



机器学习理论与工程实践 绪论

徐文星



导言

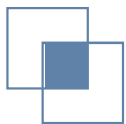


机器学习是人工智能的核心研究领域之一，其研究动机是为了让计算机系统具有人的学习能力以便实现人工智能。

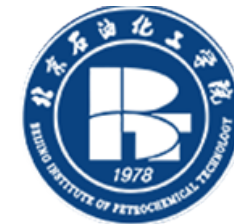
本课程将以思维和方法为目标，首先介绍模型的评估，然后和大家一起讨论学习一批经典而常用的机器学习技术。

希望大家通过本课程的学习，不仅掌握机器学习基本原理、最有效的机器学习技术和实践应用领域，而且能够在理解的基础上选择并利用机器学习的常用算法解决本专业和相关领域的实际问题的能力。更进一步地，我希望你们能够养成科学的思维方式，大胆创新、小心求证。

Wenxing Xu, 2022



目录



- 1 学科定义
- 2 课程要求
- 3 机器学习方法
- 4 应用场景与挑战



机器学习理论及工程实践



学科定义

课程要求

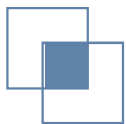
机器学习方法

应用场景与挑战

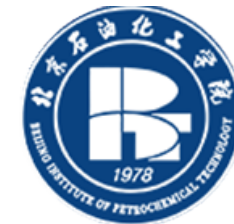


什么是机器学习 ?





什么是机器学习 ?



如何获取女神芳心



百度一下

网页

资讯

视频

图片

知道

文库

贴吧

地图

采购

更多

百度为您找到相关结果约1,160,000个

搜索工具

方法/步骤:

1. 所谓知己知彼,方能百战百胜。想要俘获女神芳心,必需了解女神的喜好,只有投其所好,...
2. 想要搞定女神,首先搞定女神闺蜜。和女神闺蜜统一战线,这样才能事半功倍。这里当然要...
3. 不定时,准备惊喜,并不一定要花很多钱。让女神觉得你是个有诚意,并愿意为她花时间,...
4. 每天在固定的时间点给女神电话、微信问候。让你们的互动成为她生活的的一种习惯。让她...
5. 和女神逛街或过马路的时候,永远走在有车的那边,当车辆比较多时,自然的揽一揽女神的肩,...

查看更多内容...

[如何俘获女神的芳心-百度经验](#)



什么是机器学习



You probably use it dozens of times a day without even knowing it. Each time you do a web search on Google or Bing, that works so well because their machine learning software has figured out how to rank what pages. When Facebook or Apple's photo application recognizes your friends in your pictures, that's also machine learning. Each time you read your email and a spam filter saves you from having to wade through tons of spam, again, that's because your computer has learned to distinguish spam from non-spam email.

Machine learning is a **science of getting computers to learn without being explicitly.**

ML@Coursera



什么是机器学习



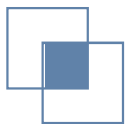
机器学习是从人工智能中产生的一个重要学科分支，是实现智能化的关键。

机器学习 (Machine Learning) 是一门多领域**交叉学科**，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

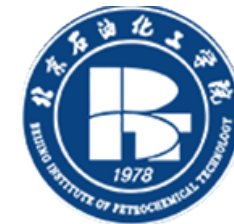
百度百科

Machine learning is the study of **algorithms** and mathematical **models** that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model of **sample data**, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Wikipedia



什么是机器学习



Machine Learning \approx Looking for Function

- Speech Recognition

$$f(\text{audio waveform}) = \text{"How are you"}$$

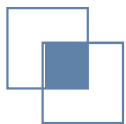
- Image Recognition

$$f(\text{cat image}) = \text{"Cat"}$$

- Playing Go

$$f(\text{Go board state}) = \text{"5-5"}_{(\text{next move})}$$

李宏毅2021/2022春机器学习课程



机器学习概论



学科定义

课程要求

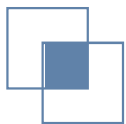
机器学习方法

应用场景与挑战



教材





课程内容



第1章：绪论

第2章：模型评估

第3章：线性学习（Linear learning）

第6章：支持向量机学习（Support vector machine learning）

第4章：决策树学习（Decision tree learning）

第5章：神经网络学习（Neural network learning）

第7章：贝叶斯学习（Bayesian learning）

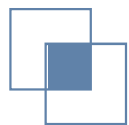
第8章：集成学习（Ensemble）

第9章：聚类（Clustering）

第10章：降维（Dimension reduction）

第13章：半监督学习（Unsupervised learning）
























32
学时

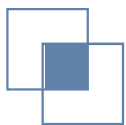


参考推荐的视频资源

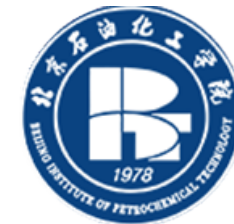


AI-MIT-Patrick H. Winston (2010)




-  1.什么是人工智能.mp4
-  2.推理：目标树与问题求解.mp4
-  3.推理：目标树与基于规则的专家系统.mp4
-  4.搜索：深度优先、爬山、束搜索.mp4
-  5.搜索：最优、分支限界、A-star.mp4
-  6.搜索：博弈、极小化极大、 α - β .mp4
-  7.约束：解释线条图.mp4
-  8.约束：搜索、域缩减.mp4
-  9.约束：视觉对象识别.mp4
-  10.学习介绍、最近邻.mp4
-  11.学习：识别树、无序.mp4
-  11.学习：识别树、无序.mp4
-  12.学习：神经网络、反向传播.mp4
-  13.学习：遗传算法.mp4
-  14.学习：稀疏空间、音韵学.mp4
-  15.学习：相近差错、妥适条件.mp4
-  16.学习：支持向量机.mp4
-  17.学习：boosting算法.mp4
-  18.表示：分类、轨迹、过渡.mp4
-  19.架构：GPS、SOAR、包容架构、心智社会.mp4
-  20.概率推理I.mp4
-  21.概率推理II.mp4
-  22.模型融合、跨通道耦合、课程总结.mp4



参考推荐的视频资源



NN通向智能之路-中科院计算所-陈云霁

-  第1集 人工神经网络基础.mp4
-  第2集 人工神经网络发展现状.mp4
-  第3集 人工神经网络的硬件实现.mp4



浙江大学-研究生机器学习课程-

35.8万 2019-12-01

校园宽带小王子

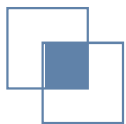


机器学习

中国地质大学（武汉） 蒋良孝、胡成玉

机器学习是人工智能的核心研究领域之一，其研究动机是为了让计算机系统具有人的学习能力以便实现人工智能。本课程将以数据挖掘中的分类任务为例，首先讲解分类模型的评估，然后讲解一...

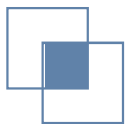
7353人参加 已结束，可查看内容



考核



| | | |
|---|--------------|-----|
| 1 | 课堂签到+讨论+随堂测试 | 10' |
| 2 | 课堂讲演 | 20' |
| 3 | 基本实验 | 20' |
| 4 | 期末考试（开卷） | 50' |



机器学习概论



学科定义

课程要求

机器学习方法

应用场景与挑战



发展历程



★ 推理期（20世纪50-70年代初）

- 认为只要给机器赋予逻辑推理能力，机器就能具有智能
- A. Newell 和 H. Simon 的“逻辑理论家”、“通用问题求解”程序，获得了1975年图灵奖

★ 知识期（20世纪70年代中期）

- 认为要使机器具有智能，就必须设法使机器拥有知识
- E.A. Feigenbaum 作为“知识工程”之父在 1994 年获得了图灵奖

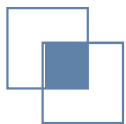
★ 学科形成（20世纪80年代）

20 世纪 80 年代是机器学习成为一个独立学科领域并开始快速发展、各种机器学习技术百花齐放

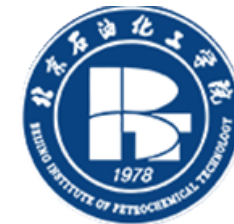
- 1980 年美国卡内基梅隆大学举行第一届机器学习研讨会
- 1990 年《机器学习:风范与方法》出版

★ 繁荣期（20世纪80年代-至今）

- 20 世纪 90 年代后，统计学习方法占主导，代表为SVM
- 2006 至今，大数据分析的需求，神经网络又被重视，成为深度学习理论的基础



机器学习方法



有监督学习 (supervised learning)：从给定的**有标注的训练数据集**中学习出一个函数（模型参数），当新的数据到来时可以根据这个函数预测结果。常见任务包括**分类**与**回归**。

分类：输出是类别标签

Classification: Y is discrete

Y: 年轻人(1), 老年人(-1)

X: x_1 黑头发的比例, 值域 (0, 1);
 x_2 行走速度, 值域 (0, 100) 米/每分钟.

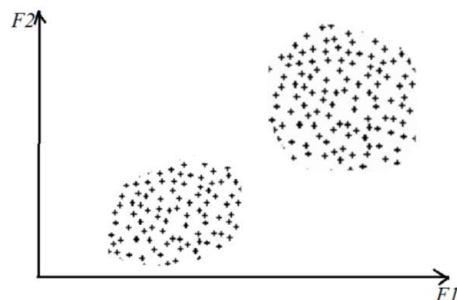
Training Data:

Y=1: (1, 99)、(0.9, 80)、(0.80, 100) ...

Y=-1: (0.2, 30)、(0.5, 50)、(0.4, 30) ...

Test:

X=(0.85, 98), Y=?



回归：输出是实数

Regression: Y is continue

Y: 房屋价钱 (万元), 值域 $Y \geq 0$.

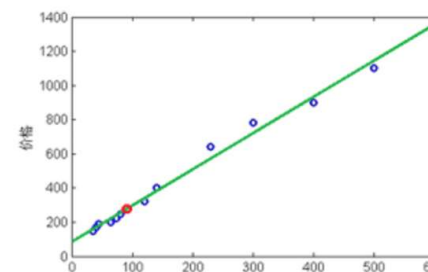
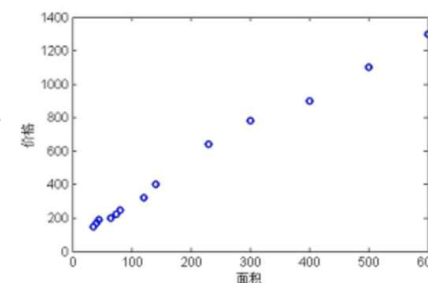
X: x_1 =房屋面积 m^2 .

Training Data:

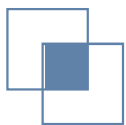
| | |
|-----|------|
| 35 | 150 |
| 40 | 170 |
| 45 | 190 |
| 65 | 200 |
| 74 | 224 |
| 80 | 245 |
| 120 | 320 |
| 140 | 400 |
| 230 | 640 |
| 300 | 780 |
| 400 | 900 |
| 500 | 1100 |
| 600 | 1300 |

Test: X=90

Y=?



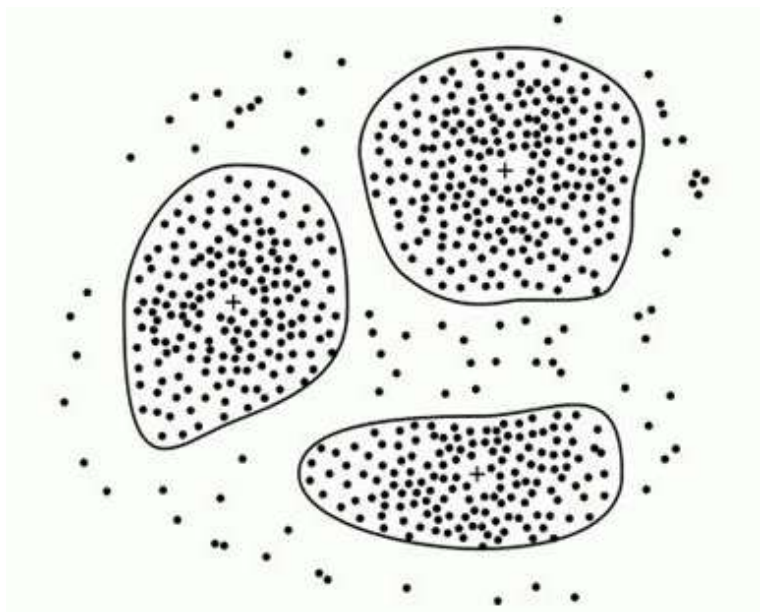
$$y = ax + b$$



机器学习方法



无监督学习 (unsupervised learning) : 没有标注的训练数据集, 需要根据样本间的统计规律对样本集进行分析, 常见任务如**聚类**等。



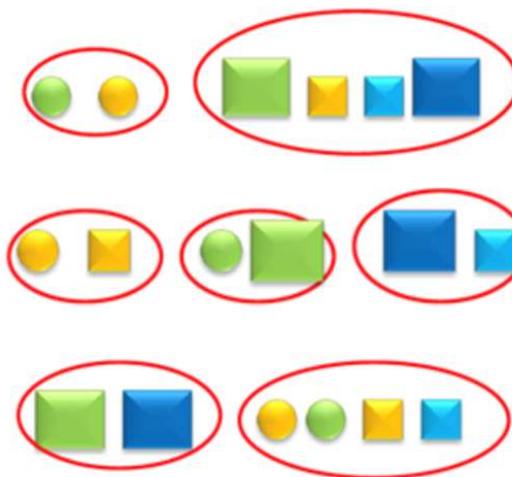
Clustering:

X: (颜色, 形状, 大小)

Data:



For all the data, $Y=?$

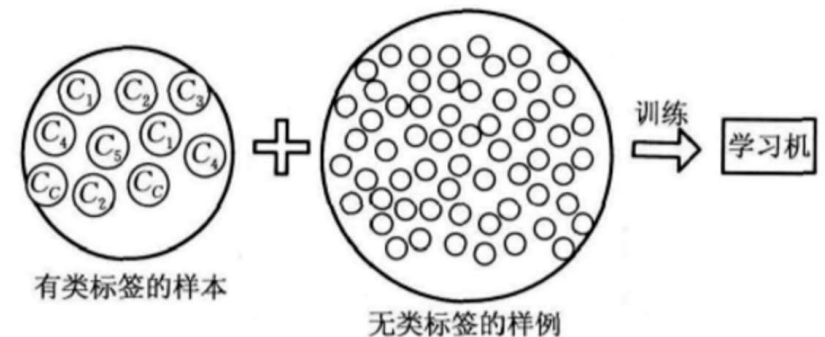




机器学习方法



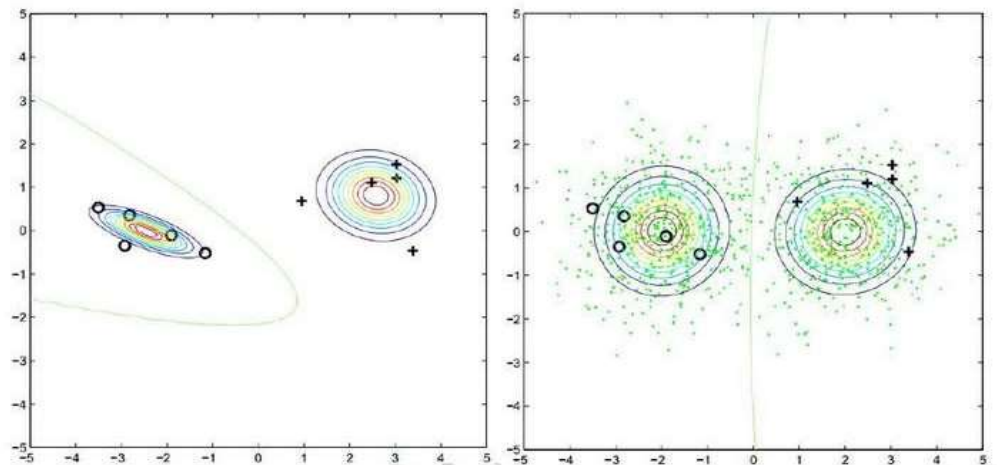
半监督学习 (Semi-supervised learning)：结合 **(少量的) 标注训练数据** 和 **(大量的) 未标注数据** 来进行数据的分类学习。

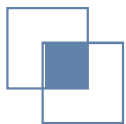


两个基本假设：

- **聚类假设**：处在相同聚类中的样本示例有较大的可能拥有相同的标记。
- **流形假设**：处于一个很小的局部区域内的样本示例具有相似的性质，因此，其标记也应该相似。

※ 相似的样本具有相似的输出





机器学习方法



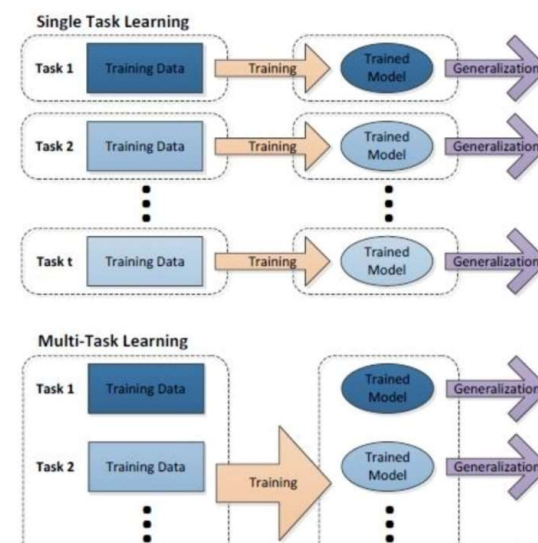
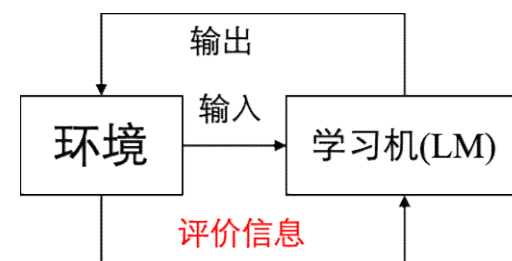
增强学习 (Reinforcement Learning) : 外部环境对输出只给出评价信息而非正确答案, 学习机通过强化受奖励的动作来改善自身的性能。

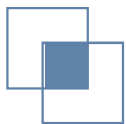
如: **让计算机学着去玩Flappy Bird**

我们不需要设置具体的策略, 比如先飞到上面, 再飞到下面, 我们只是需要给算法定一个“小目标”! 比如当计算机玩的好的时候, 我们就给它一定的奖励, 它玩的不好的时候, 就给它一定的惩罚, 在这个算法框架下, 它就可以越来越好, 超过人类玩家的水平。

多任务学习 (Multi-task Learning) : 把多个相关 (related) 的任务放在一起同时学习。

单任务学习时, 各个任务之间的模型空间 (Trained Model) 是相互独立的, 但现实世界中很多问题不能分解为一个一个独立的子问题, 且这样忽略了问题之间所包含的丰富的关联信息。多任务学习就是为了解决这个问题而诞生的。多个任务之间共享一些因素, 它们可以在学习过程中, 共享它们所学到的信息, 相关联的**多任务学习**比单任务学习具备更好的泛化 (generalization) 效果。





机器学习概论



学科定义

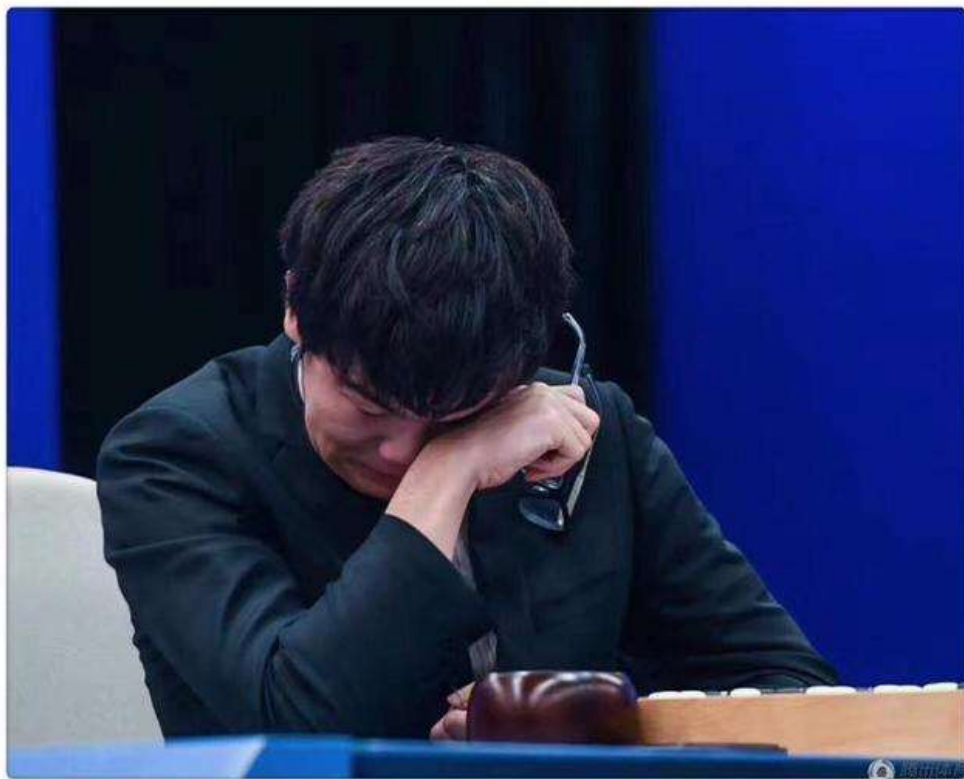
发展历程

机器学习方法

应用场景与挑战



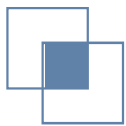
3:0 ! AlphaGo 完胜柯洁



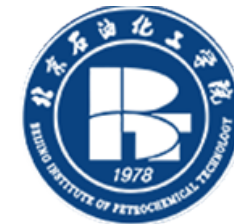
柯洁：中国围棋职业九段棋手，世界排名第一

AlphaGo：Google DeepMind 开发的机器学习围棋程序

AlphaGo使用**蒙特卡罗树**搜索与两个**深度神经网络**相结合的方法，其中一个以估值网络来评估大量的选点，一个以走棋网络来选择落子。

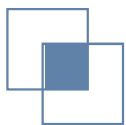


无人驾驶车队亮相2018春晚



百度发布 “**Apollo (阿波罗)**”
软件平台，向汽车行业及自动
驾驶领域提供一套完整的平台。

无人驾驶主要包括三个环节：
感知、**决策**、和控制
核心技术：异步多传感器同
步+深度**数据融合**



机器学习已无处不在

搜索引擎：网页、图片、视频、新闻、学术、地图

信息推荐：新闻、商品、游戏、书籍

图片识别：人像、用品、动物、交通工具

用户分析：社交网络、影评、商品评论

机器翻译、摘要生成.....

生物信息学.....



150x106

相似图片

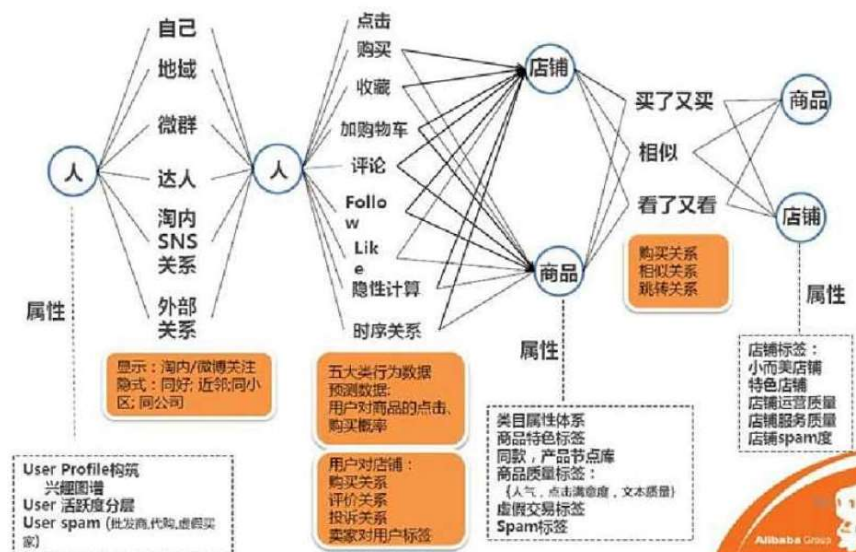


Google的成功，使得Internet搜索引擎成为一个新兴的产业

不仅有众多专营搜索引擎的公司出现（例如专门针对中文搜索的就有慧聪、百度等），而且Microsoft等巨头也开始投入巨资进行研发

Google掘到的第一桶金，来源于其创始人Larry Page和Sergey Brin提出的PageRank算法

机器学习技术正在支撑着各类搜索引擎（尤其是贝叶斯学习技术）





机器学习无所不能?

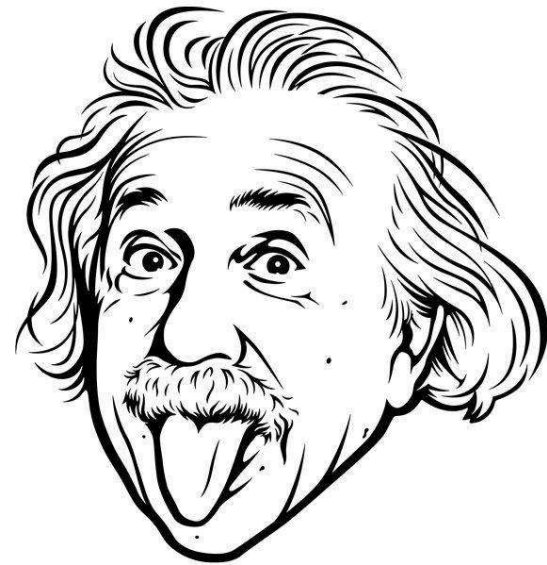


◆**问题思考**: 机器学习是否无所不能?



规则、计算、模式

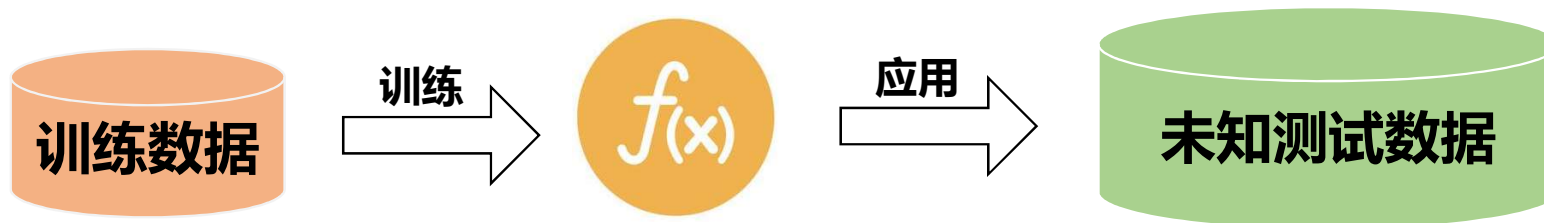
V.S.



思想、创意、情感



机器学习面临的难题与挑战



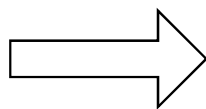
- ◆ **数据稀疏性：** 训练一个模型，需要大量（标注）数据，但是数据往往比较稀疏。
- ◆ **高数量和高质量标注数据需求：** 获取标定数据需要耗费大量人力和财力。而且，人会出错，有主观性。
- ◆ **冷启动问题：** 对于一个新产品，在初期，要面临数据不足的冷启动问题。
- ◆ **泛化能力问题：** 训练数据不能全面、均衡的代表真实数据。



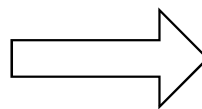
机器学习面临的难题与挑战



模型



策略



算法

- ◆ **模型抽象困难：** 总结归纳实际问题中的数学表示非常困难。
- ◆ **模型评估困难：** 在很多实际问题中，很难形式化的、定量的评估一个模型结果的好坏。
- ◆ **寻找最优解困难：** 要解决的实际问题非常复杂，将其形式化后的目标函数也非常复杂，往往在目前还不存在一个有效的算法能找到目标函数的最优值。



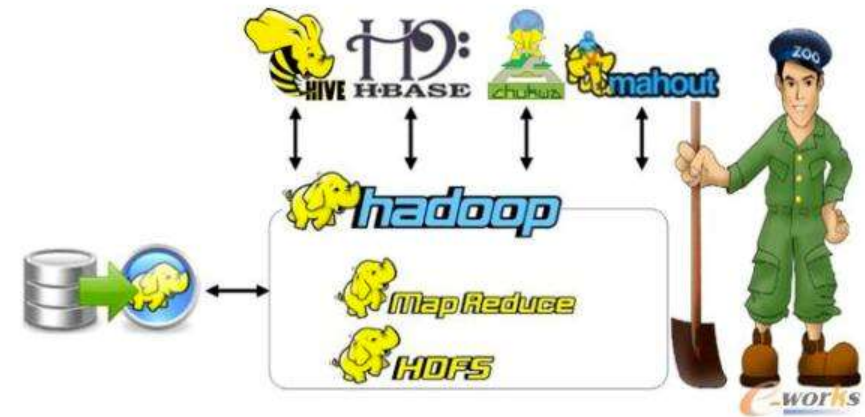
机器学习面临的难题与挑战



◆ **Scalability** 是互联网的核心问题之一。搜索引擎索引的重要网页超过 100 亿: 如果1台机器每秒处理1000 网页, 需要至少100天。

◆ **Quick response** 是互联网核心的用户体验。线下模型训练可以花费很长时间: 比如, Google 某个模型更新一次需要几千台机器, 大约训练半年时间。但是, 线上使用模型的时候要求一定要 “快, 实时 (real-time)”

◆ **Online learning**: 互联网每时每刻都在产生大量新数据, 要求模型随之不停更新, 所以online learning是机器学习的一个重要研究方向。





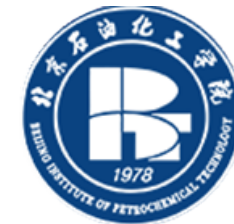
感谢聆听



有什么问题吗？



考考大家



机器学习 VS

数据挖掘

人工智能

统计学习

