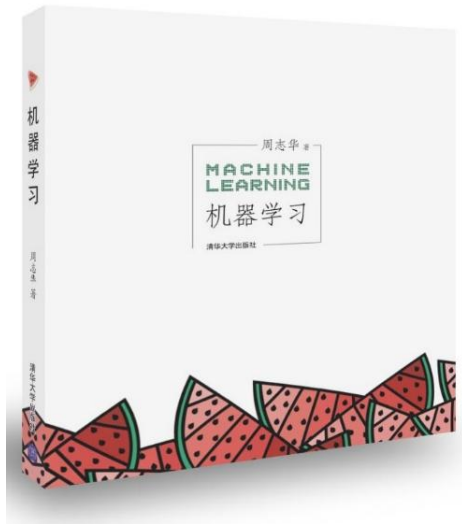
机器学习导论

授课材料

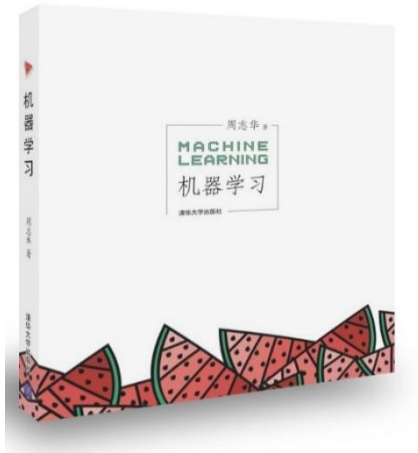
周志华 著. 机器学习,
北京: 清华大学出版社,
2016年1月.
425页，62.6万字
16 章，3 附录
ISBN: 978-7-302-206853-6
2016年1月第1次印刷
2020年11月第35次印刷



附录请自行阅读

本学期讲授前 9-10 章

建议使用方式



1. 初学机器学习的第一本书：
   通读、速读；细节不懂处略过
   了解机器学习的疆域和基本思想，理解基本概念
   “观其大略”

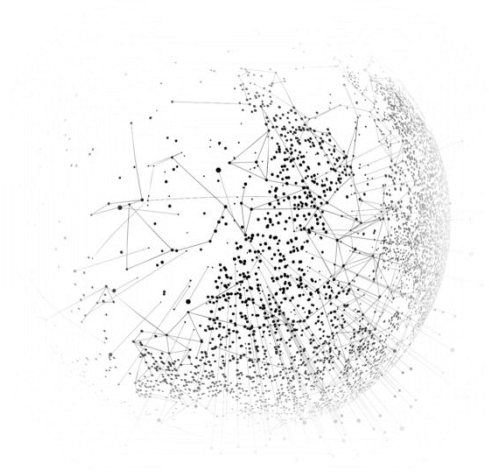
2. 阅读其他关于机器学习具体分支的读物（三月、半年）

3. 再读、对“关键点”的理解：
   理解技术细冗后的本质，升华认识
   “提纲挈领”

4. 对机器学习多个分支有所了解（1-3年）

5. 再读、细思：
   不同内容的联系，不同的描述方式、出现位置蕴涵的意义、……个别字句的启发可能自行摸索数年不易得
   “疏通经络”

http://www.lamda.nju.edu.cn/zhouzh/zhouzh.files/publication/MLbook2016.htm



一、绪论

机器学习

经典定义：利用经验改善系统自身的性能 [T. Mitchell 教科书, 1997]

经验 → 数据

随着该领域的发展，目前主要研究智能数据分析的理论和方法，并已成为智能数据分析技术的源泉之一

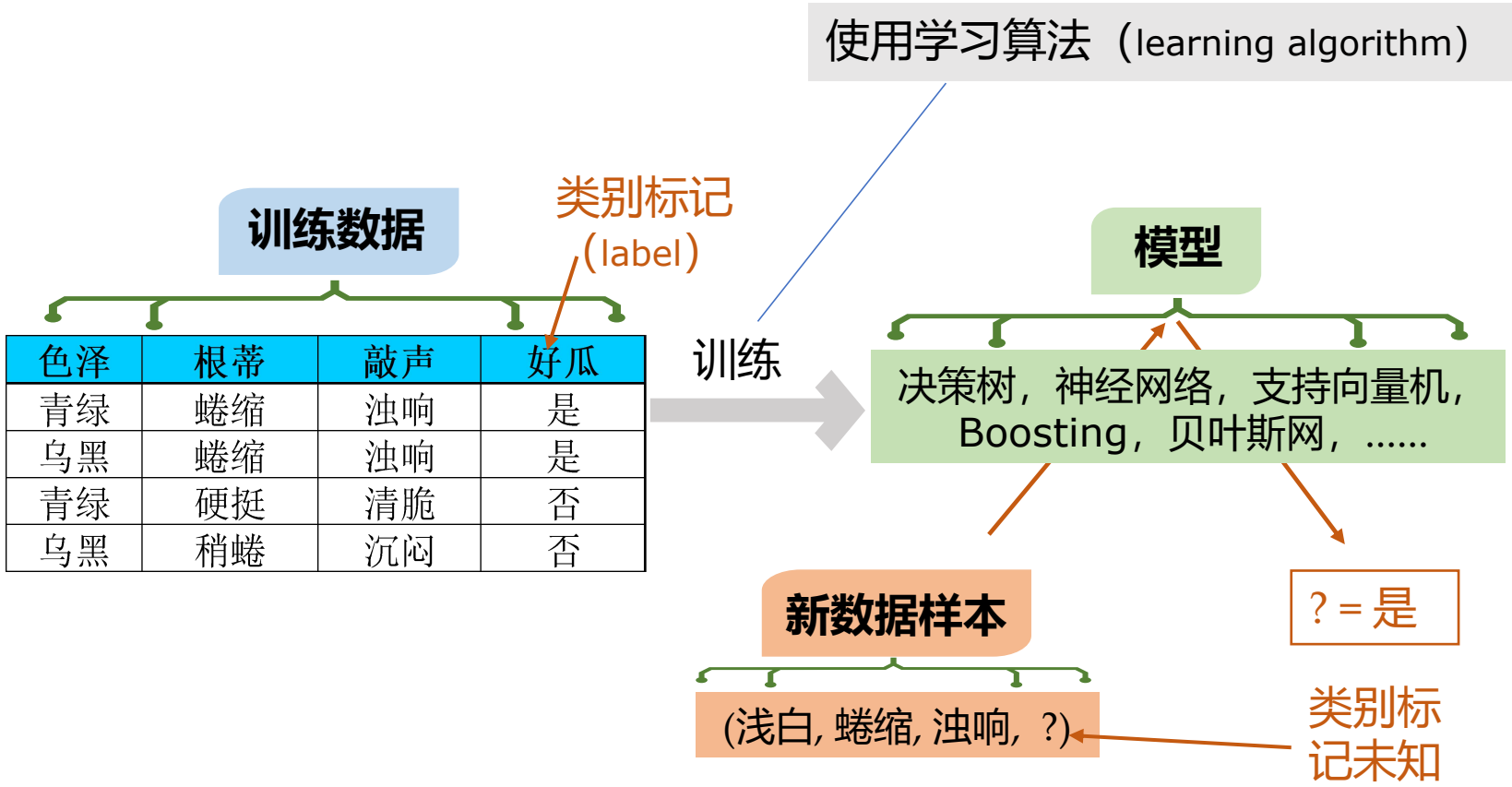
大数据时代

大数据 ≠ 大价值

智能
数据分析

机器学习

典型的机器学习过程


使用学习算法（learning algorithm）
训练数据
类别标记
(label)
色泽 根蒂 敲声 好瓜
青绿 蜷缩 浊响 是
乌黑 蜷缩 浊响 是
青绿 硬挺 清脆 否
乌黑 稍蜷 沉闷 否
训练
模型
决策树，神经网络，支持向量机，
Boosting，贝叶斯网，……
新数据样本
(浅白, 蜷缩, 浊响，?)
? = 是
类别标记未知

机器学习有坚实的理论基础

Leslie Valiant
(莱斯利•维利昂特)
(1949-)
2010年图灵奖

计算学习理论

Computational learning theory

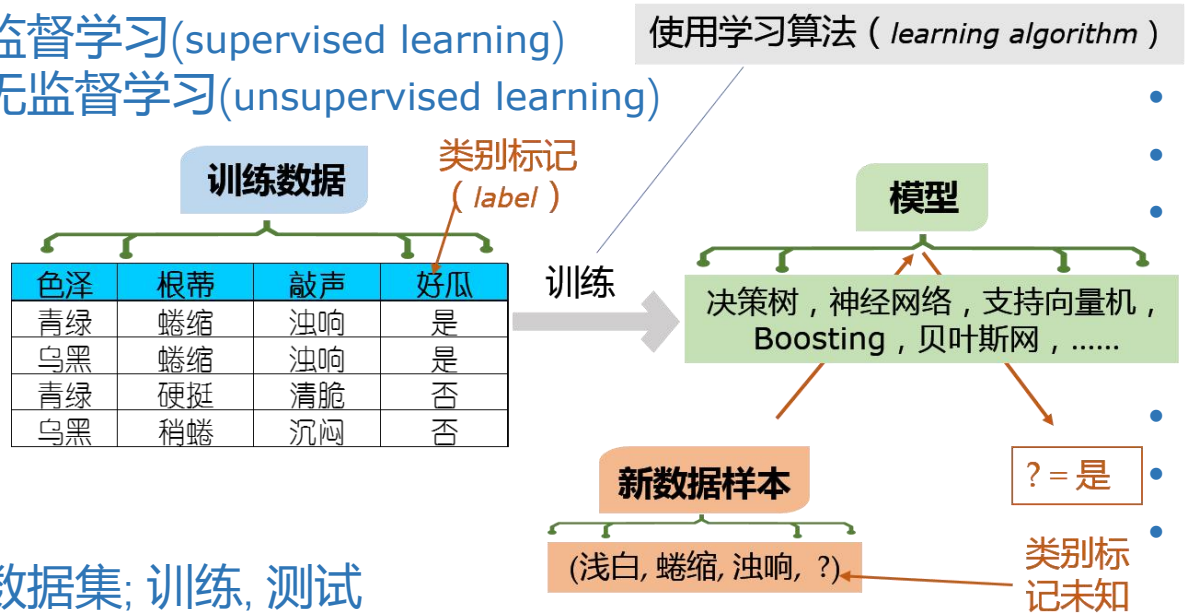
最重要的理论模型:

PAC (Probably Approximately Correct，概率近似正确) learning model [Valiant, 1984]

$$P(|f(\boldsymbol{x}) - y| \leq \epsilon) \geq 1 - \delta$$

基本术语

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)



- 假设(hypothesis)
- 真相(ground-truth)
- 学习器(learner)

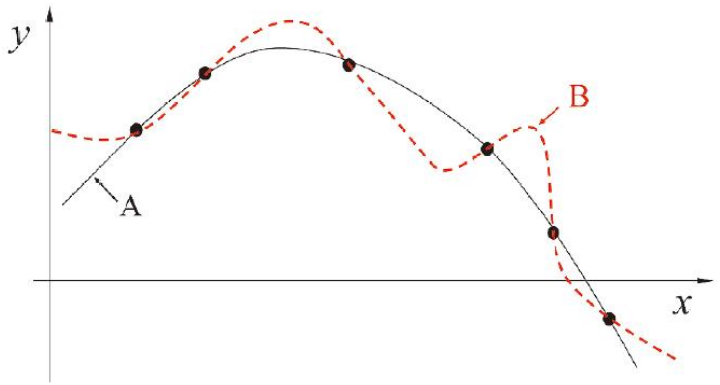
- 分类, 回归
- 二分类, 多分类
- 正类, 反类

- 数据集; 训练, 测试
- 示例(instance), 样例(example)
- 样本(sample)
- 属性(attribute), 特征(feature); 属性值
- 属性空间, 样本空间, 输入空间
- 特征向量(feature vector)
- 标记空间, 输出空间

- 未见样本(unseen instance)
- 未知 “分布”
- 独立同分布(i.i.d.)
- 泛化(generalization)

归纳偏好
(Inductive Bias)

机器学习算法在学习过程中对某种类型假设的偏好



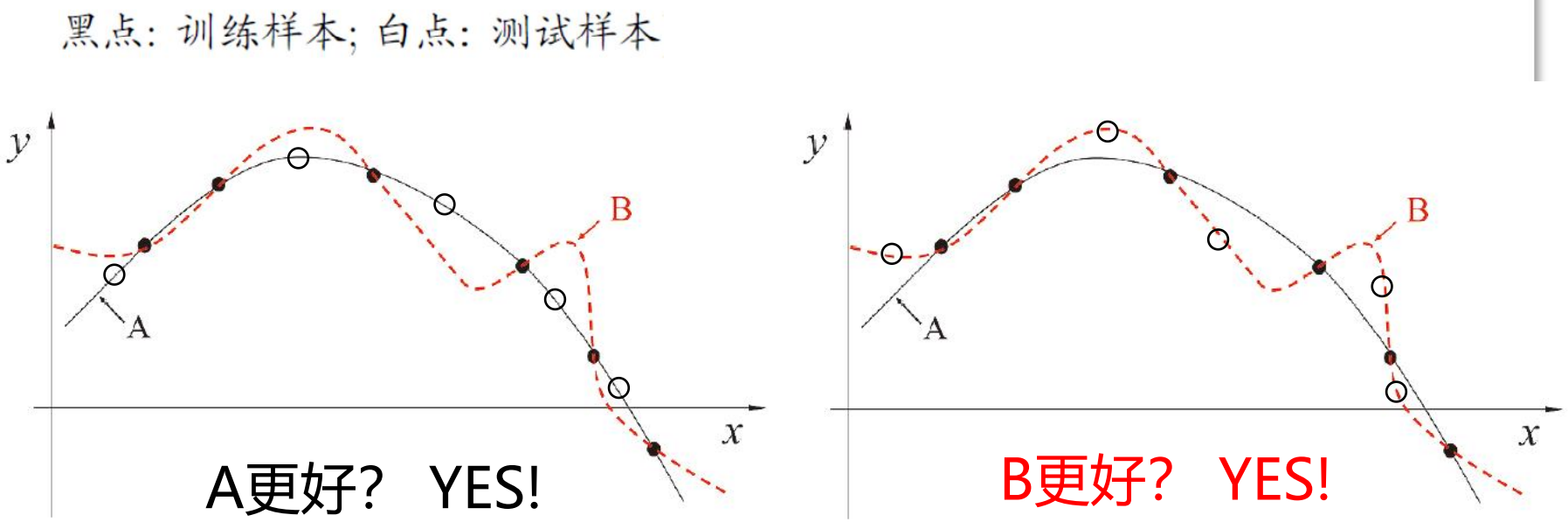
A更好?
B更好?

一般原则:
奥卡姆剃刀
(Ocam´s Razor)

任何一个有效的机器学习算法必有其偏好

学习算法的归纳偏好是否与问题本身匹配,
大多数时候直接决定了算法能否取得好的性能!

哪个算法更好?



没有免费的午餐!

NFL定理：一个算法 $\mathfrak{L}_a$ 若在某些问题上比另一个算法 $\mathfrak{L}_b$ 好，必存在另一些问题 $\mathfrak{L}_b$ 比 $\mathfrak{L}_a$ 好

NFL定理的寓意

NFL定理的重要前提：

所有“问题”出现的机会相同、或所有问题同等重要

实际情形并非如此；我们通常只关注自己正在试图解决的问题

脱离具体问题，空泛地谈论“什么学习算法更好”毫无意义！

具体问题，具体分析！

现实机器学习应用中 ……

把机器学习的“十大算法” “二十大算法”都弄熟，逐个试一遍，是否就“止于至善” 了?

NO!

机器学习并非“十大套路” “二十大招数” 的简单堆积

现实任务千变万化，

以有限的“套路”应对无限的“问题” ，焉有不败?

最优方案往往来自：按需设计、度身定制