華東師範大學
EAST CHINA NORMAL UNIVERSITY

第一章 机器学习基础

Machine Learning Basis

第1.1节 机器学习简介

第1.2节 机器学习基本概念

第1.3节 机器学习性能度量

第1.1节 机器学习简介

一、机器学习概念

二、机器学习与其他概念的区别

三、机器学习挑战

四、机器学习的历史和未来

一、 机器学习概念

机器学习是近40多年兴起的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，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

机器学习是人工智能(Artificial Intelligence, AI）的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，它主要使用归纳、综合而不是演绎。

1. 什么是机器学习

学习:

- 赫伯特·西蒙(1959)：如果一个系统，能够通过执行某个过程，就此改进了它的性能，那么这个过程就是学习。

一、 机器学习概念

机器学习:

赫伯特·西蒙(1959)：机器学习的目的是让计算机拥有自主学习的能力，而无须对其进行事无巨细的编程。（强调学习的自主性）

Langley(1996)：机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能

Tom Mitchell(1997): 机器学习是对能通过经验自动改进的计算机算法的研究。（强调学习的效果）

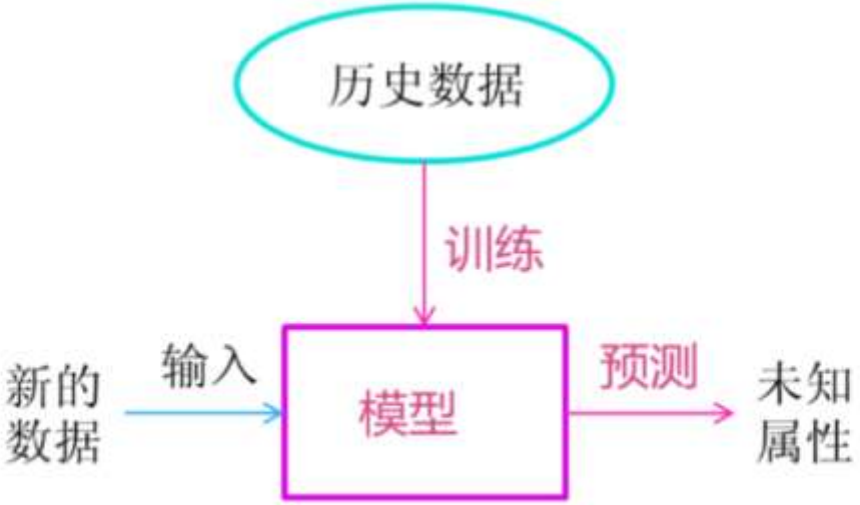
- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E
- 如果一个程序在使用既有的经验E(Experience)来执行某类任务T(Task)的过程中被认为是具备学习能力的，那么它一定要展现出：利用现有的经验E，不断改善其完成既定任务T的性能(Performance)的特质。

一、 机器学习概念

特雷弗·哈斯蒂《统计学习基础》：机器学习就是抽取重要的模式和趋势，理解数据的内涵表达。（强调从数据中学习）

弗拉基米尔·万普尼克《统计学习理论的本质》:机器学习就是一个基于经验数据的函数估计问题。（侧重可操作性）

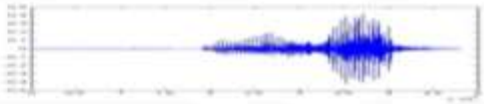
探究和开发一系列算法来如何使计算机不需要通过外部明显的指示，而可以自己通过数据来学习，建模，并且利用建好的模型和新的输入来进行预测的学科。（侧重机器学习的流程）



一、 机器学习概念

机器学习的本质: 构建一个映射函数

语音识别:

$f($  $)=$ “你好”

图像识别:

$f($  $)=$ “猫”

围棋:

$f($  $)=$ “5-5” (落子位置)

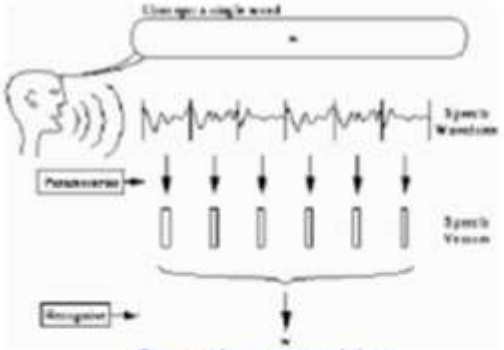
对话系统:

$f($ “你好” $)=$ “今天天气真不错”

用户输入　　机器

一、 机器学习概念

2. 机器学习的应用

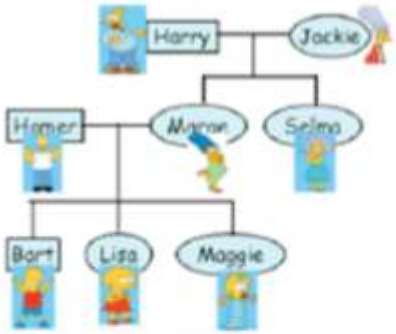


Speech recognition

Information retrieval

Computer vision

Games



Pedigree

Evolution

Robotic control

Planning

二、 机器学习与其他概念的区别

1. 机器学习与人工智能

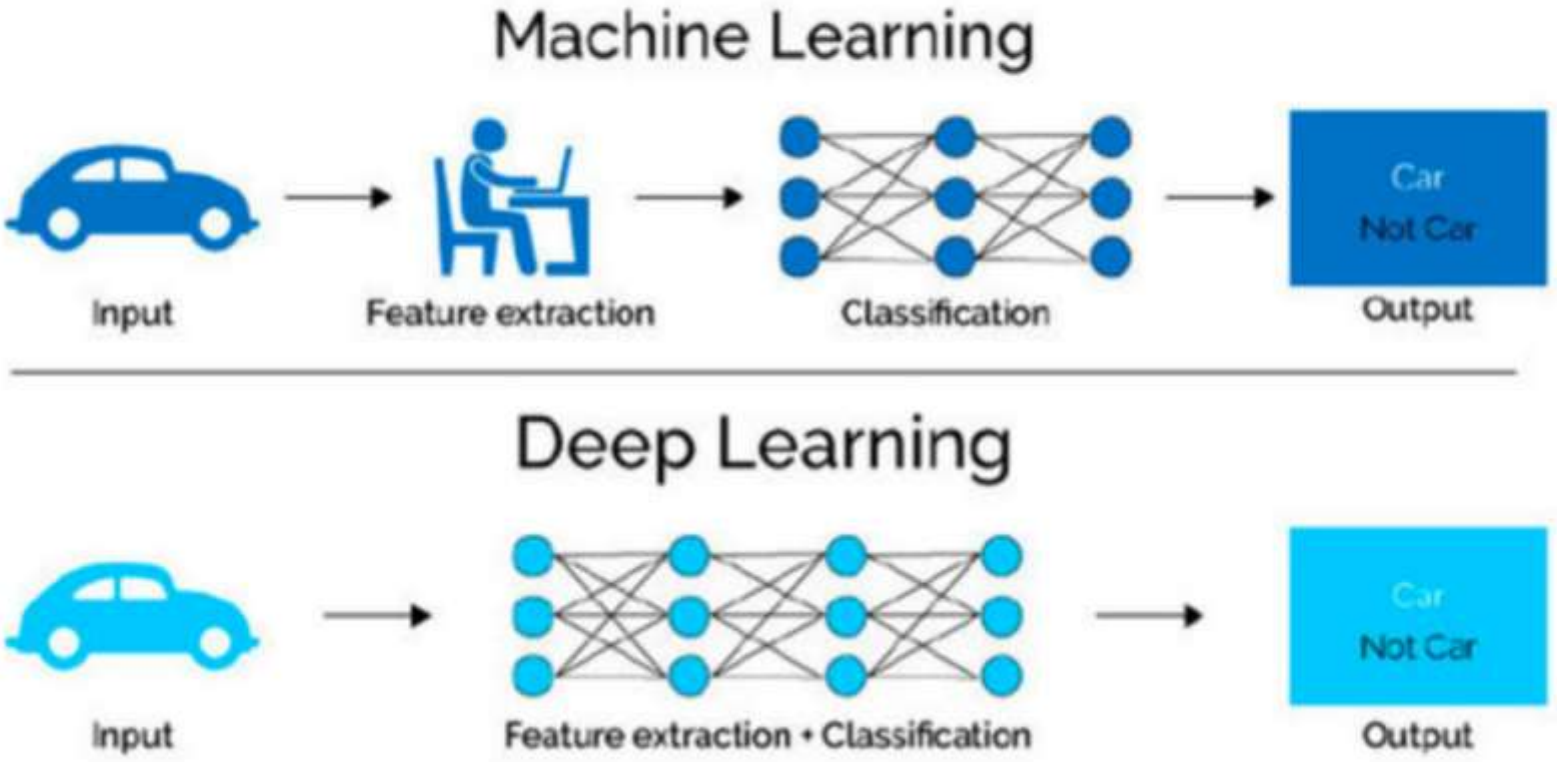
机器学习：实现人工智能的一种方法

- 人工智能: 让机器展现人类智能
- 机器学习：实现人工智能的一种方法
- 深度学习：实现机器学习的一种技术

二、 机器学习与其他概念的区别

2. 机器学习与深度学习

深度学习是机器学习的一个子领域，特征提取更依赖于隐层模型，解释性弱，趋于黑盒子，对数据依赖性更强，更擅长处理高维度大数据。



二、 机器学习与其他概念的区别

3. 机器学习和数据挖掘

机器学习是数据挖掘的重要工具。

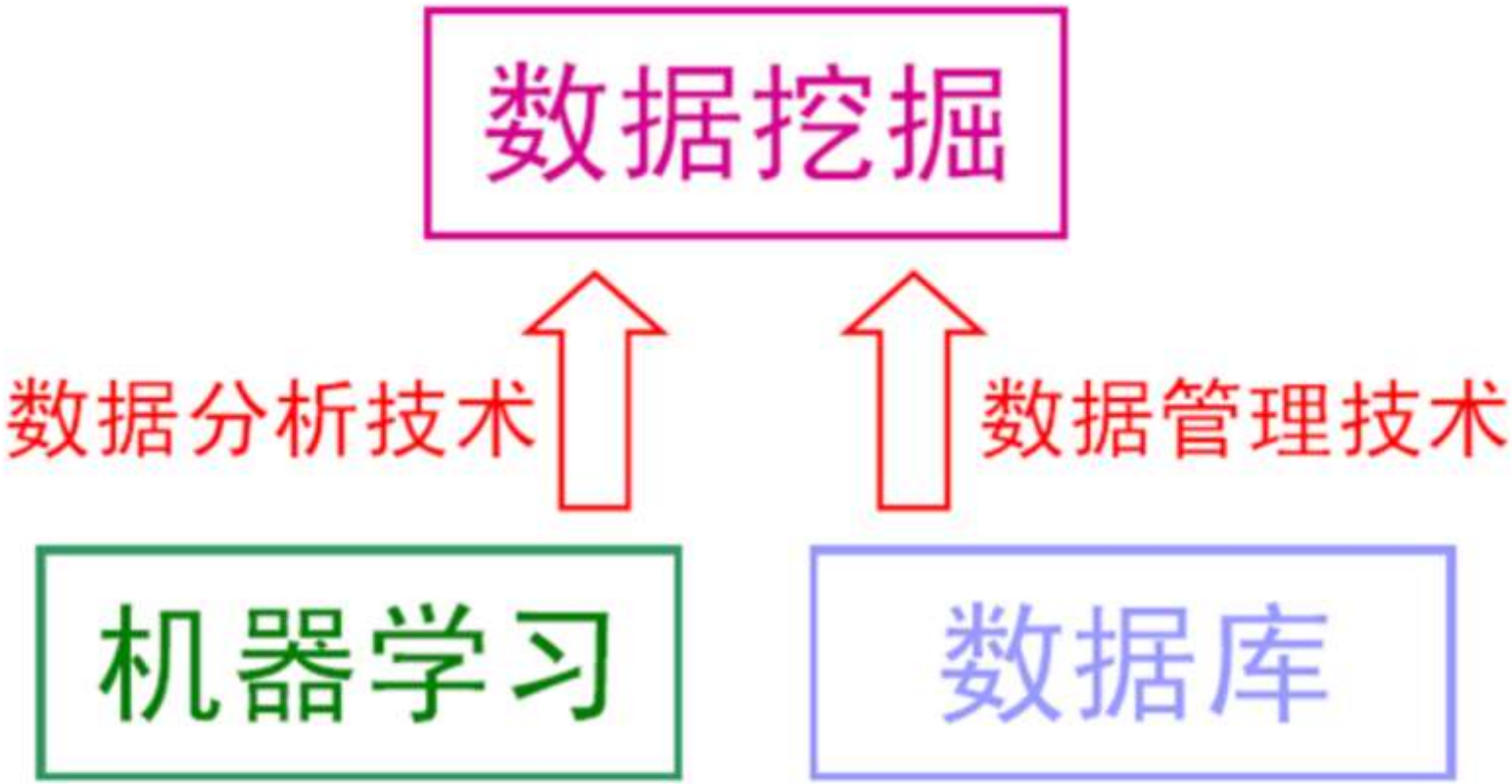
数据挖掘不仅仅要研究、拓展、应用一些机器学习方法，还要通过许多非机器学习技术解决数据仓储、大规模数据、数据噪音等等更为实际的问题。

机器学习的涉及面更宽，常用在数据挖掘上的方法通常只是“从数据学习”，然则机器学习不仅仅可以用在数据挖掘上，一些机器学习的子领域甚至与数据挖掘关系不大，例如增强学习与自动控制等等。

数据挖掘试图从海量数据中找出有用的知识。

大体上看，数据挖掘可以视为机器学习和数据库的交叉，它主要利用机器学习界提供的技术来分析海量数据，利用数据库界提供的技术来管理海量数据。

二、 机器学习与其他概念的区别



4. 机器学习和统计学习

统计学习是theory-driven，对数据分布进行假设，以强大的数学理论支撑解释因果，注重参数推断（Inference）

机器学习是data-driven，依赖于大数据规模预测未来，弱化了收敛性问题，注重模型预测（Prediction）

- 理解和预测
  - 解释因果：统计学习（theory drive）
    - 回归和假设检验
  - 预测未来：机器学习（data drive）
    - 优化问题

5. 机器学习与传统编程

机器学习通过程序让计算机来模拟人的学习过程

例：通过身高x，预测体重y

传统编程:

(1)确定输入x, 输出y

(2)[根据已有数据集]，通过人的经验或者查询资料，确定x和y的关系：$y = 0.9x - 90$

机器学习:

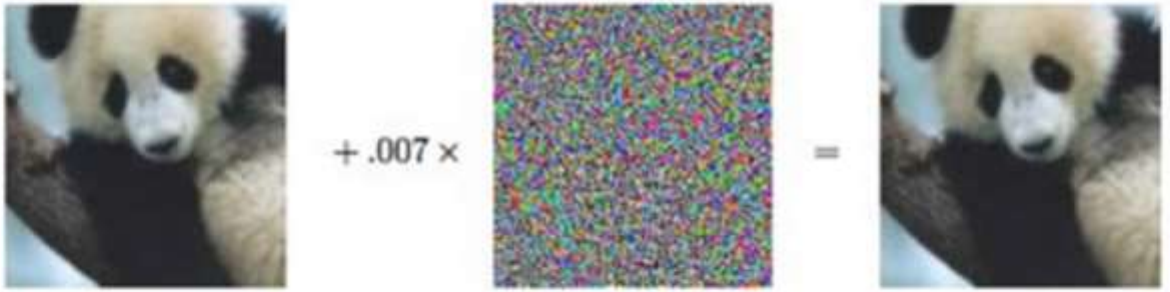
(2a) 设计模型为$y = ax + b$，编写学习算法，对已有数据集进行训练，得到预测模型$y = 0.8x - 100$

6. 机器学习的适用条件

- 适用条件
  - 事物本身存在某种潜在规律
  - 某些问题难以使用普通编程解决（图像识别、语音识别）
  - 有大量的数据样本可供使用
- 大数据
  - Web：Google index 包括大约450亿页面
  - Click-stream data: 10-100TB/天
  - Transaction data：5-50TB/天
  - TV：2TB/天/频道；YouTube 4TB/天 的上传量
  - Photos：15亿张/周的上传量
  - 数字电话：100 PB/天

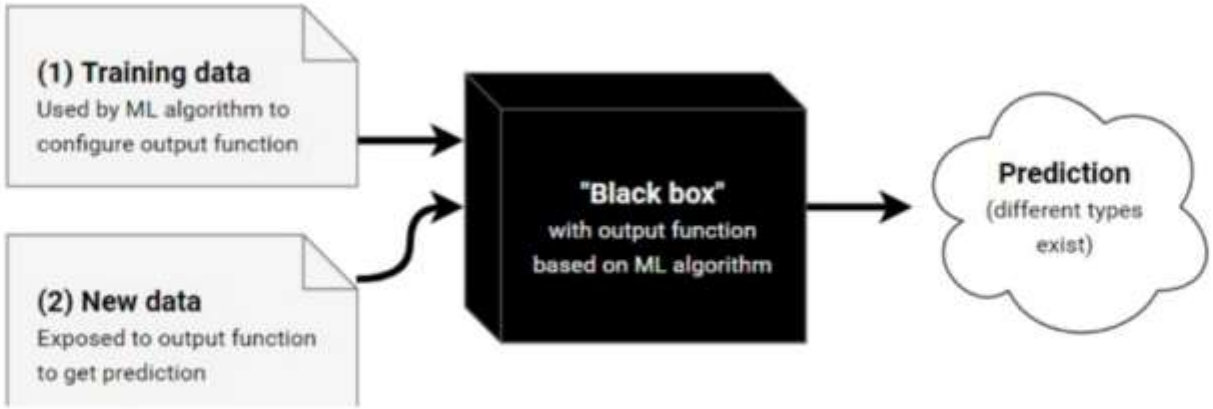
三、 机器学习挑战

- 模型的预测效果
- 模型的稳定性
  - 对抗样本：攻击者通过在源数据上增加人类难以通过感官辨识到的细微改变，却可以让机器学习模型接受并做出错误的分类决定。
  - 典型的场景就是图像分类模型的对抗样本，通过在图片上叠加精心构造的变化量，在肉眼难以察觉的情况下，让分类模型产生误判。



三、 机器学习挑战

- 模型结果的可解释性
  - 算法歧视



四、 机器学习的历史和未来

1. 机器学习的历史

- 符号主义(symbolism): 20世纪80年代到90年代中期
  - 所有的智能行为都可以被简化成在一个逻辑系统中的符号操作过程
  - 符号主义可以学习产生明确的概念表示.
    - 有强知识表示能力, 可以较容易地表达出复杂数据关系, 而且领域知识通常可方便地通过逻辑表达式进行描述
    - 不仅可利用领域知识辅助学习, 还可以通过学习对领域知识进行精化和增强.
    - 表示能力太强, 直接导致学习过程面临的假设空间太大、复杂度极高, 因此问题规模稍大就难以有效进行学习.
  - 自动定理证明、专家系统、知识图谱
  - 符号系统的学习是感知智能走向认识智能的基础

四、 机器学习的历史和未来

- 贝叶斯学派: 20世纪90年代中期至21世纪初.
  - 概率图模型：通过引入概率工具来描述事件的不确定性
    - 随机变量：描述事件的随机性
    - 条件概率：描述事件之间的关系
    - 推理过程：进行后验概率的计算
    - 利用数据来调整模型参数
  - 推理过程比较复杂

四、 机器学习的历史和未来

- 连接主义(connectionism): 20世纪50年代中后至90年代中期、21世纪初至今.
  - M-P 神经元模型、误差逆传播 (error back-propagation, BP)算法、深度学习 (deep learning)
  - 优点:
    - 对使用者要求不高.
    - 模型复杂度高、容量大、学习能力强, 可以在很多现实问题 (尤其是设计语音、图像等复杂对象的应用) 上发挥作用.
  - 局限:
    - 连接主义学习产生的是黑箱模型,缺乏严格的理论基础
    - 学习过程设计大量超参数, 参数设置缺乏指导, 主要靠手工调参
    - 模型需要连续可微, 难以处理符号化、离散数据

2. 机器学习的历史机器学习的未来

弱监督条件下的学习

- 在许多现实任务中，既缺乏大量有标记数据，又难以通过无成本探索获得大量训练样本
- 弱监督信息
  - 监督信息不完全. 例如医学图像中，少部分得到了专家标注，而大部分没有标记. 这种情况下可以进行半监督学习和主动学习.
  - 监督信息不具体. 例如医学图像中，某图像被标记出有病灶，但是未具体标出病灶在哪. 这种情况下可以进行多示例学习
  - 监督信息不精确. 例如医学图像中，专家由于疲劳、疏忽等原因出现标记错误. 这种情况下可以进行带噪声学习和众包学习.

四、 机器学习的历史和未来

多样化的深度学习算法

- 深度神经网络要求计算单元是连续可微的
- 但是，现实任务中的数据并不都是实值的，如何利用深度学习处理符号数据、离散数据，是未来的一个研究方向
- 图神经网络 (graph neuron networks, GNN)

开放动态环境下的学习

- 在许多现实任务中，经常会遇到开放动态环境，高风险应用
- 要求学得的模型具有很高的稳健性 (robust)

因果学习

- 现在的机器学习算法大多只能推断相关性 (correlation)，而不能得到因果 (casuality)
- 因果学习得到的模型具有很好的可解释性

不同视角的机器学习

不同行业的人以为我做的事情

父母以为我做的事情

朋友以为我做的事情

$$\frac{\partial}{\partial w}L(w,b,\alpha) = w - \sum \alpha_i y_i x_i = 0,\ w = \sum \alpha_i y_i x_i$$
$$\frac{\partial}{\partial b}L(w,b,\alpha) = \sum \alpha_i y_i = 0$$

代入$L(w,b,\alpha)$

$$\begin{aligned}\min\ L(w,b,\alpha) &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m}\alpha_i(-y_i(w^T x_i + b) + 1)\\ &= \frac{1}{2}w^T w - \sum_{i=1}^{m}\alpha_i y_i w^T x_i - b\sum_{i=1}^{m}\alpha_i y_i + \sum_{i=1}^{m}\alpha_i\\ &= \frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y_i x_i - \sum_{i=1}^{m}\alpha_i y_i w^T x_i + \sum_{i=1}^{m}\alpha_i\\ &= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i w^T x_i\\ &= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i x_j)\end{aligned}$$

将此 max 问题转成 min 问题：

$$\max\ \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i x_j) = \min\ \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i x_j) - \sum_{i=1}^{m}\alpha_i$$

$$\text{s.t.}\ \sum_{i=1}^{m}\alpha_i y_i = 0,$$

程序员以为我做的事情

我自己以为我做的事情

```
import xgboost as xgb
import numpy as np
```

实际上我做的事情

Which of the following is best suited for machine learning?

1. predicting whether the next cry of the baby girl happens at an even-numbered minute or not
2. determining whether a given graph contains a cycle
3. deciding whether to approve credit card to some customer
4. guessing whether the earth will be destroyed by the misuse of nuclear power in the next ten years