中国科学技术大学

University of Science and Technology of China

人工智能讲义

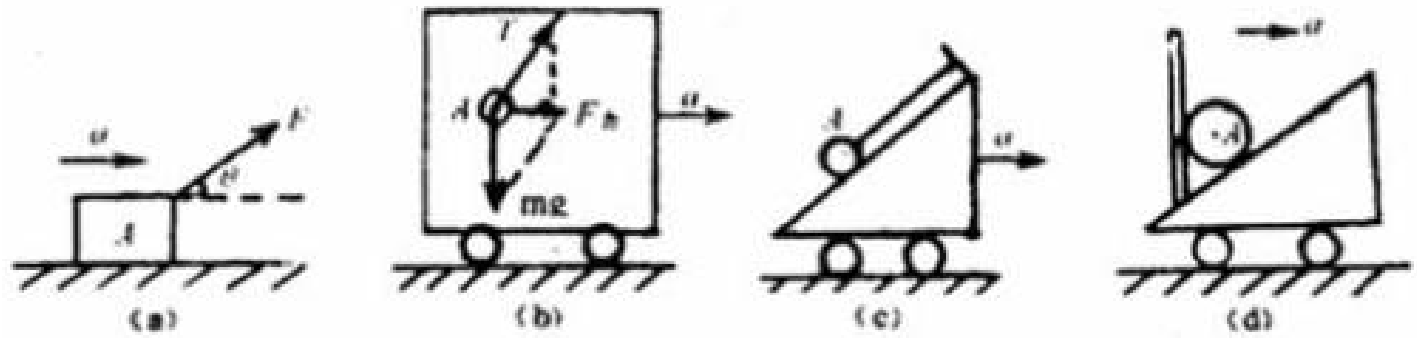
机器学习概念

March 27, 2018

Outline

1. 什么是机器学习
2. h 的评估
3. h 的获得

牛顿力学定理的发现



牛顿力学：力、质量、速度的函数关系

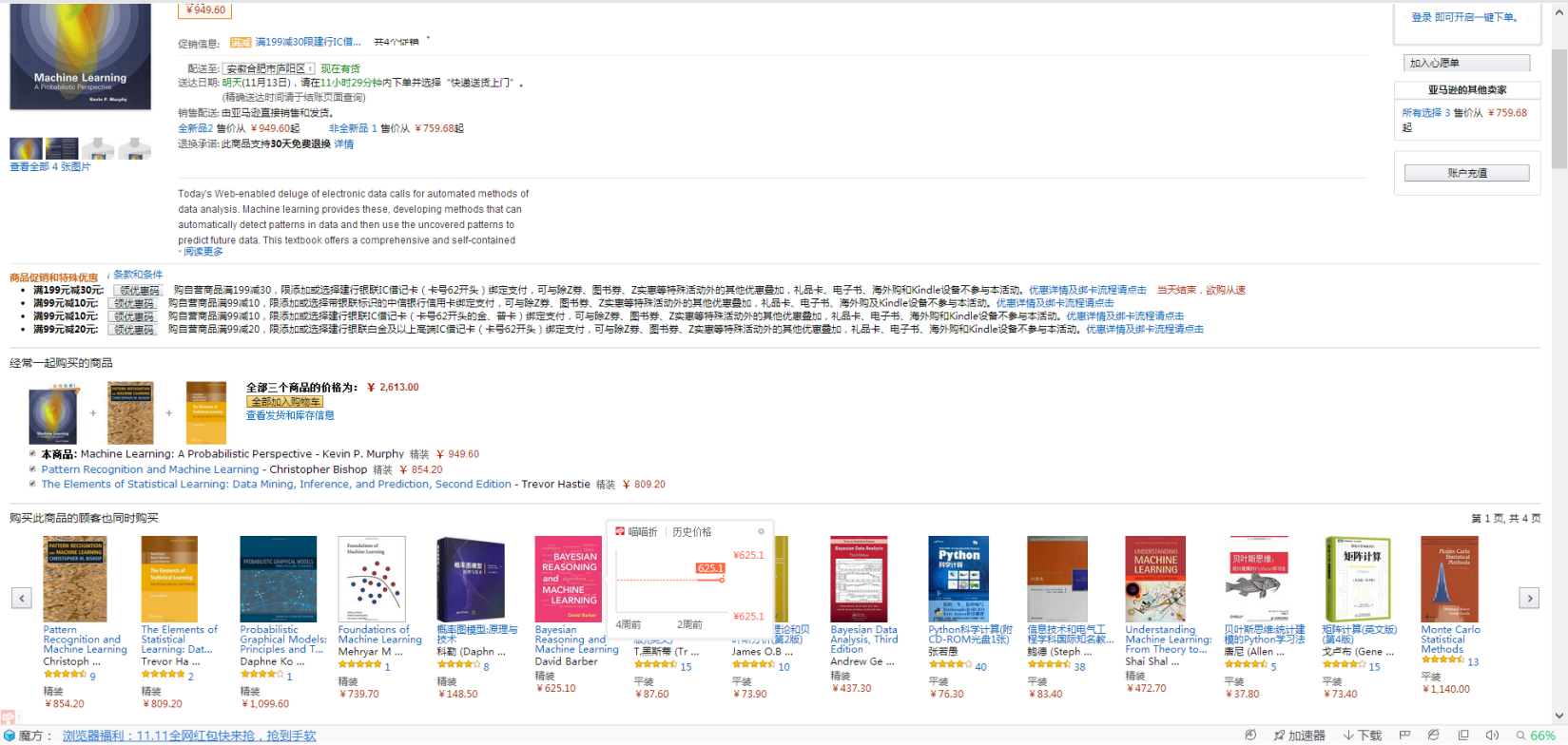
- 如何实验发现的?
- 收集实验数据，填入“表格”(不完整的函数关系)
- 假设函数关系，验证函数关系
- 发现的函数关系是“知识”，规律，是压缩实验数据表格的“压缩方法”
- 函数、知识和函数描述的复杂性

无人自动驾驶

自动驾驶技术

- 面对各种不同情况/路况，自动选择驾驶策略（速度/方向）等；路径规划问题的在线版本；
- 路况无法穷举，状态数目近乎无穷，最佳应对如何实现？

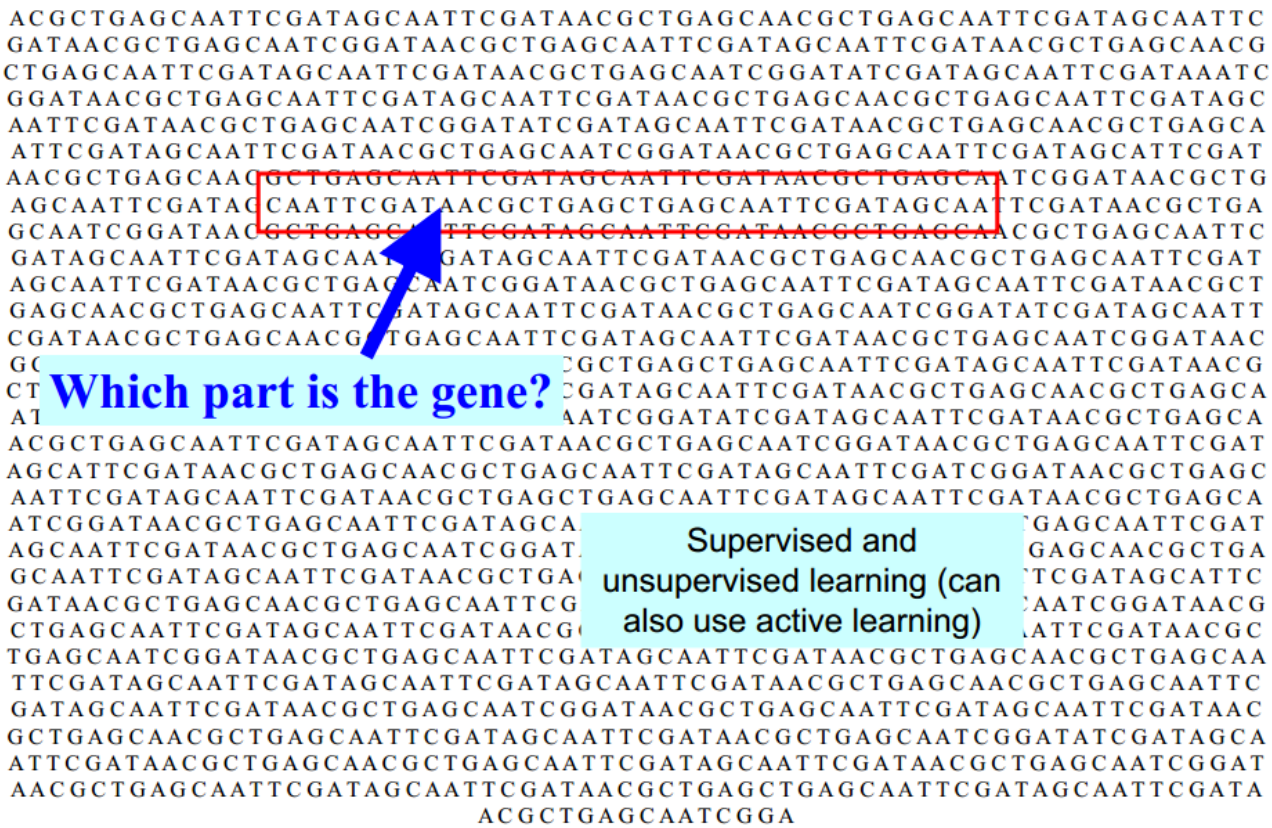
推荐系统



推荐系统：依据用户的习惯和爱好，给出商品的推荐品

- 将所有的商品和用户对，施加一个判别：“喜欢”和“不喜欢”
- 如何在很少的已知数据（用户喜欢和不喜欢某些商品）条件下，实现对极大数量的用户和商品对的判别？规律？知识？

Biology



生物学：在 DNA 链/字符串中找出一个不连续的片段，判定它是否是基因

- 列表枚举所有可能的基因，然后可能性太多，能获得的表格中的数据太少；
- 专业领域内的知识 + 数据分析技术手段；生物学 + 机器学习 $\Longrightarrow$ 生物信息学；机器学习称为通用的“科学发现”手段。

NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

So far, NELL has accumulated over 50 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 2,628,200 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.

Recently-Learned Facts twitter

Refresh

instance	iteration	date learned	confidence
biker_rings is a personal care product	955	20-oct-2015	93.6
drafter is a job position	955	20-oct-2015	97.6
obrien_county_cpc_administrator is a kind of office held by a politician	955	20-oct-2015	91.6
johnny_hawksworth is a musician	955	20-oct-2015	99.1
key_interest_rate is an arachnid	955	20-oct-2015	100.0
weekly_standard is a company that has an office in the city new_york	955	20-oct-2015	93.8
wwor is a TV station in the city new_york	959	07-nov-2015	100.0
cisco has acquired linksys	955	20-oct-2015	100.0
flowers is an agricultural product produced in austria	958	03-nov-2015	100.0
newsweek is a company in the economic sector of news	955	20-oct-2015	100.0

语言：句子是单词的排列

- 自动阅读或意思识别，可以用表格将所有的句子都枚举出来（有限）
- 但是能否压缩描述？找到规律或知识？

西洋双陆棋

设计一个程序打败人类

- 从失败中逐步学会调整策略，现在能够战胜人；
- 设计程序思路：一个映射表，枚举了所有的棋局和最佳应对，在失败中不断调整失败的应对。

机器学习的含义

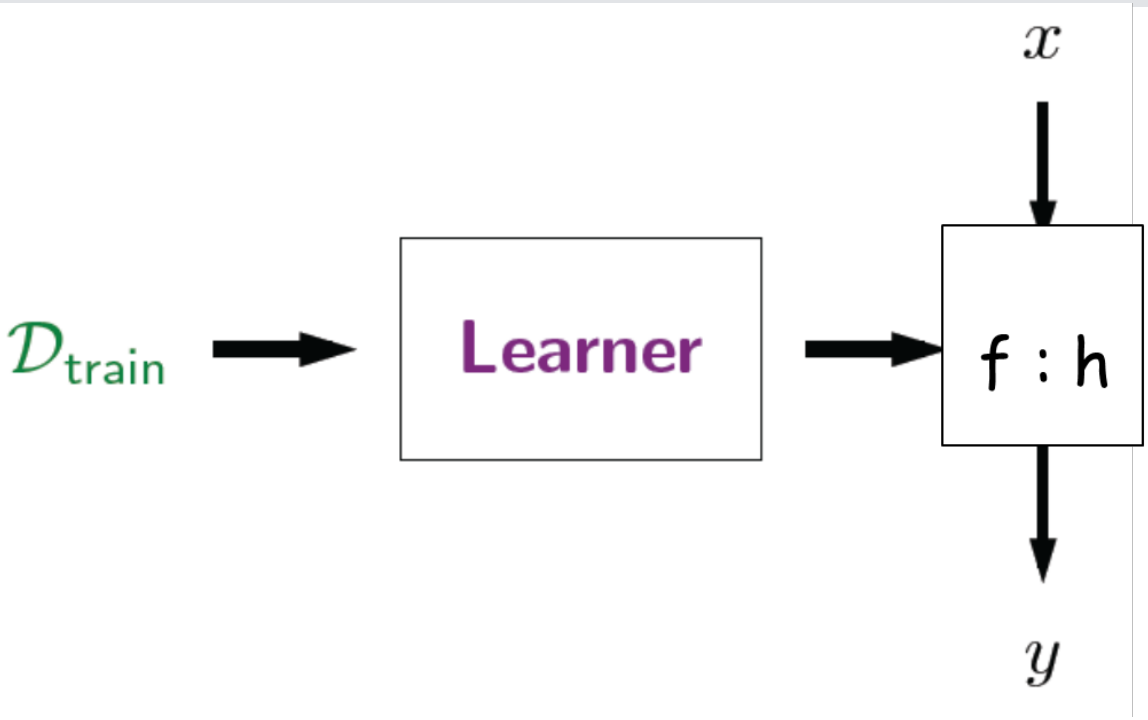
含义 1：获得完整的数据表格

- 现实中，总是能针对某个事物收集到一些数据
- 数据总是能填写到表格中去
- 得到的表格一般都是不完整的，有缺失的
- 机器学习：获得完整的表格。 (表格对应一个函数 f)

含义 2：压缩完整表格的存储空间

- 完整表格的行数通常是列数的指数函数
- 当列数较大时，在计算机内存储完整的表格通常是不可能的
- 压缩表示完整表格，即用描述长度较小的函数来表示描述长度最大（描述复杂性最高）的完整表格；
- 机器学习：获得描述长度较小的函数 f。

机器学习定义



机器学习：寻找 f 近似值 h 的过程

- D_{train}：训练数据集
- $Learner$: 学习器，从训练数据集中归纳出 h 的机器/程序/算法等
- f: 函数，$y = f(x)$，可以是现实世界的一个过程/机制/方法等的抽象
- h: 函数 f 的近似值，模型/假设
- x: 输入变量/自变量
- y: 输出变量/响应/因变量

假设输入是一个向量，不妨设 $X = (X_1, X_2, \ldots, X_n)$，第 i 个分量的值域大小记为 $|X_i|$, 其第 j 个取值为 v_{ij} ，则如下表格完全表示一个函数：

X_1	X_2	$\ldots$	X_n	$f(X_1, X_2, \ldots, X_n)$
v_{11}	v_{21}	$\ldots$	v_{n1}	y_1
v_{11}	v_{21}	$\ldots$	v_{n2}	y_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$v_{1\|X_1\|}$	$v_{2\|X_2\|}$	$\ldots$	$v_{n\|X_n\|}$	$y_{\|X_1\|\|X_2\|\ldots\|X_n\|}$

Table: n 元函数的表示

函数 f 的近似值 h

机器学习的对象与结果：h

- 又称“假设”或“模型”，包括两层含义：
  - 预定义在 h 中的“知识”，用于压缩函数的描述长度；
  - 对所有输入都定义了输出/响应

建模/学习：模型选择与训练

- 确定 h 的过程，更精确地讲，分为两步：
- 模型选择：
  - 枚举模型/表格模型，用完整表格，最大的复杂性来定义 h，一般认为此时在模型中没有预定义任何“知识”，也因此它是最通用的模型；
  - 用对数据的先验知识，假定数据输出和输入之间呈线性关系；此时预定义的知识就是“输入”和“输出”之间体现为线性关系。
- 训练：确定模型的参数，比如把表格不全；或确定线性表达式的系数等。

如何评价 h 的优劣

准确性：h 和 f 之间的差异

- h 近似 f，近似的准确程度是多少？这是最关键的评估标准

复杂性：h 的描述长度是多少？

- h 的描述长度通常和计算的时空代价相关，因此我们要确保 h 的描述长度在合理、有效的范围内。

完备性：h 是否对 f 的每个输入都定义了一个响应？

- 机器学习一般情形下要求结果具备完备性；而数据挖掘一般没有此要求。

h 准确性的定义

- 绝对误差：$err_1(h,f)=\sum_{i=1}^{|X|}|f(x_i)-h(x_i)|$，适用于 y 取连续值的情形
- 平方误差：$err_2(h,f)=\sum_{i=1}^{|X|}(f(x_i)-h(x_i))^2$，适用于 y 取连续值的情形
- 误差计数：$err_0(h,f)=\sum_{i=1}^{|X|}1[f(x_i)\neq h(x_i)]$，适用于 y 取离散值的情形

准确性计算方法及存在问题

- 对所有的输入，比较 f 和 h 输出的差异，然后求和；
- 问题 1：h 的输入 x 的值域太大，通常时间上不允许遍历 x 的值域，所以精确计算准确性存在困难。
- 问题 2：上述误差和 x 的值域大小相关，通常将上述误差除以 x 值域大小，得到误差均值，并用之来评估准确性。

h 的准确性

准确性的近似计算方法的思想

- 从完整映射表中随机抽样若干行，检测在这些行中 h 的错误率，用该错误率近似真实的错误率。

实际应用中

- 从已知 y 真实值的行中，选择保留一部分行，不参与训练（寻找 h），这部分被保留的行，被称之为“测试数据集”，用测试数据集的错误率，近似估计 h 的错误率
- 已知数据集/真实数据划分为：
  - 训练数据集 D_{train}
  - 测试数据集 D_{test}

准确性的近似值

- $err_2(h, f) = \sum_{x \in D_{test}} (f(x) - h(x))^2$，类似可重新定义 err_1, err_0

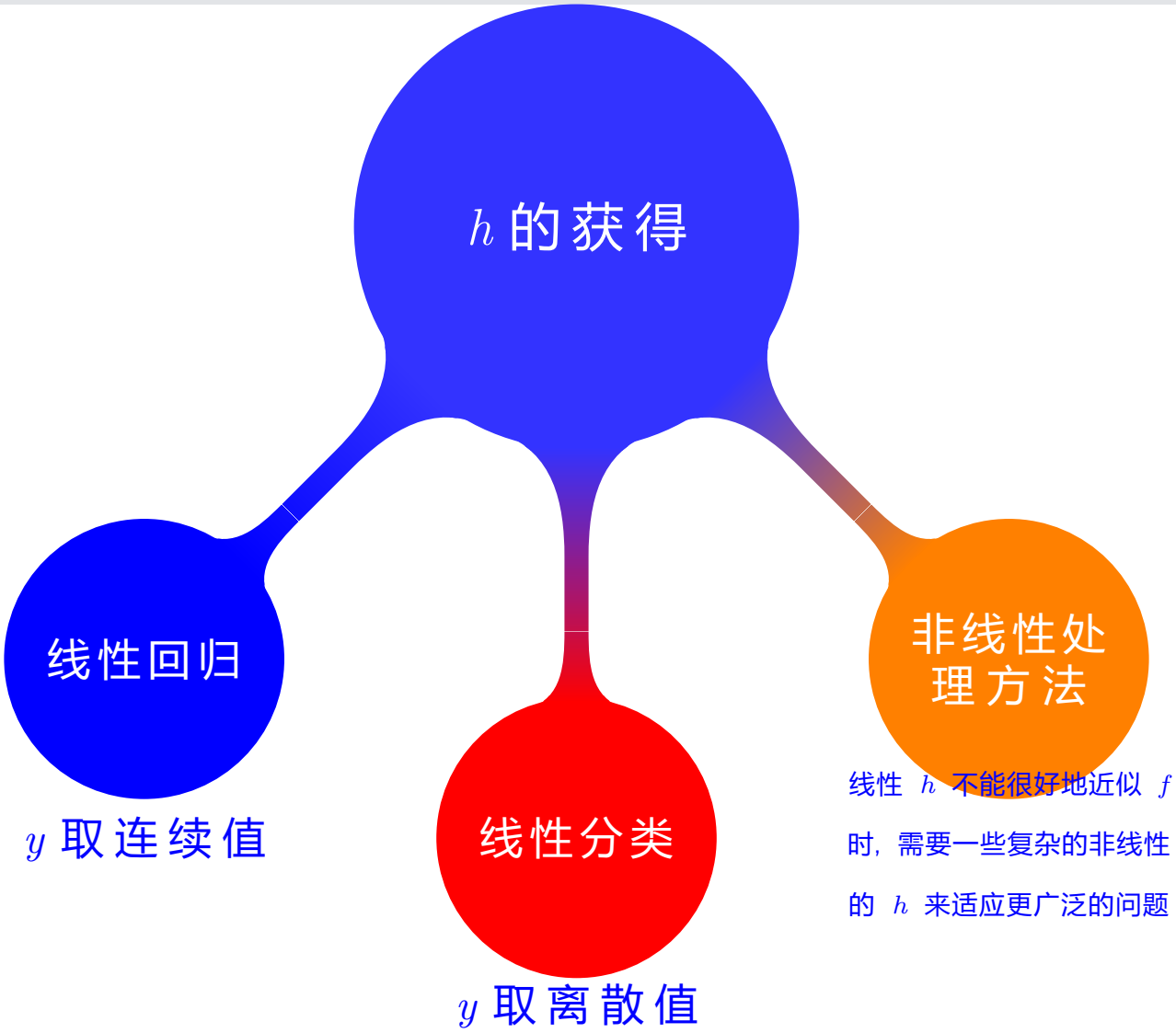
数学上最简单的表达式为线性

- $y = ax + b$

机器学习中最重要的、最简单的 h 也是线性模型

- 机器学习中用线性模型来近似 $y = f(x)$
- $y = Wx$，x 是标量
- $y = \mathbf{W} \cdot \mathbf{x} = \Sigma_x wx$，$\mathbf{x}$ 是向量
- 上面的 $W, \mathbf{W}$ 是线性表达式的系数，$\cdot$ 表示内积/点积

我们从线性 h 开始了解机器学习


h 的获得
线性回归
y 取连续值
线性分类
y 取离散值
非线性处理方法
线性 h 不能很好地近似 f 时，需要一些复杂的非线性的 h 来适应更广泛的问题