zhangy-lab

机器学习

从“西瓜书”开始

潘贤润

2024/04/21

第一章

绪 论

1.1 引言
1.2 基本术语
1.3 假设空间
1.4 归纳偏好
1.5 发展历程和应用现状
1.6 课后习题

zhangy-lab

问题1

为什么叫西瓜书?

问题2

什么是机器学习?

问题3

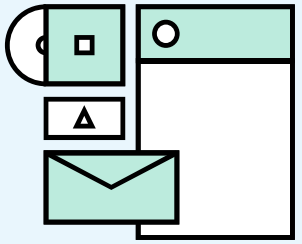
为什么我们要学机器学习，机器学习到底有什么用?

1.1 引言

•什么是机器学习？

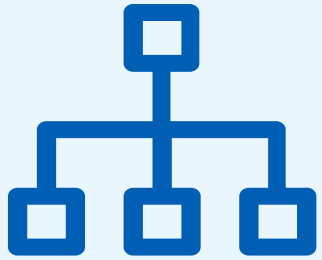
色泽	根蒂	敲声	
青绿	蜷缩	浊响	→ 好瓜
……	……	……	
乌黑	硬挺	沉闷	→ 坏瓜

经验 → 决策



数据

训练/学习 → 学习算法



模型/学习器

决策树是一种经典的监督式学习算法，
其基本原理是通过训练数据集学习出一棵决策树，用于对新数据进行分类或回归预测。

如果一个计算机程序对于某项任务T和某项性能衡量指标 P，其性能随着经验E的提高而提高，则可认为该程序可以从经验E中学习。

计算机程序：模型
P：性能度量：买到好瓜的比例
T：任务：买西瓜
E：经验；数据

1.2 基本术语

(色泽=青绿;根蒂=蜷缩;敲声=浊响), (色泽=乌黑;根蒂=稍蜷;敲声=沉闷), (色泽=浅自;根蒂=硬挺;敲声=清脆).....

取值为

属性/特征

属性值

样本/示例
/特征向量

数据集

列向量

行向量

属性空间/
样本空间/
输入空间

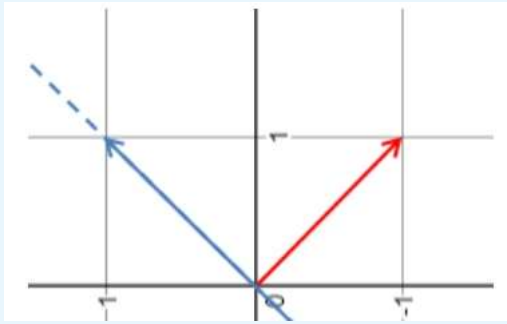
训练样本

测试样本

模型

"假设"

潜在规律
"真相"



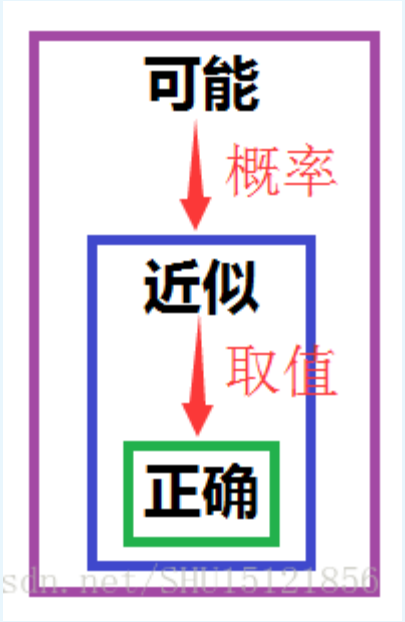
数据集通常用集合来表示,
令集合 D = {x1, x2, ..., xm} 表示包含 m 个样本的数据集,
假设此数据集中的每个样本都含有 d 个特征,
则第 i 个样本的数学表示为 d 维向量:xi = (xi1; xi2; ...; xid),
其中 xij 表示第i个样本 在第 j 个属性上的取值。

1.2 基本术语

什么是机器学习?

PAC（Probably Approximate Correct）

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



(色泽=青绿;根蒂=蜷缩;敲声=浊响)，(好瓜)

样例(xi, yi)　　标记空间/输出空间

标记

任务

有无标记

离散型　　连续型

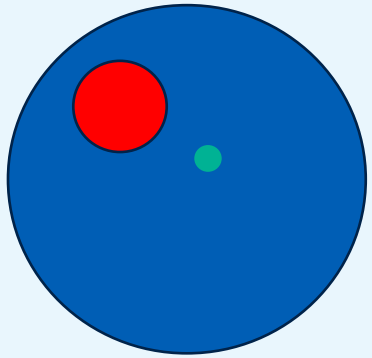
监督学习　　无监督学习

二分类　　多分类　　回归

正类1　　反类0

聚类/降维

分布



泛化能力

数据决定模型的上限，而算法则是让模型无限逼近上限

1.3 假设空间

假设空间部分如下，
1 色泽 = *，根蒂 = *，敲声 = *
2 色泽 = 青绿，根蒂 = *，敲声 = *
3 色泽 = 乌黑，根蒂 = *，敲声 = *
4 色泽 = *，根蒂 = 蜷缩，敲声 = *
5 色泽 = *，根蒂 = 硬挺，敲声 = *
…
36 色泽 = 乌黑，根蒂 = 稍蜷，敲声 = *
37 ∅

归纳：特殊到一般
演绎：一般到特殊
→ 归纳学习

广义
狭义：概念学习

- 概念学习最基本的就是布尔概念学习
  - 从有关某个布尔函数的输入输出的训练集中，推测出这个布尔函数。
- 概念可被看作是一个布尔函数

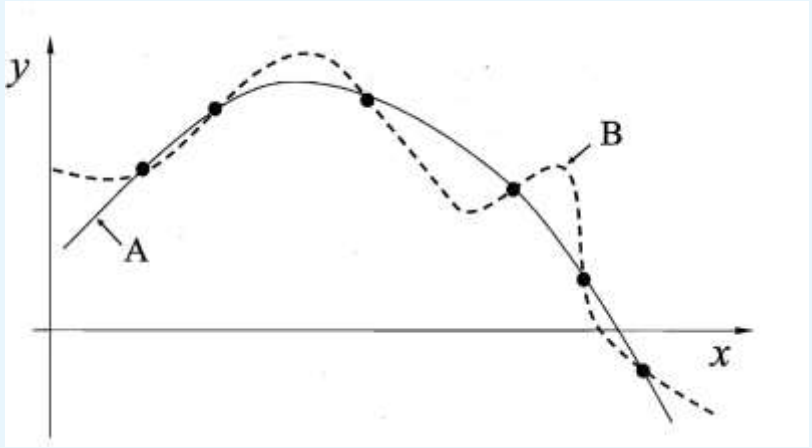
中国人 美国人 日本人
是 否 是

是亚洲人嘛?

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

- 假设空间
  - 是一个函数空间，即由函数所构成的集合
  - 假设空间由形如 “(色泽=？) ∧ (根蒂=？) ∧ (敲声=？)” 的所有假设组成。如果属性色泽、根蒂、敲声分别有2 、3 、3种可能取值，还要考虑到一种属性可能无论取什么值都合适（用通配符*表示），另外有一种情况就是好瓜这个概念根本不成立（用∅表示），则假设空间大小为 （2 + 1）×（3 + 1）×（3 + 1）+ 1 = 49
- 版本空间就是与所学习概念一致的所有假设所构成的集合。从假设空间中删除与正例不一致的假设、和与反例一致的假设，最终将会获得与训练集一致（即对所有训练样本能够进行正确判断）的假设

1.4 归纳偏好



- 奥卡姆剃刀定律

  若有多个假设与观察一致，则选最简单的那个

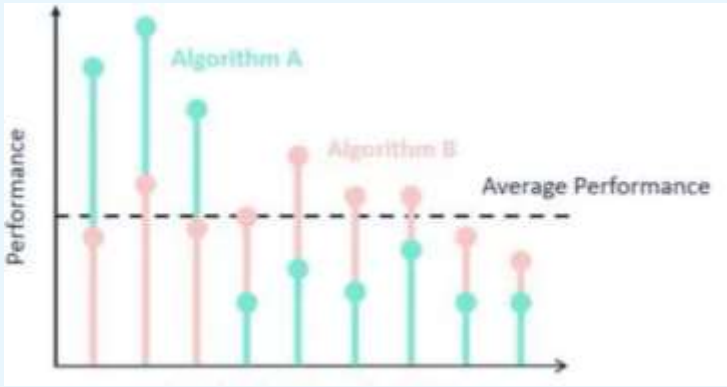
- No Free Lunch Theorem (NFL)

  算法的期望性能是相同的

  要针对具体的问题具体地分析

y1=ax^2+bx+c

y2=ax^3+c



$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\,P(h\mid X,\mathfrak{L}_a)\ ,$$

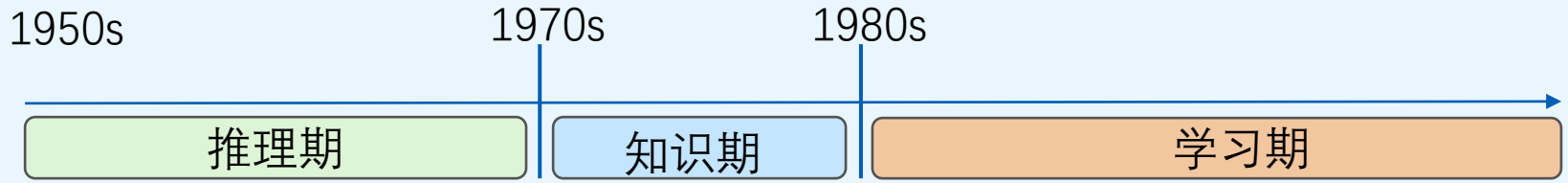
$$\begin{aligned}\sum_f E_{ote}(\mathfrak{L}_a|X,f)&=\sum_f\sum_h\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\,P(h\mid X,\mathfrak{L}_a)\\&=\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h\mid X,\mathfrak{L}_a)\sum_f\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\\&=\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h\mid X,\mathfrak{L}_a)\frac{1}{2}2^{|\mathcal{X}|}\\&=\frac{1}{2}2^{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h\mid X,\mathfrak{L}_a)\end{aligned}$$

$$=2^{|\mathcal{X}|-1}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\cdot 1\ . \tag{1.2}$$

式(1.2)显示出，总误差竟然与学习算法无关！对于任意两个学习算法 $\mathfrak{L}_a$ 和 $\mathfrak{L}_b$，我们都有

$$\sum_f E_{ote}(\mathfrak{L}_a|X,f)=\sum_f E_{ote}(\mathfrak{L}_b|X,f)\ , \tag{1.3}$$

- “推理” ——“知识”——“学习”



- 数据分析
- 生物信息学：生命现象——规律发现
- 数据科学：理论+实验+计算
- 数据挖掘：数据库+机器学习
- 天气预报，图片搜索……

1.6 课后习题

查漏补缺

用合取联结词“∧(并且)”将两个或两个以上的命题联结起来而形成的命题形式

若干简单合取式的析取“∨ (或)”

1.1 表 1.1 中若只包含编号为 1 和 4 的两个样例, 试给出相应的版本空间.

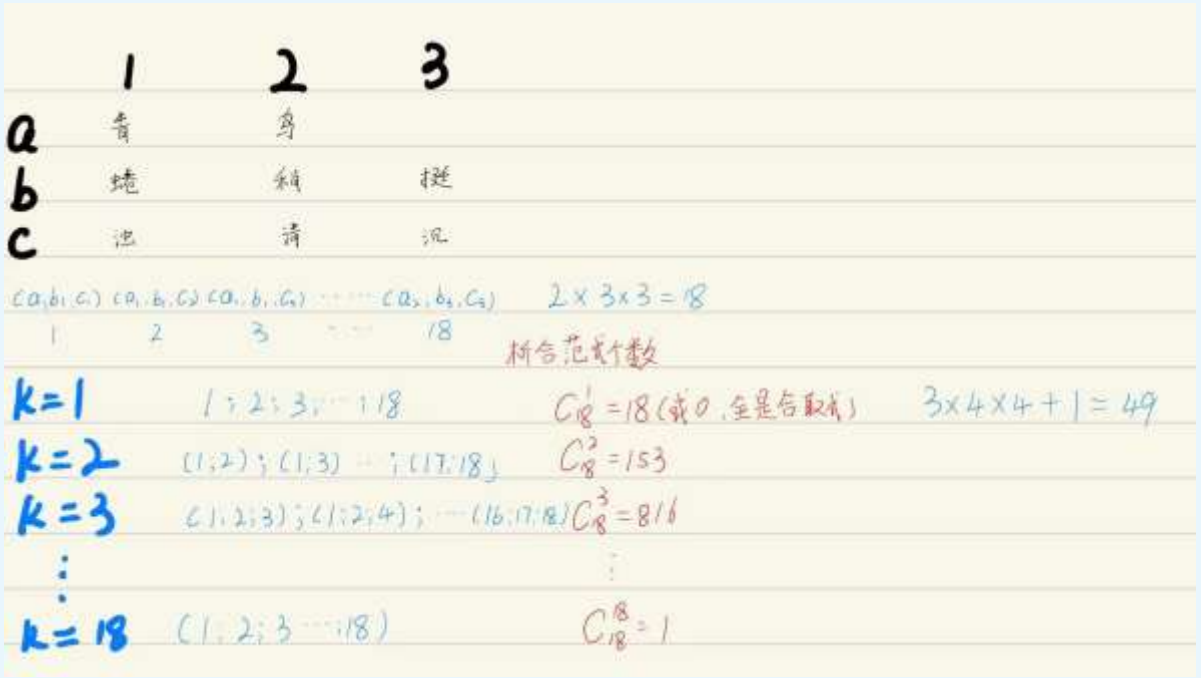
1.2 与使用单个合取式来进行假设表示相比, 使用“析合范式”将使得假设空间具有更强的表示能力. 例如

$$\text{好瓜} \leftrightarrow \big((\text{色泽}=*) \wedge (\text{根蒂}=\text{蜷缩}) \wedge (\text{敲声}=*)\big) \vee \big((\text{色泽}=\text{乌黑}) \wedge (\text{根蒂}=*) \wedge (\text{敲声}=\text{沉闷})\big),$$

会把“(色泽=青绿) ∧ (根蒂=蜷缩) ∧ (敲声=清脆)”以及“(色泽=乌黑) ∧ (根蒂=硬挺) ∧ (敲声=沉闷)”都分类为“好瓜”. 若使用最多包含 k 个合取式的析合范式来表达表 1.1 西瓜分类问题的假设空间, 试估算共有多少种可能的假设.

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

可能的假设有 $C_{18}^0 + C_{18}^1 + \cdots + C_{18}^{18} = 2^{18}$ 种，这是可能假设数的上限。



1.6 课后习题

查漏补缺

1.3 若数据包含噪声，则假设空间中有可能不存在与所有训练样本都一致的假设. 在此情形下，试设计一种归纳偏好用于假设选择.

1.4* 本章 1.4 节在论述“没有免费的午餐”定理时，默认使用了“分类错误率”作为性能度量来对分类器进行评估. 若换用其他性能度量 ℓ，则式(1.1)将改为

$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\ell\left(h\left(\boldsymbol{x}\right),f\left(\boldsymbol{x}\right)\right)P\left(h\mid X,\mathfrak{L}_a\right)\ ,$$

试证明“没有免费的午餐定理”仍成立.

噪声数据（Noisy Data）就是无意义的数据

正则化是一种修改学习算法的方式，旨在减少模型的泛化误差。正则化的基本思想是在损失函数中增加一项额外的项，用于惩罚模型的复杂度。简单而不容易捕捉到噪声的模型假设有助于减少过拟合，增加泛化能力。

一种常见的正则化方法是L1或L2正则化，它们通过在损失函数中增加正则化项以限制模型参数的大小

$$\begin{aligned}\sum_f E_{\text{ote}}\left(\mathfrak{L}_a\mid X,f\right)&=\sum_f\sum_h\sum_{x\in\mathcal{X}-X}p(x)\cdot\ell(h(x),f(x))\cdot p\left(h\mid X,\mathfrak{L}_a\right)\\&=\sum_{x\in\mathcal{X}-X}p(x)\sum_h p\left(h\mid X,\mathfrak{L}_a\right)\sum_f\ell(h(x),f(x))\\&=\sum_{x\in\mathcal{X}-X}p(x)\sum_h p\left(h\mid X,\mathfrak{L}_a\right)\cdot 2^{|x|-1}\left(\ell(h(x)=f(x))+\ell(h(x)\neq f(x))\right)\\&=2|x|-1\cdot A\sum_{x\in\mathcal{X}-X}p(x)\cdot 1\end{aligned}$$

要求性能度量 ℓ 满足$\ell(h(x)=f(x))+\ell(h(x)\neq f(x))=A$为常数，对于二分类问题，即要求 $\ell(0,0)+\ell(0,1)=\ell(1,1)+\ell(1,0)=A$.

西瓜书

为什么叫西瓜书?

本书贯穿以西瓜为例，一则因为瓜果中笔者尤喜西瓜，二则因为西瓜在笔者所生活的区域有个有趣的蕴义. 朋友小聚、请客吃饭，菜已全而主未知，或馔未齐而人待走，都挺尴尬. 于是聪明人发明了“潜规则”：席终上西瓜. 无论整盘抑或小碟，宾主见瓜至，则心领神会准备起身，皆大欢喜. 久而久之，无论菜肴价格贵贱、场所雅鄙，宴必有西瓜. 若将宴席比作(未来)应用系统，菜肴比作所涉技术，则机器学习好似那必有的西瓜，它可能不是最“高大上”的，但却是离不了的、没用上总觉得不甘心的.

感谢倾听

See you next.