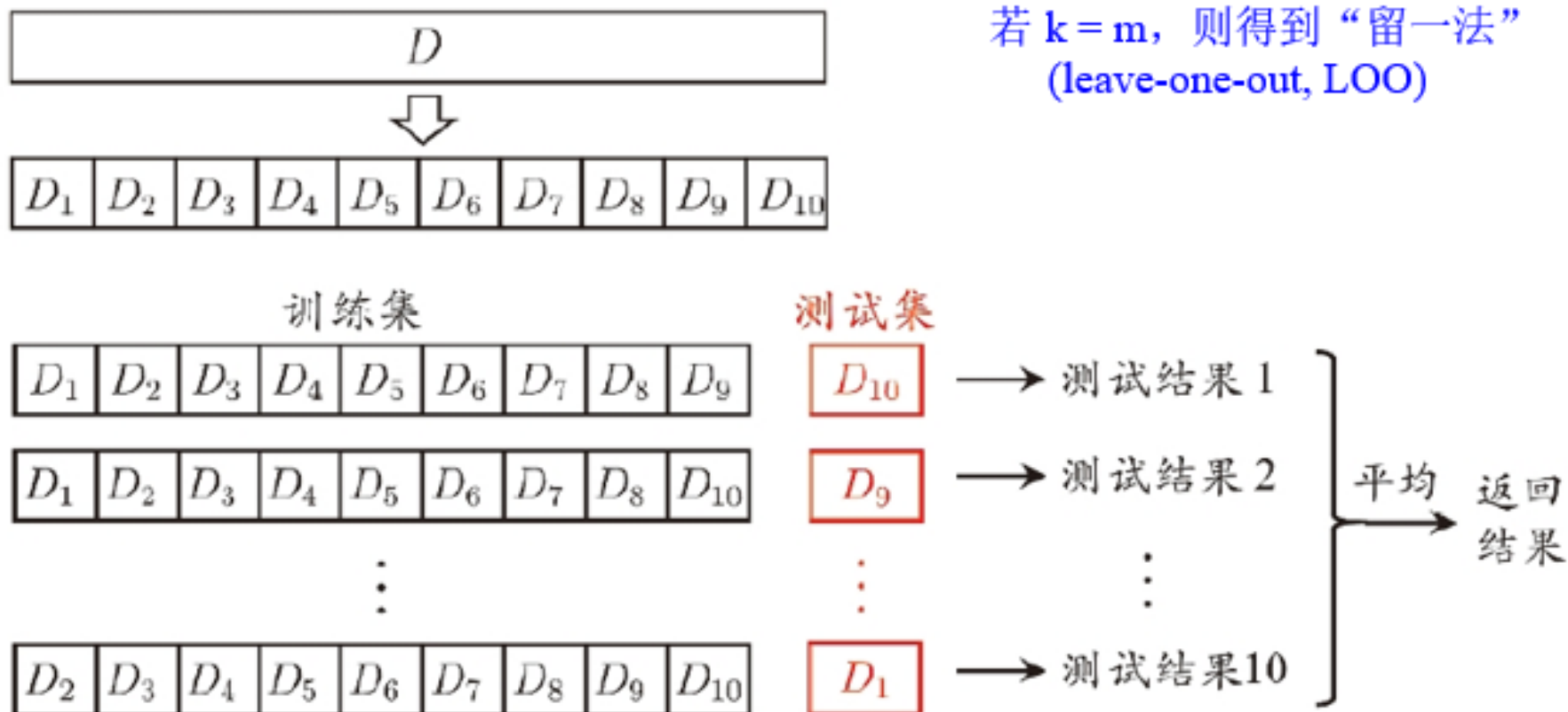
比较检验

- 关键：怎么获得“测试集” (test set) ?
 - 测试集应该与训练集“互斥”
- 常见方法：
 - 留出法 (hold-out)
 - 交叉验证法 (cross validation)
 - 自助法 (bootstrap)

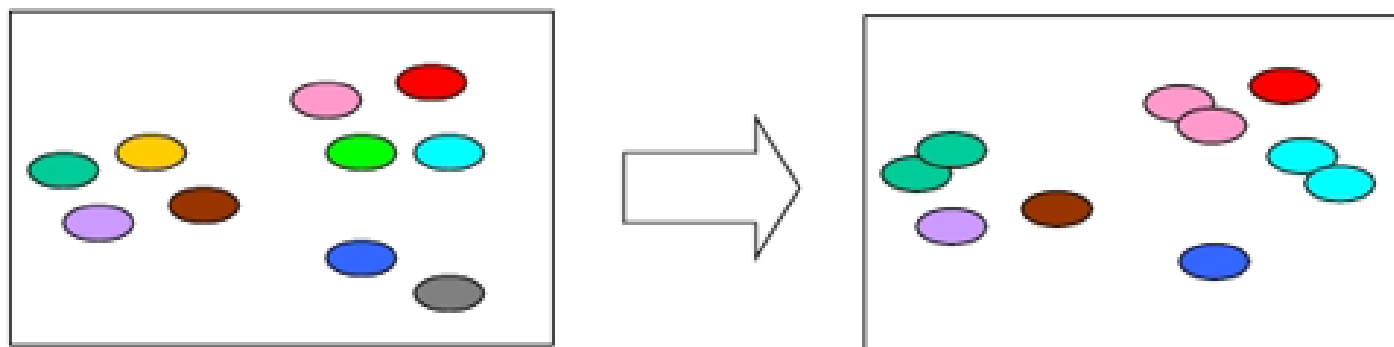


注意：

- 保持数据分布一致性（例如：分层采样）
- 多次重复划分（例如：100次随机划分）
- 测试集不能太大、不能太小（例如：1/5~1/3）



基于“自助采样” (bootstrap sampling)
亦称“有放回采样”、“可重复采样”



约有 36.8% 的样本不出现

$$\downarrow \lim_{m \rightarrow \infty} \left(1 - \frac{1}{m}\right)^m \rightarrow \frac{1}{e} \approx 0.368$$

训练集与原样本集同规模

数据分布有所改变

“包外估计” (out-of-bag estimation)

算法的参数：一般由人工设定，亦称“超参数”

模型的参数：一般由学习确定

调参过程相似：先产生若干模型，然后基于某种评估方法进行选择

参数调得好不好对性能往往对最终性能有关键影响

区别：训练集 vs. 测试集 vs. 验证集 (validation set)

算法参数选定后，要用“训练集+验证集”重新训练最终模型



目录

Contents

1. 机器学习简介

1.1 概念与原理

1.2 机器学习算法分类

1.3 常用定理

2. 模型评估与选择

2.1 评估方法

2.2 性能度量

2.3 比较检验

性能度量(performance measure)是衡量模型泛化能力的评价标准, 反映了任务需求

使用不同的性能度量往往会导致不同的评判结果

什么样的模型是“好”的, 不仅取决于算法和数据, 还取决于任务需求

回归(regression) 任务常用均方误差:

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

错误率:

$$E(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{N}(f(x_i) \neq y_i)$$

精度:

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

查准率 vs. 查全率

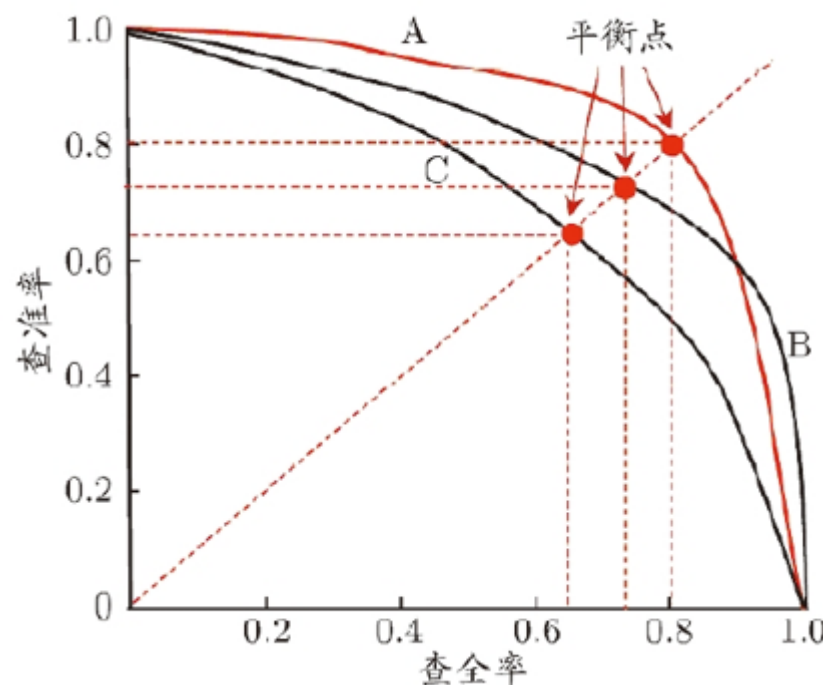


真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

查准率: $P = \frac{TP}{TP+FP}$

查全率: $R = \frac{TP}{TP+FN}$

根据学习器的预测结果按正例可能性大小对样例进行排序，并逐个把样本作为正例进行预测



PR图:

- 学习器 A 优于 学习器 C
- 学习器 B 优于 学习器 C
- 学习器 A 优于 学习器 B

BEP:

- 学习器 A 优于 学习器 B
- 学习器 A 优于 学习器 C
- 学习器 B 优于 学习器 C

比 BEP 更常用的 F1 度量:

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

$$\frac{1}{F1} = \frac{1}{2} \times \left(\frac{1}{P} + \frac{1}{R} \right)$$

若对查准率/查全率有不同偏好:

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

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

$\beta > 1$ 时查全率有更大影响; $\beta < 1$ 时查准率有更大影响



目录

Contents

1. 机器学习简介

1.1 概念与原理

1.2 机器学习算法分类

1.3 常用定理

2. 模型评估与选择

2.1 评估方法

2.2 性能度量

2.3 比较检验

在某种度量下取得评估结果后，是否可以直接比较以评判优劣？

- NO ! 因为：**
- 测试性能不等于泛化性能
 - 测试性能随着测试集的变化而变化
 - 很多机器学习算法本身有一定的随机性

机器学习 \Rightarrow “概率近似正确”

□ PAC: **Probably Approximately Correct**

□ 根据大数定律，当训练集大小 $|D|$ 趋向无穷大时，泛化错误趋向于0，即经验风险趋近于期望风险。

$$\lim_{|D| \rightarrow \infty} \mathcal{R}(f) - \mathcal{R}_D^{emp}(f) = 0$$

□ PAC学习

$$P\left((\mathcal{R}(f) - \mathcal{R}_D^{emp}(f)) \leq \epsilon\right) \geq 1 - \delta$$

近似正确, $0 < \epsilon < 0.5$

可能, $0 < \delta < 0.5$

统计假设检验 (hypothesis test) 为学习器性能比较提供了重要依据



统计显著性

□ 两学习器比较

□ 交叉验证 t 检验 (基于成对 t 检验)

k 折交叉验证; 5x2交叉验证

□ McNemar 检验 (基于列联表, 卡方检验)

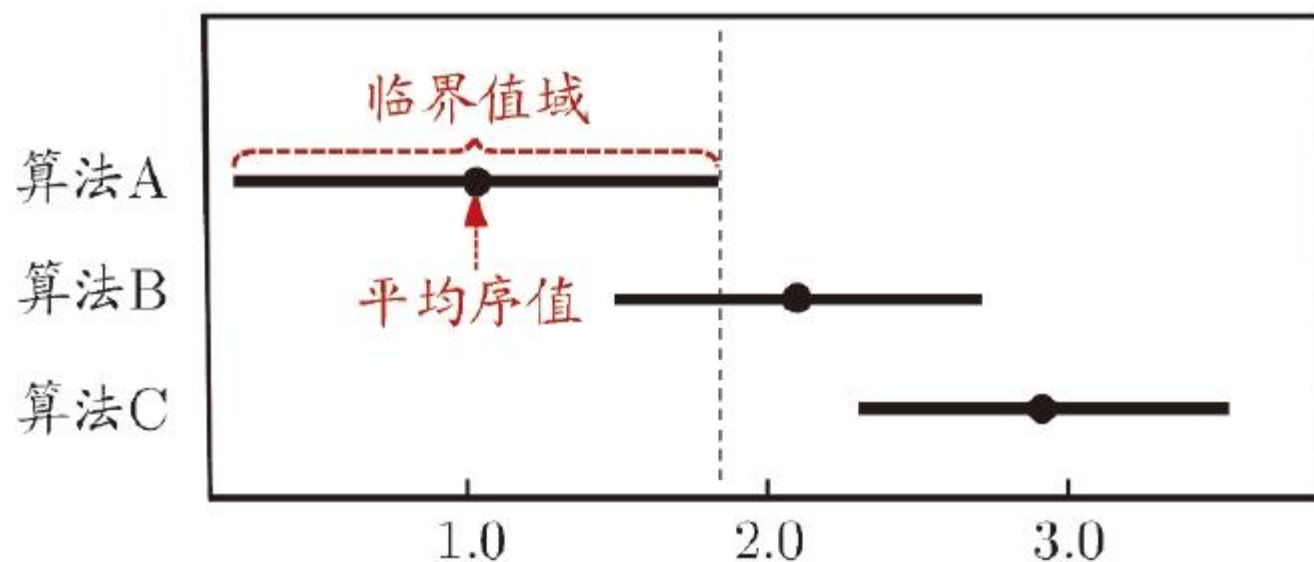
□ 多学习器比较

□ Friedman + Nemenyi

□ Friedman检验 (基于序值, F检验; 判断“是否都相同”)

□ Nemenyi 后续检验 (基于序值, 进一步判断两两差别)

横轴为平均序值，每个算法圆点为其平均序值，线段为临界阈值的大小



若两个算法有交叠 (A 和 B)，则说明没有显著差别；
否则有显著差别 (A 和 C)，算法 A 显著优于算法 C



中國農業大學
China Agricultural University

数据如海藏玄机，**模型**筑梦绘天机。
训练方法精雕琢，**智见**未来展新奇。

