教材《机器学习与大语言模型》配套电子教案

机器学习

第一讲: 机器学习概述

顾小东
上海交通大学计算机学院

内容纲要

AI发展史

为什么要学习机器学习(Why)?

机器学习是什么(What)?

- 机器学习核心要素

如何构建机器学习系统(How)?

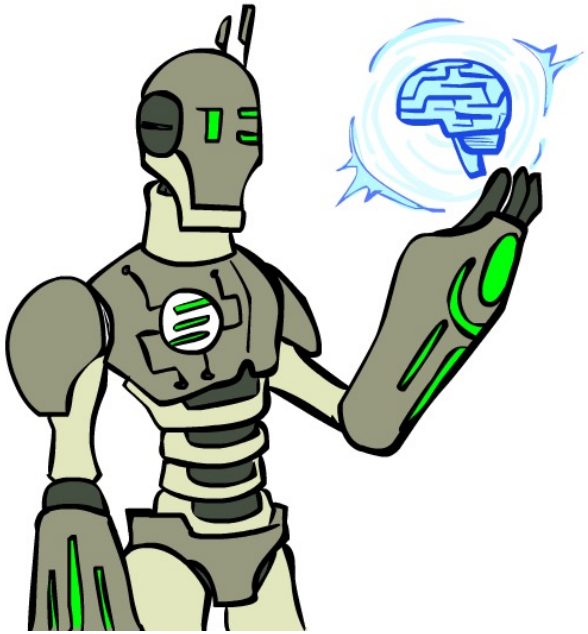
- 通用流程
- 运行示例(working example)

机器学习算法总览

- 分类
- 典型算法

学习or记忆？泛化性与过拟合

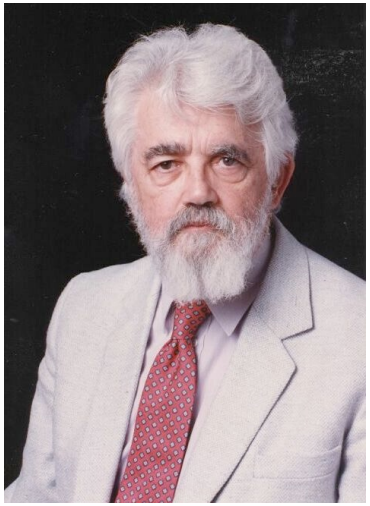
- 泛化性、过拟合、正则化



什么是人工智能 (AI)？

什么是智能?

- 拥有逻辑、理解、自我意识、学习、推理、规划、创造、批判思考和问题解决的能力。



John McCarthy
one of the founders of the discipline of AI

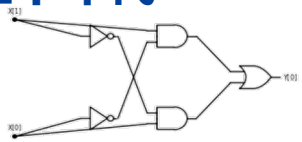
什么是人工智能?

- 用计算机程序构造智能系统

AI发展简史

二十世纪四五十年代

Early Days

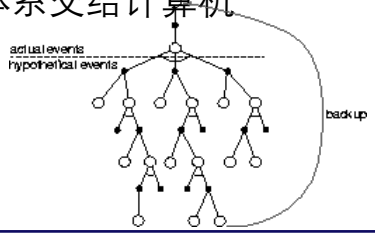


- 1943: Mcculloch & Pitts: Boolean circuit model of brain
- 1950: Turning’ s “Computing Machinery and Intelligence”

二十世纪五十至七十年代

Excitement: Look, Ma, no hands!

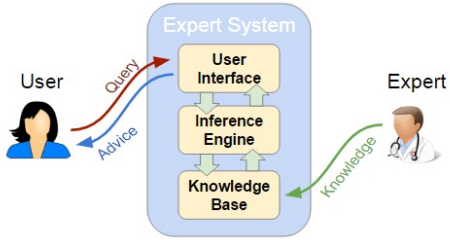
把人总结的知识体系交给计算机
专家系统



二十世纪七十至九十年代

Knowledge-based Approaches

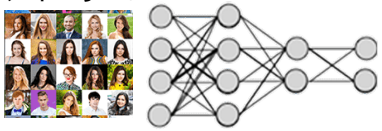
从样例中学习
符号主义、联结主义
统计学习



二十一世纪

深度学习

深度神经网络
表征学习



人工智能 vs 机器学习 vs 深度学习

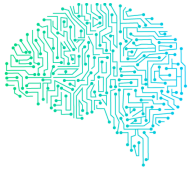
人工智能

能够处理一般需要人类知识和智力介入任务的计算机系统

机器学习

从若干输入输出样例中学习数据规律的统计技术

深度学习



一种自动学习数据表征的机器学习技术

机器学习发展简史

1950-1960s

逻辑推理

机器学习概念
图灵测试
人工神经网络

1970-1980s

知识系统

把人总结的知识体系交给计算机
专家系统

1980-1990s

归纳和统计学习

Learning from examples
BP Neural Networks
Statistical Learning

2012-

深度学习

深度神经网络
卷积

2022-

大语言模型

Transformer
ChatGPT

为什么要机器学习 (Why)？

机器学习 vs 传统编程

传统编程

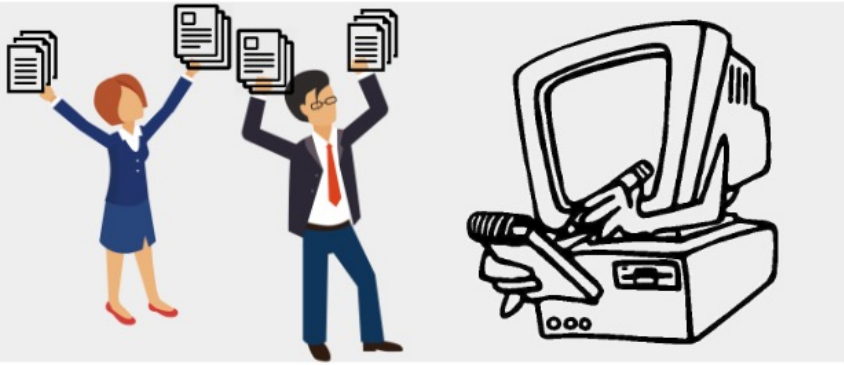
编写一段程序，按照显式定义的一组指令执行

机器学习

编写一段程序，自动从样例中总结处理过程

```
if email contains “发票”
  then mark is-spam;
if email contains …
if email contains …
```

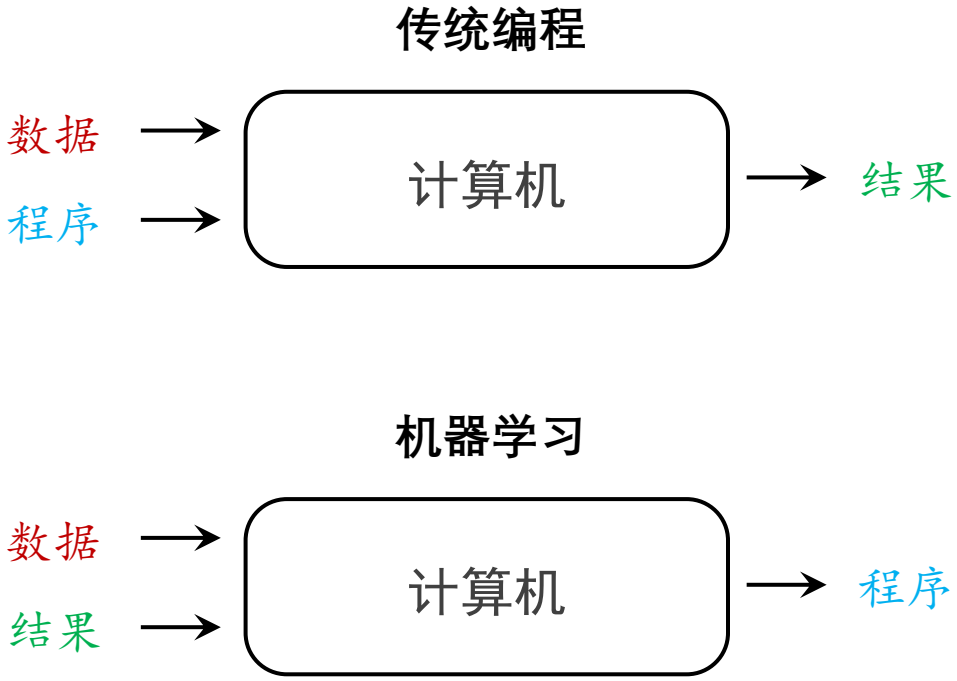
太多决策规则…

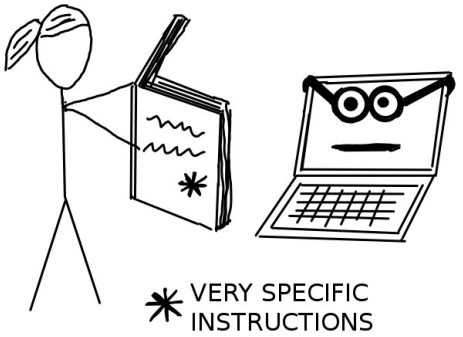


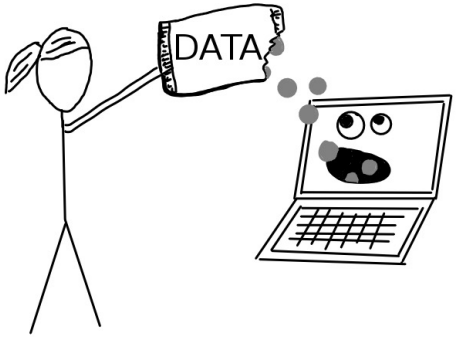
Magic!

机器学习 = “学习不用编程完成任务的能力”

机器学习 vs 传统编程

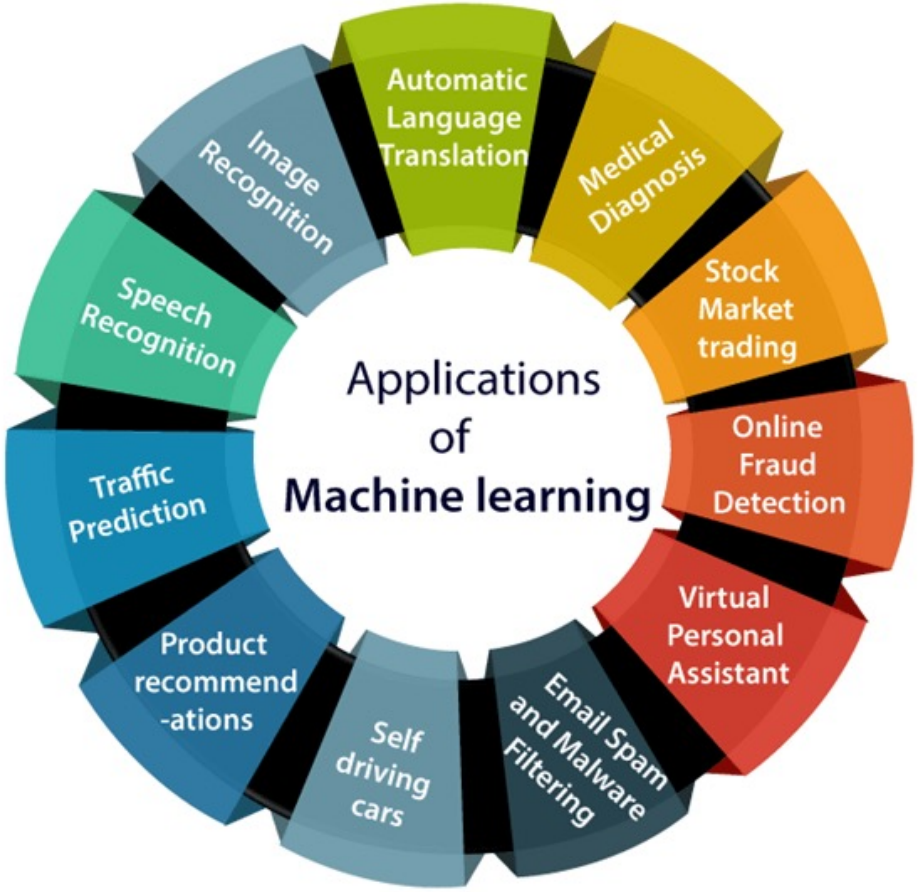

传统编程
数据
程序
计算机
结果
机器学习
数据
结果
计算机
程序


VERY SPECIFIC
INSTRUCTIONS


DATA

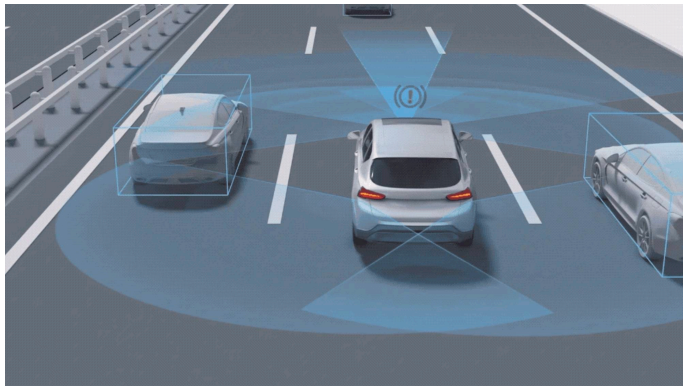
机器学习的应用

无处不在的机器学习



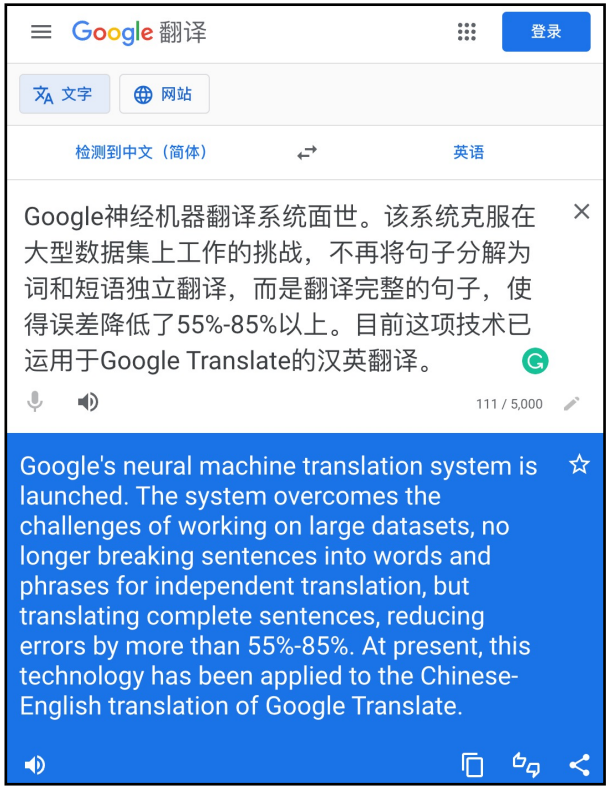
机器视觉

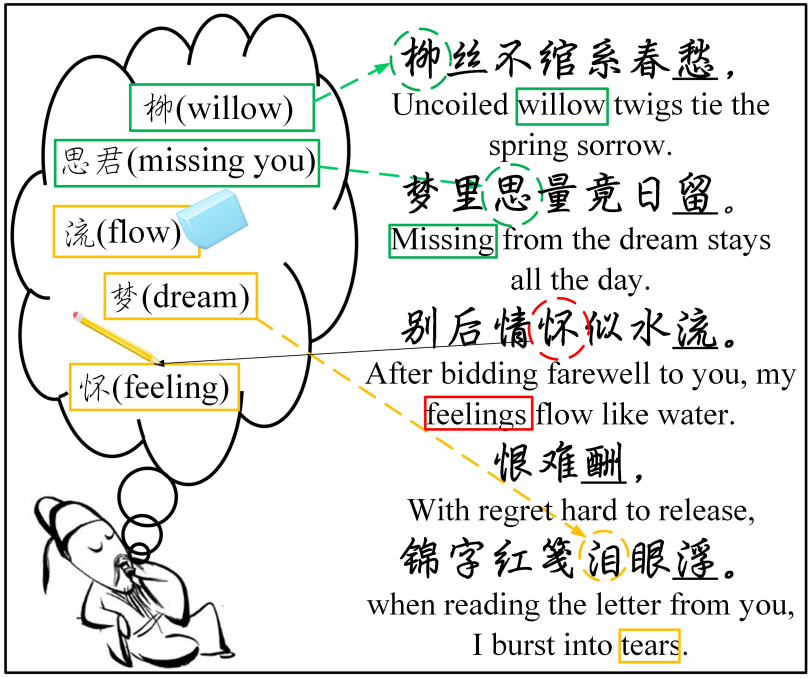
目标检测、语义分割、医疗诊断、自动驾驶…



自然语言处理

翻译、问答、摘要、报告….



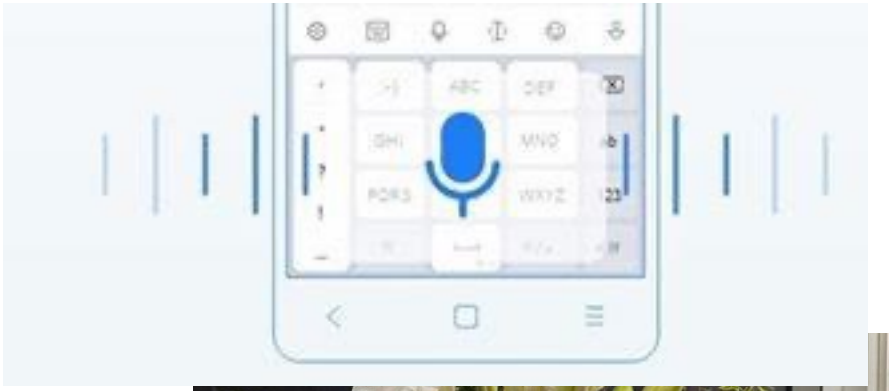


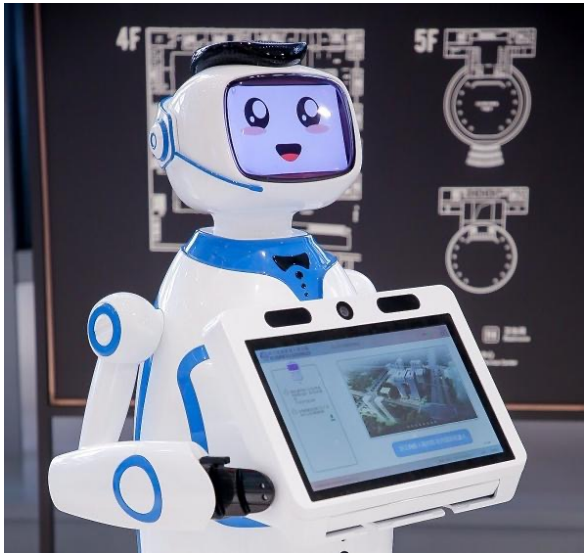
Chinese poem generation. Can you distinguish?

Yi et al. Chinese Poetry Generation with a Working Memory Model. IJCAI'18

语音助手

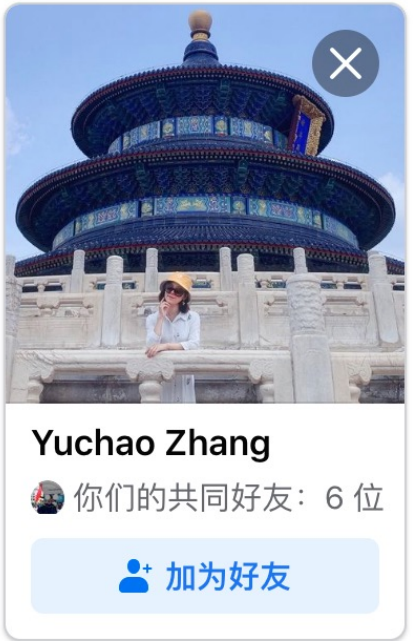
语音识别，智能助理





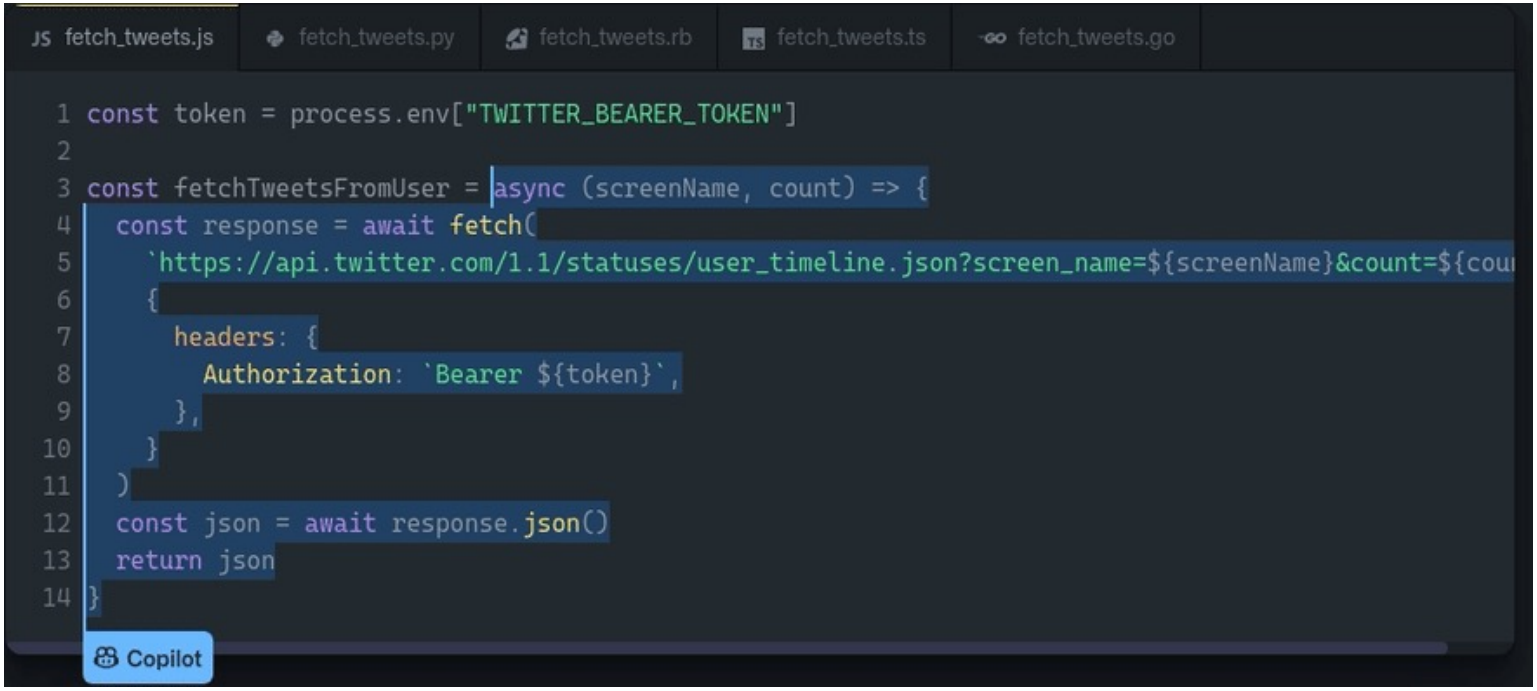
推荐系统

短视频推荐、商品推荐、好友推荐



AI编程

自动生成程序


fetch_tweets.js
fetch_tweets.py
fetch_tweets.rb
fetch_tweets.ts
fetch_tweets.go
1 const token = process.env["TWITTER_BEARER_TOKEN"]
2
3 const fetchTweetsFromUser = async (screenName, count) => {
4 const response = await fetch(
5 `https://api.twitter.com/1.1/statuses/user_timeline.json?screen_name=${screenName}&count=${cou
6 {
7 headers: {
8 Authorization: `Bearer ${token}`,
9 },
10 }
11)
12 const json = await response.json()
13 return json
14 }
Copilot

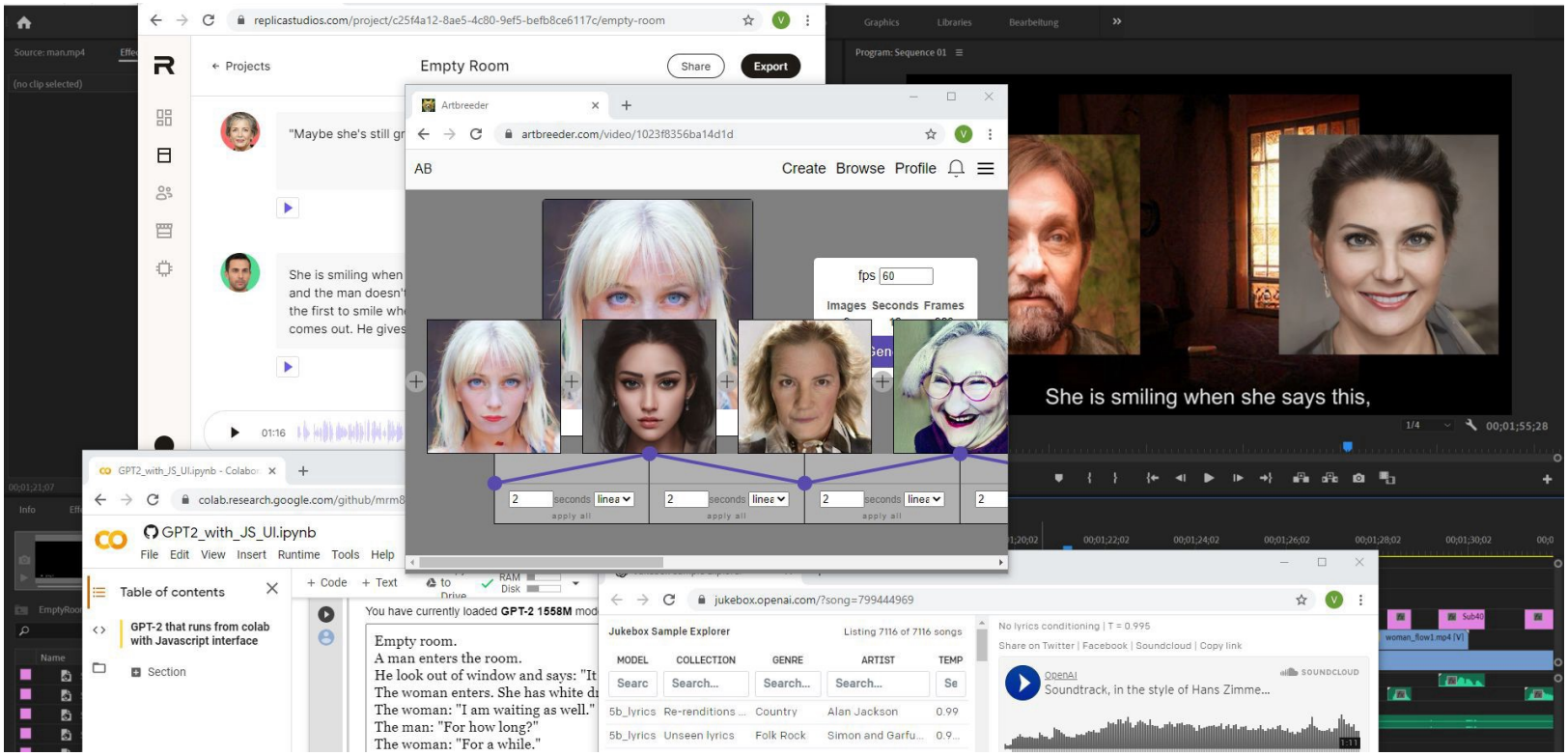
AI作画

DALL.E

https://openai.com/index/dall-e-3/

Original: *Girl With a Pearl Earring* by Johannes Vermeer

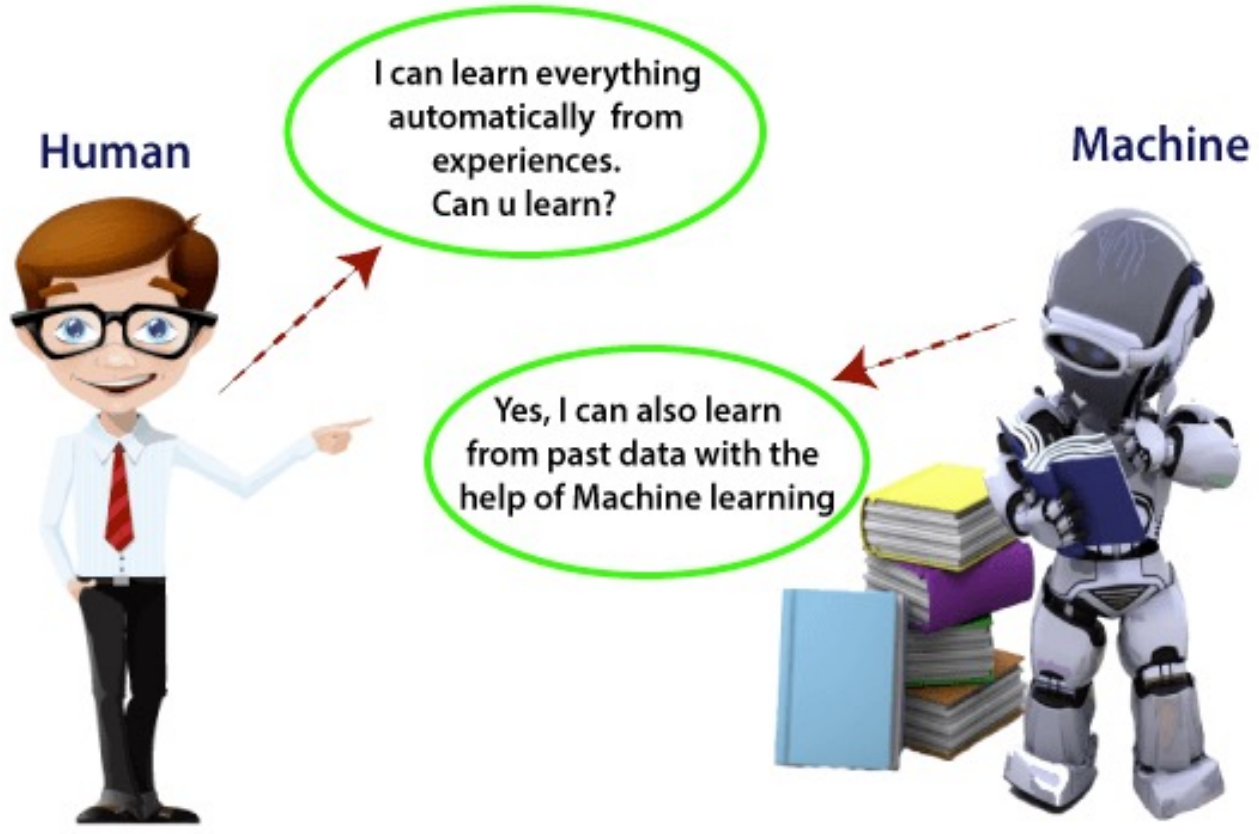
AI影片


replicastudios.com/project/c25f4a12-8ae5-4c80-9ef5-befb8ce6117c/empty-room
Projects
Empty Room
Share
Export
Artbreeder
artbreeder.com/video/1023f8356ba14d1d
AB
Create Browse Profile
fps 60
Images Seconds Frames
She is smiling when she says this,
GPT2_with_JS_UI.ipynb
File Edit View Insert Runtime Tools Help
Table of contents
GPT-2 that runs from colab with Javascript interface
Section
Empty room.
A man enters the room.
The woman enters. She has white
The woman: "I am waiting as well."
The man: "For how long?"
The woman: "For a while."
jukebox.openai.com/?song=799444969
Jukebox Sample Explorer
Listing 7116 of 7116 songs
MODEL COLLECTION GENRE ARTIST TEMP
5b_lyrics Re-renditions ... Country Alan Jackson 0.99
5b_lyrics Unseen lyrics Folk Rock Simon and Garfu... 0.9...
Soundtrack, in the style of Hans Zimme...

什么是机器学习(What)？

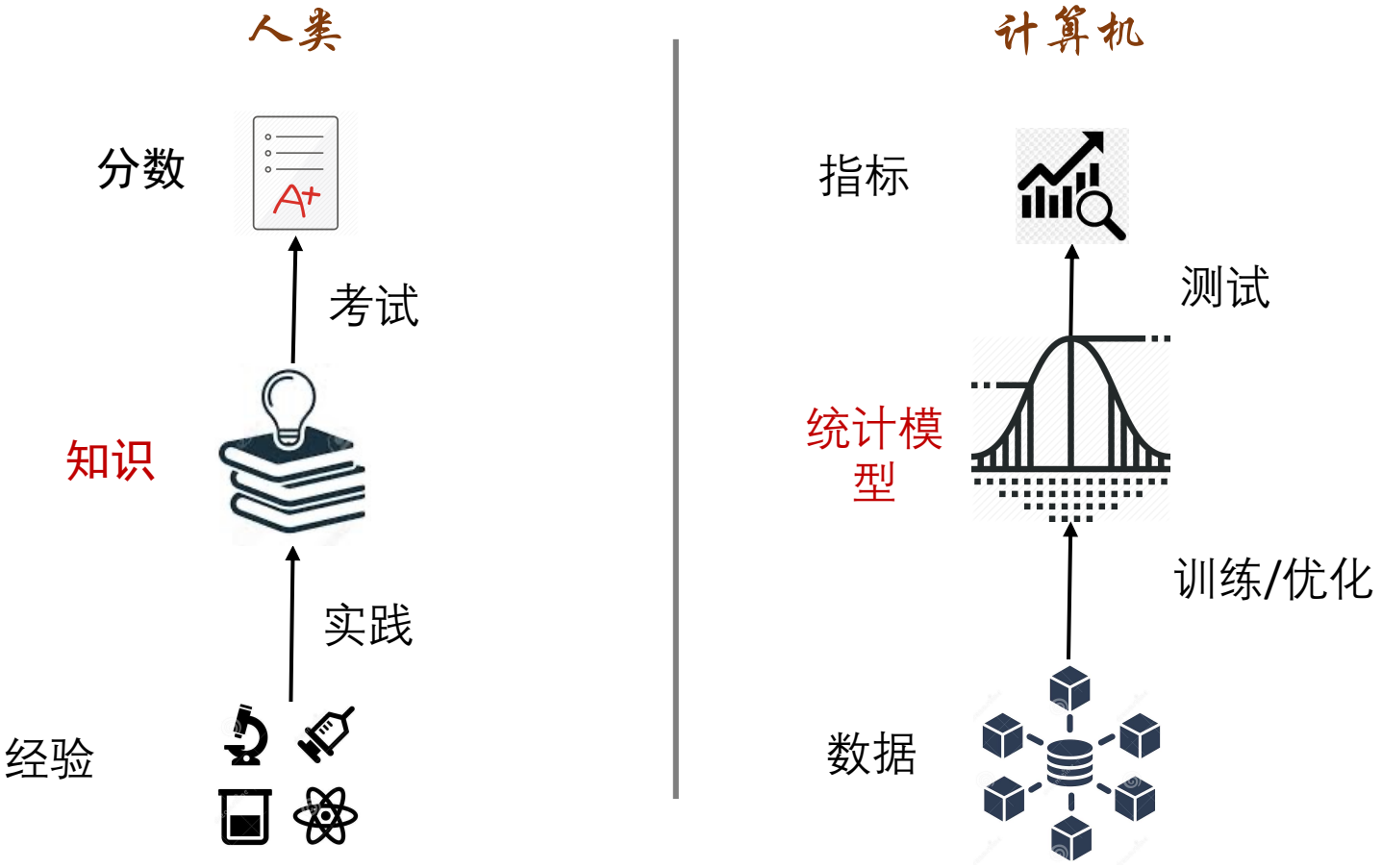
什么是机器学习？

人类学习 vs. 机器学习



一个简单的类比

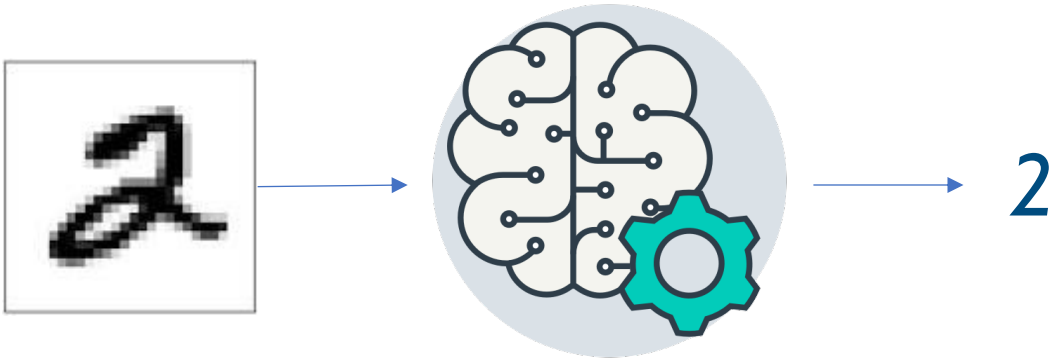
人类学习 vs. 机器学习



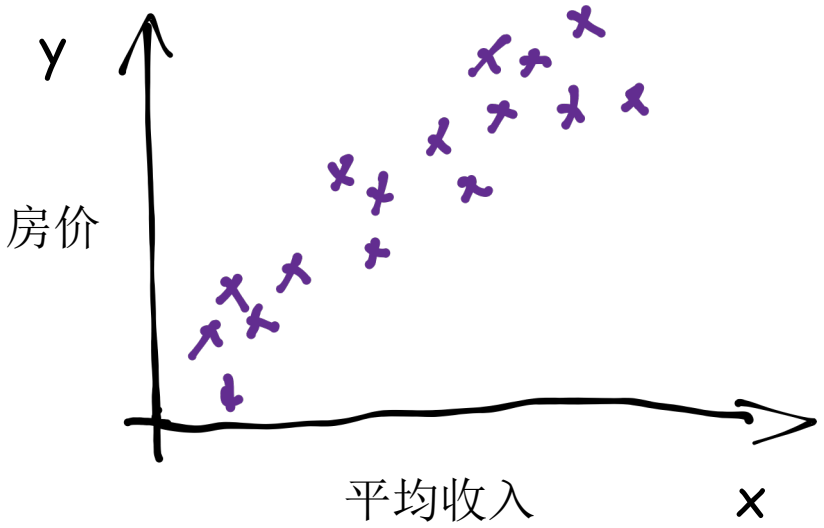
机器学习是什么？

定义

机器学习是一类通过从数据中发掘潜在规律来不断优化自身性能，以实现特定目标的计算机算法。



机器学习的核心要素



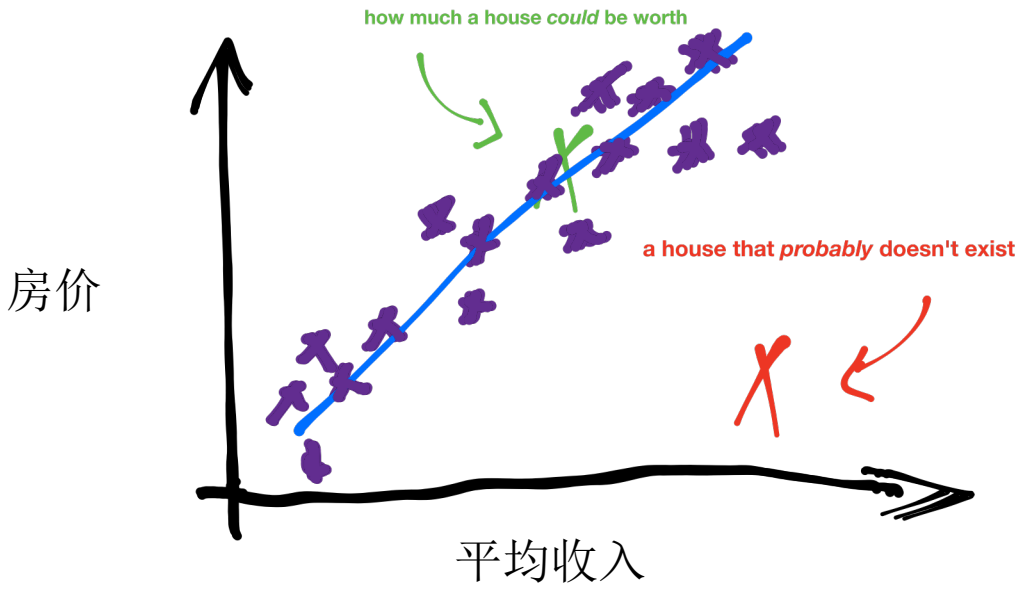
要素:

#1 数据(经验)

$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\}$

机器学习的核心要素

仅凭数据不能学习
数据+假设才能学习



从数据中发现什么基本规律？

要素：

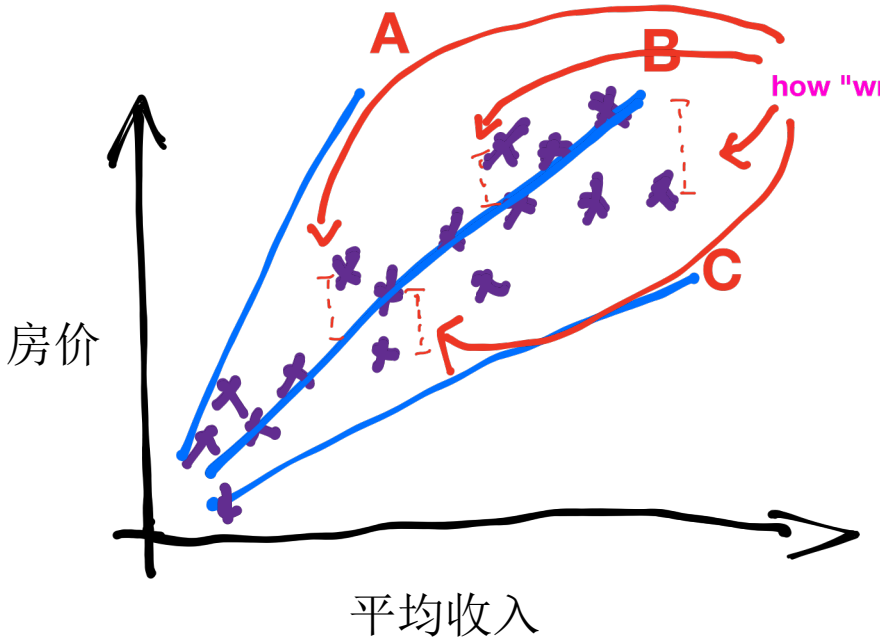
#1 数据（经验）

#2 模型（假设）

$$y = g(x|\theta) + \epsilon$$

e.g. $\hat{y} = \boldsymbol{w}x + \boldsymbol{b}$
where θ (e.g., W and b) denotes the parameters

机器学习的核心要素



how "wrong" the prediction is

要素:

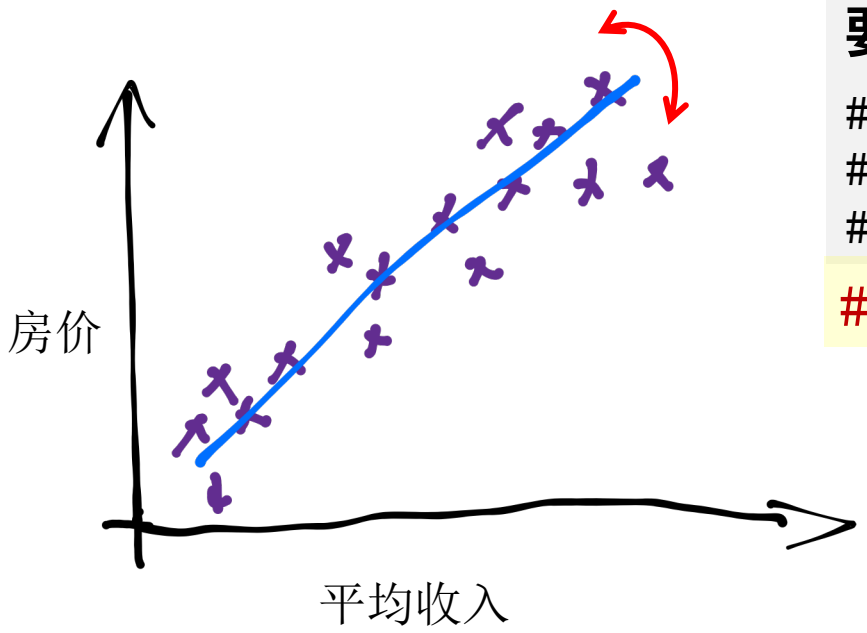
#1 数据（经验）
#2 模型（假设）

#3 损失函数（目标）

$$\mathcal{L}(\theta|\mathcal{D}) = \frac{1}{2}\sum_{n=1}^{N}[wx_n + b - y_n]^2$$

如何评估模型的效果?(性能指标)

机器学习的核心要素

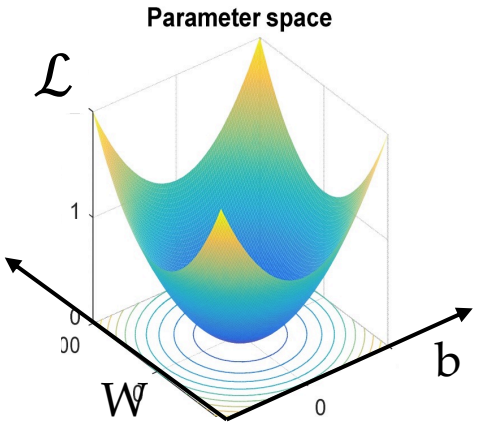


要素:

#1 数据（经验）
#2 模型（假设）
#3 损失函数(目标)

#4 优化算法(提升)

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta|\mathcal{D})$$

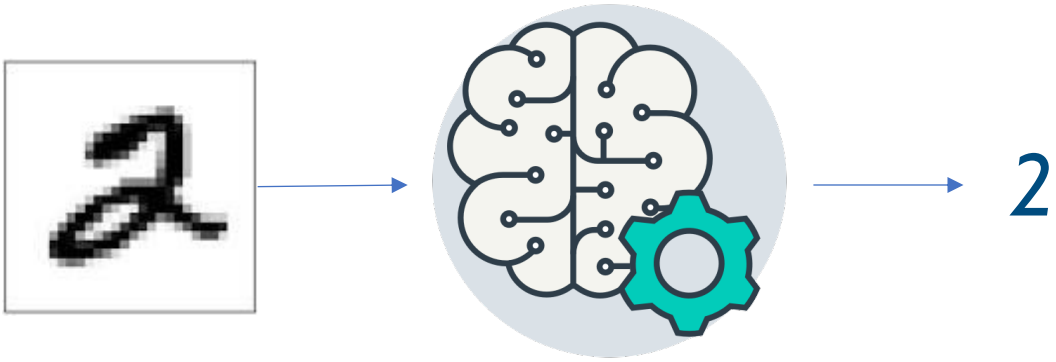


如何找到最佳的模型？

归根结底，机器学习是什么？

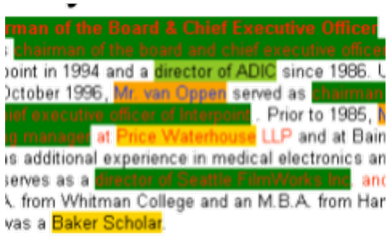
定义

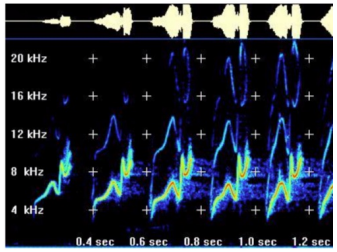
机器学习是一种人工智能技术，使机器能够在无需显式人工编程的情况下，通过从数据中自动发现潜在的统计规律（模型），并不断优化自身性能（优化），以实现特定的目标（如最小化损失函数）。

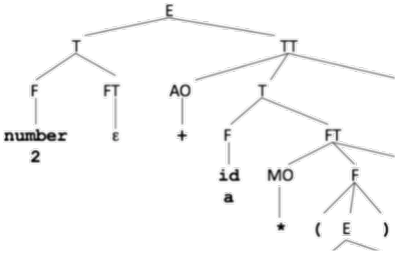


推广到其他类型的数据

其他形式的数据该如何处理?

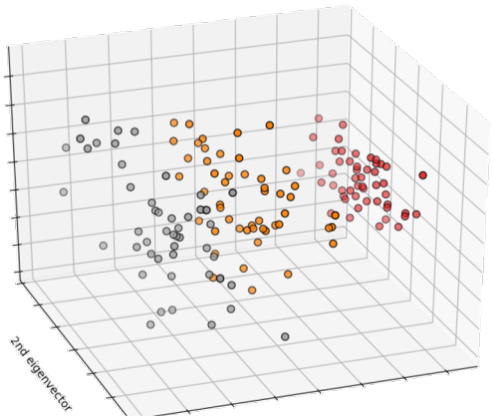






现实世界的复杂数据

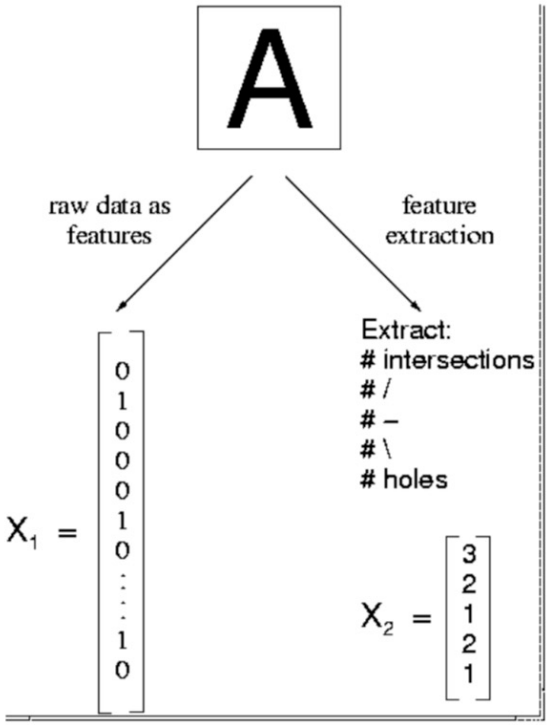
向量化/特征提取

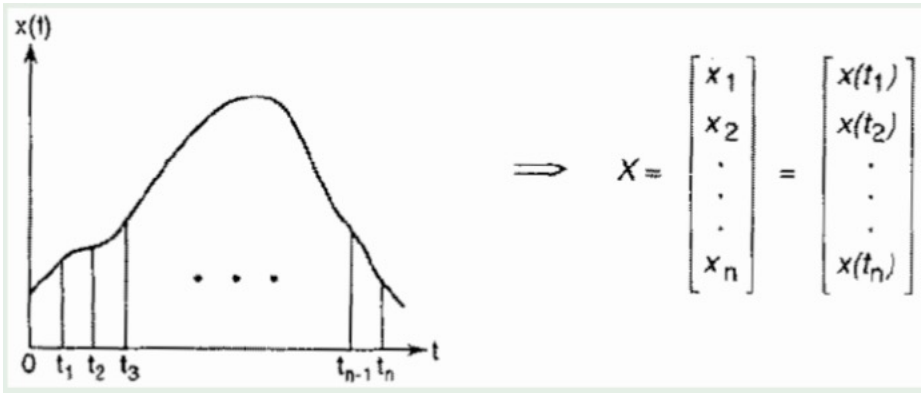


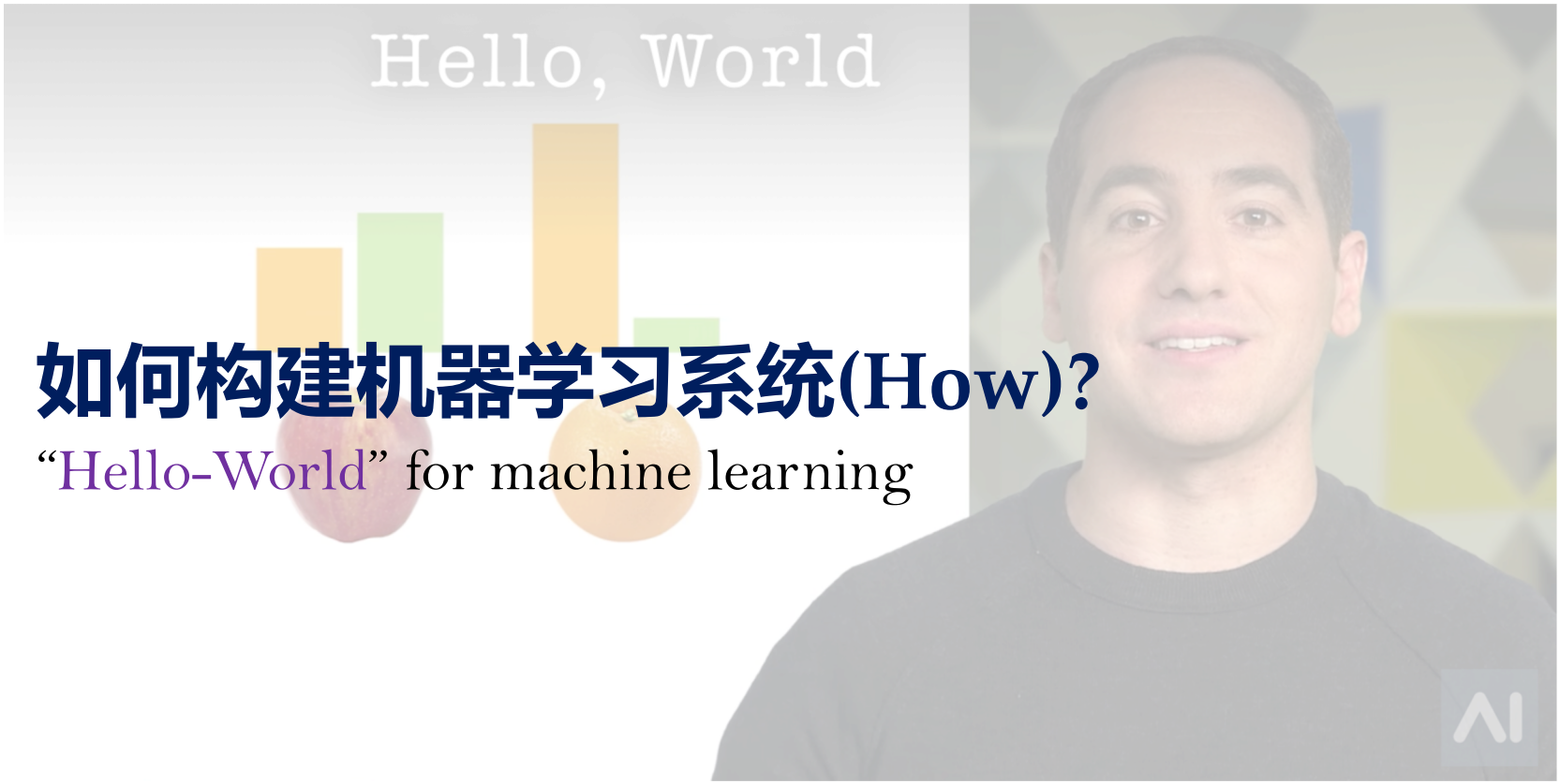
高维向量空间中的点

推广到其他类型的数据

向量化/特征提取







如何构建机器学习系统(How)?

"Hello-World" for machine learning

如何构建一个机器学习系统？

① 考虑数据可行性和规模

- 能获得多少数据？需要多少代价 (时间、算力、人力成本)?

② 选择输入数据的合理表示形式

- 数据预处理, 特征提取, 等等.

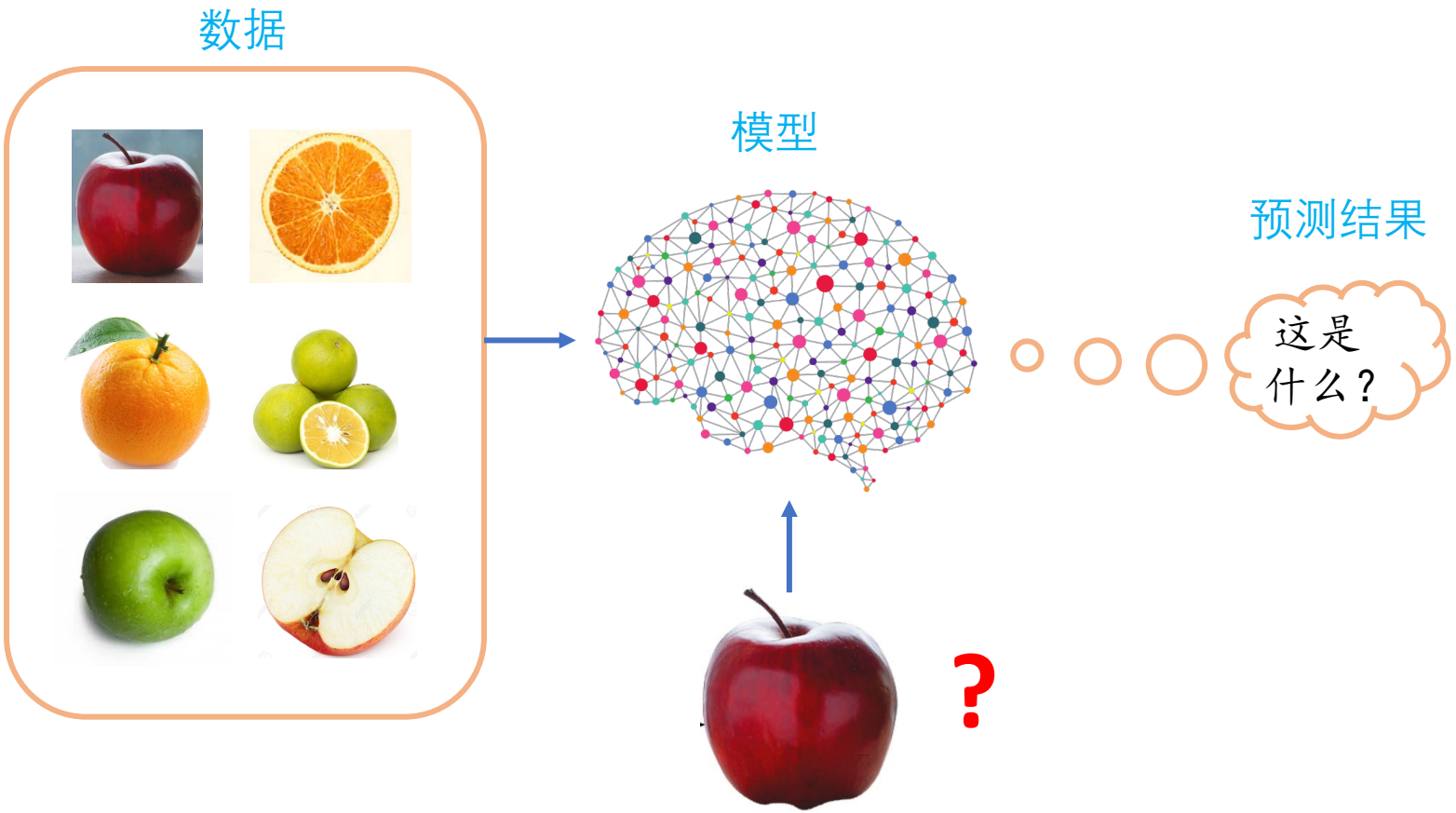
③ 选择合理的模型假设

④ 选择合适的损失函数

⑤ 选择或设计一个学习算法

演示实例 (A Working Example)

任务: 自动水果分拣系统



步骤一：数据收集

构造训练数据

- Tell me which type a specific fruit belongs to

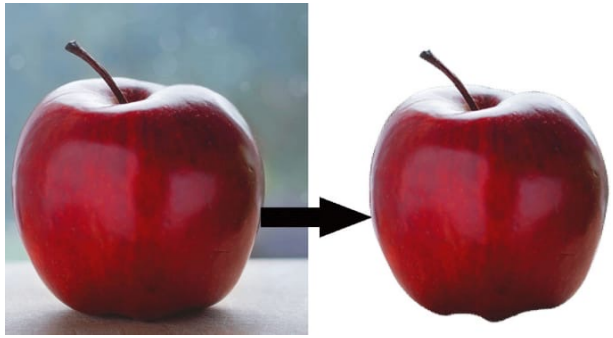
数据样本	标签
	Apple
	Orange
	Apple
…	…

步骤二：数据预处理

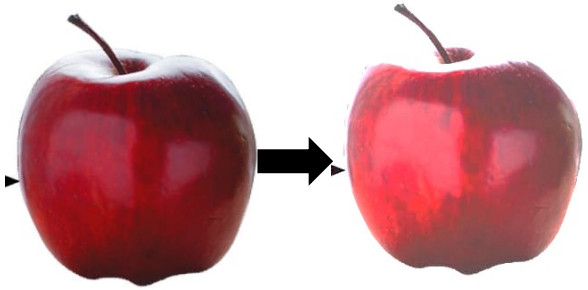
Different types of fruit differ in size, lightness, position, etc.

需要预处理！

举例:



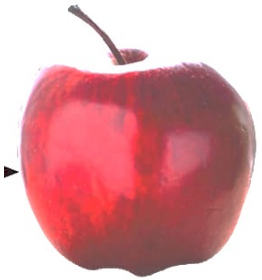
去除背景



调整光照亮度

步骤三：特征提取

将每个水果视为特征空间的一个点



特征:

颜色：红色
形状：圆形
大小：3
...

$$\begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}$$

特征:

颜色：橙色
形状：圆
大小：2
...

$$\begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}$$

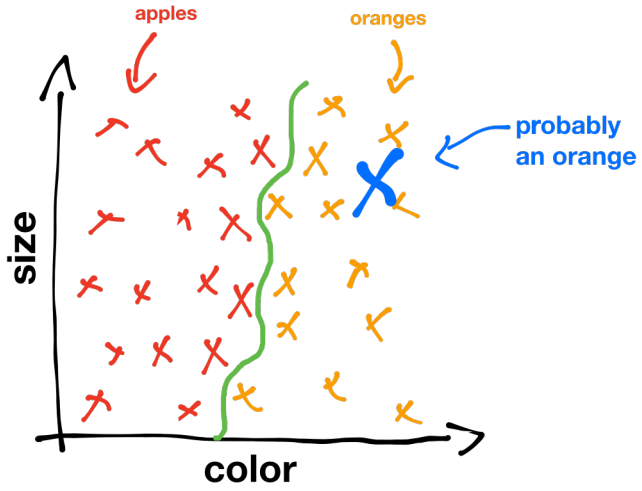
步骤四：学习 (训练)

Partition the feature space into 2 regions, one for each type of fruit

- decision boundary

步骤五：测试、评估

How to handle new images?



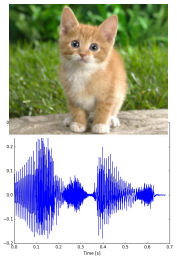
Classifying a new image using the trained classifier .

How to measure the classifier performance?

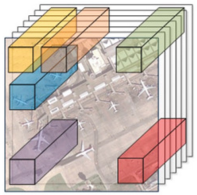
- classification error rate
  - % patterns that are assigned to the wrong category
- other aspects may be important too
  - e.g. computational complexity
  - e.g. user-friendliness

完整流程

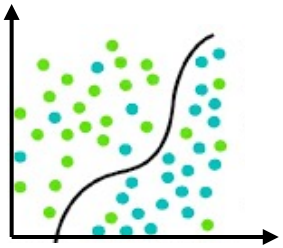
数据准备 | 训练 | 评估



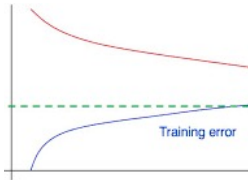
数据收集



特征提取

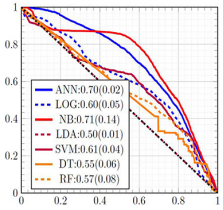


训练



验证

预测

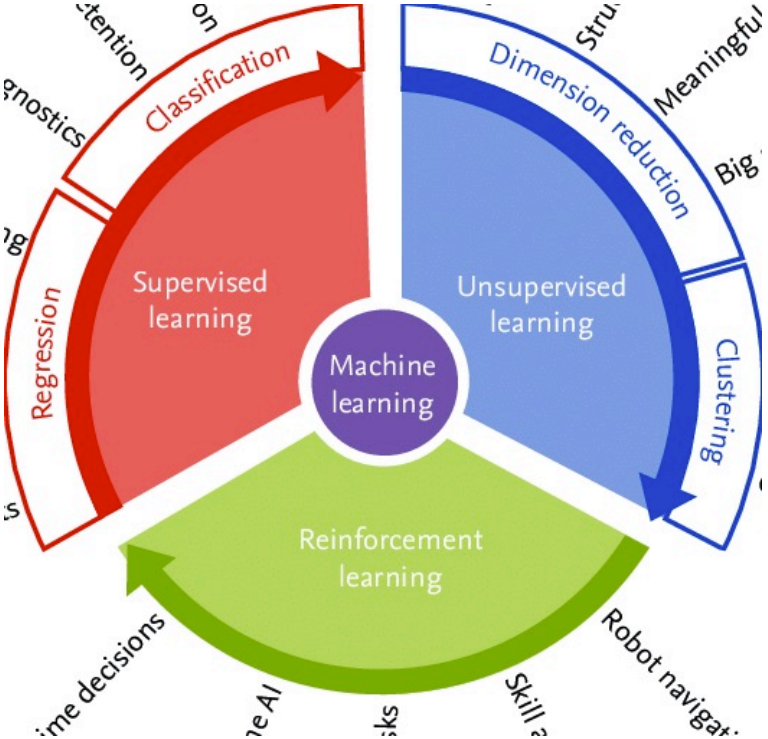


评估

机器学习算法总览

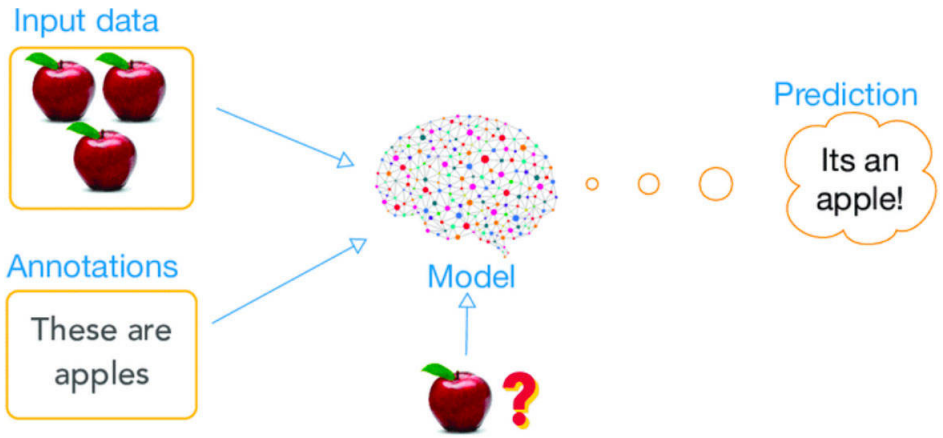
机器学习分类

- 监督学习
- 无监督学习
- 强化学习



监督学习

- 学习器的每个输入样本都有一个目标输出标签.



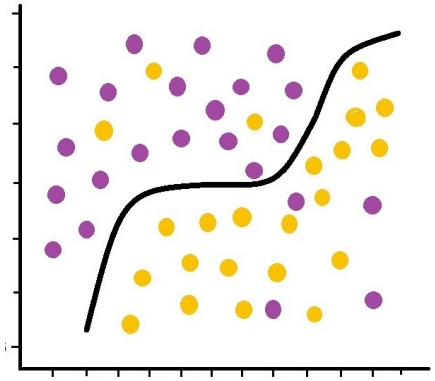
- 有 “老师”的学习

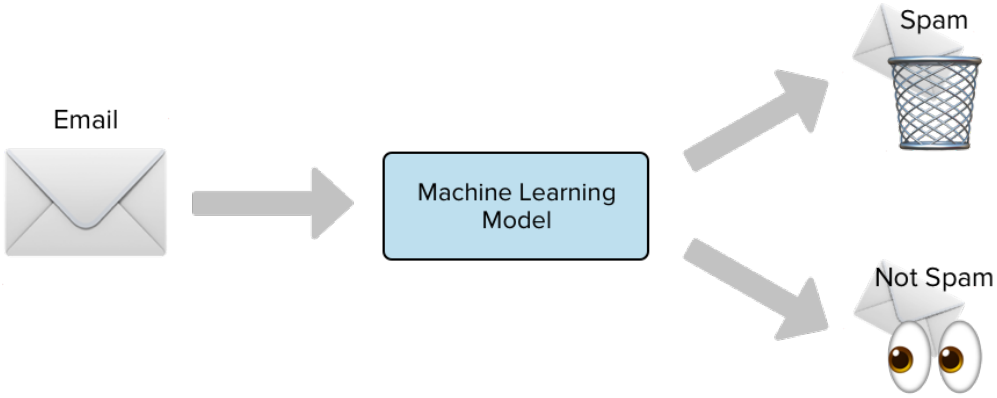
举例

- 老师教学生识别各种类型的动物。
- 练习题提供标准答案.

监督学习

分类问题

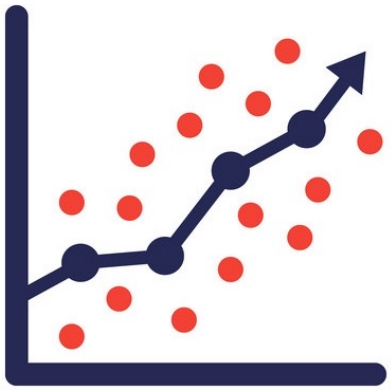


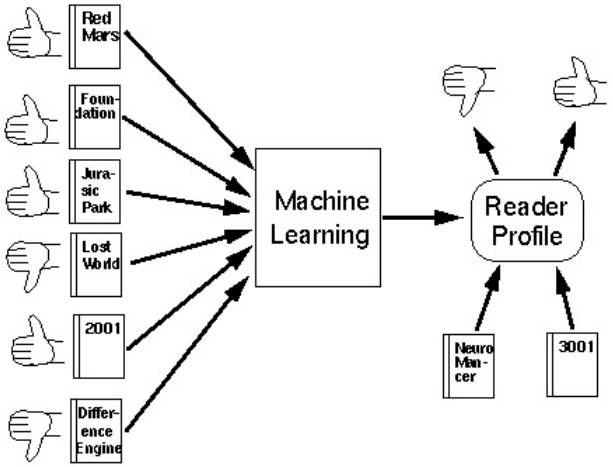


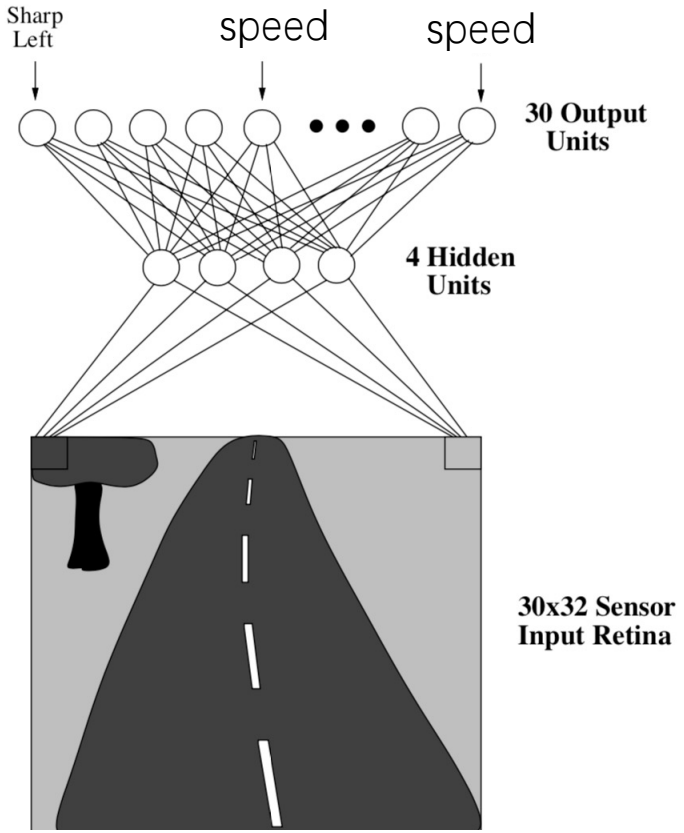
Sentiment analysis

监督学习

回归问题

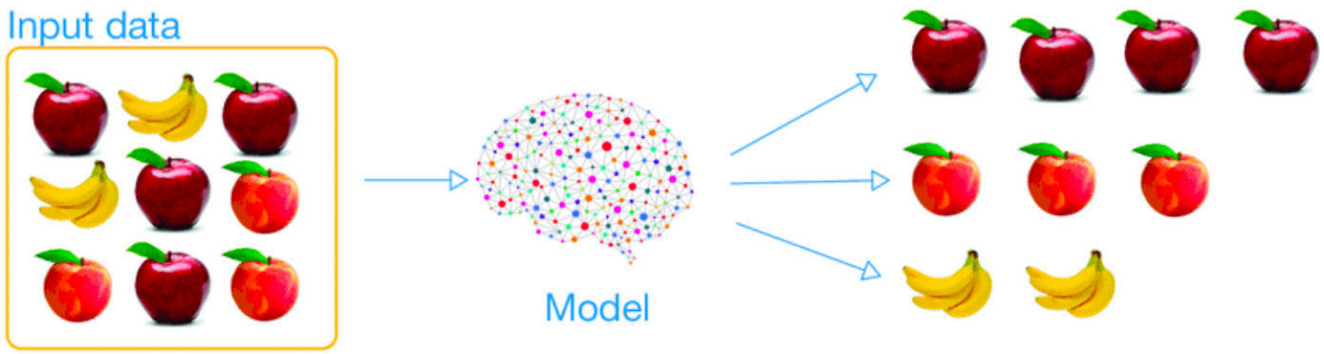



Red Mars
Foun-dation
Jura-sic Park
Lost World
2001
Differ-ence Engine
Machine Learning
Reader Profile
Neuro Man-cer
3001


Sharp Left
speed
speed
30 Output Units
4 Hidden Units
30x32 Sensor Input Retina

无监督学习

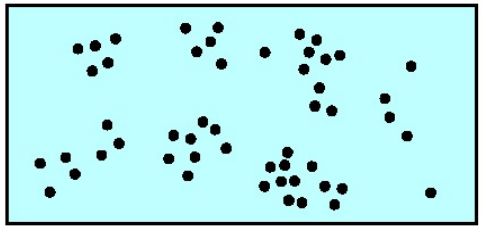
训练样本只作为数据特征输入，没有相对应的标签.



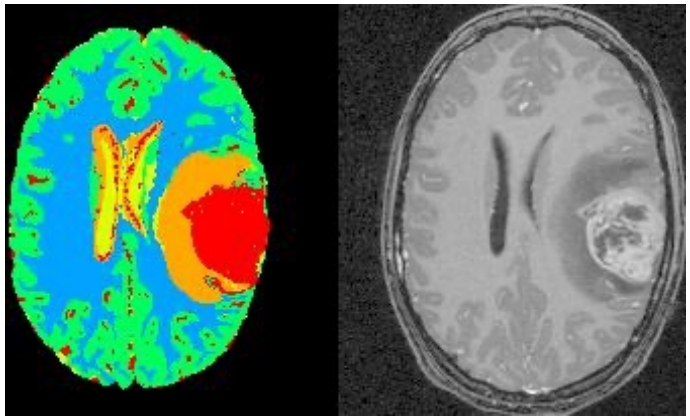
- 无老师监督
- 例：相似数据分组、聚类

无监督学习

聚类



- in the early stages of an investigation, it may be helpful to perform exploratory data analysis to gain some insight into the nature or structure of the data.

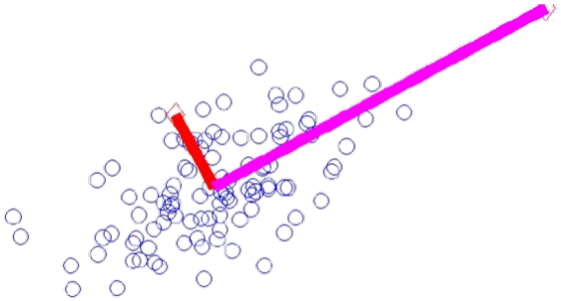


无监督学习

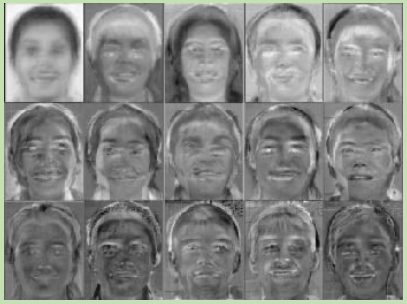
降维和特征选择

- 选择关键的属性和特征，降低数据维度

例：主成分分析 (PCA)

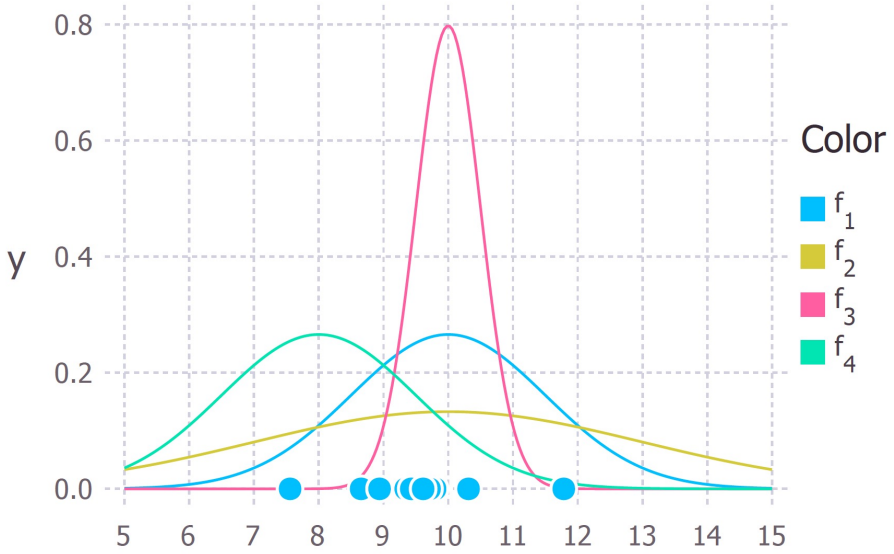


例子 (特征脸)



无监督学习

数据的概率密度估计



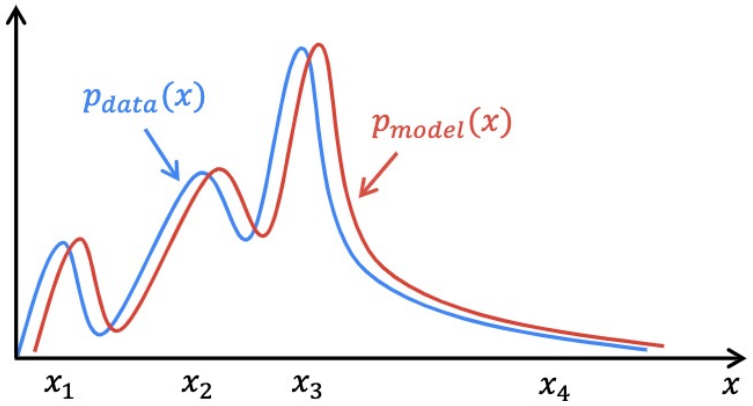
Applications

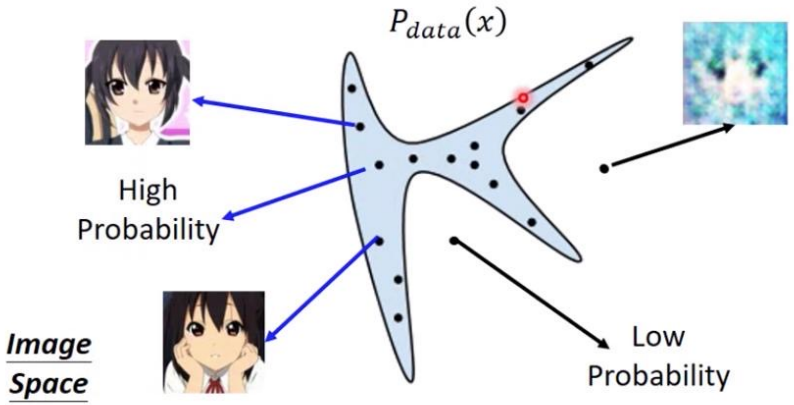
- Density estimation
- Data generation
- Abnormal Detection

无监督学习

数据生成

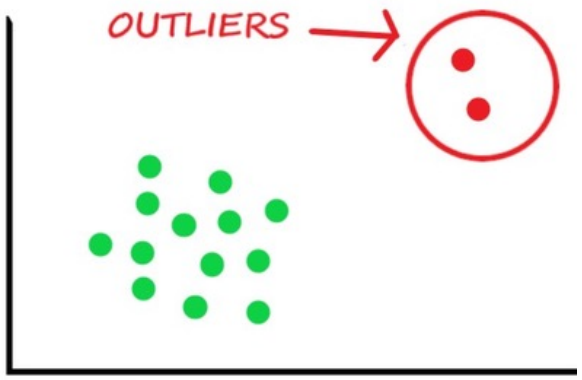
寻找模型构造的概率分布 $p_{model}(x)$ 近似真实数据分布 $p_{data}(x)$.





无监督学习

异常检测 – find the least likely observations from a dataset.



Example: network intrusion detection

NETWORK SECURITY FUNCTION	INGRESS	EGRESS	EAST-WEST
Web Application Firewall (WAF)	✔	-	**
Intrusion Detection/ Prevention (IDS/IPS)	✔	✔	✔
AntiVirus Detection/ Blocking	✔	✔	**
URL/FQDN Filtering (includes Explicit and Category based profiles)	-	✔	✔
Data Loss Prevention (DLP)	-	✔	✔
Layer7 DoS	✔	-	✔
Malicious IP Blocking	✔	✔	-
GeoIP Blocking	✔	-	**
Threat Packet Captures	✔	✔	✔

监督学习 vs. 无监督学习

监督学习

数据: (x, y)
x表示数据, y表示标签

学习目标: 学习函数映射
$x \rightarrow y$

典型任务: 分类、回归、目标检测、语义分割等

无监督学习

数据: x
x表示数据, 无标签

学习目标: 学习数据隐含或潜在的结构

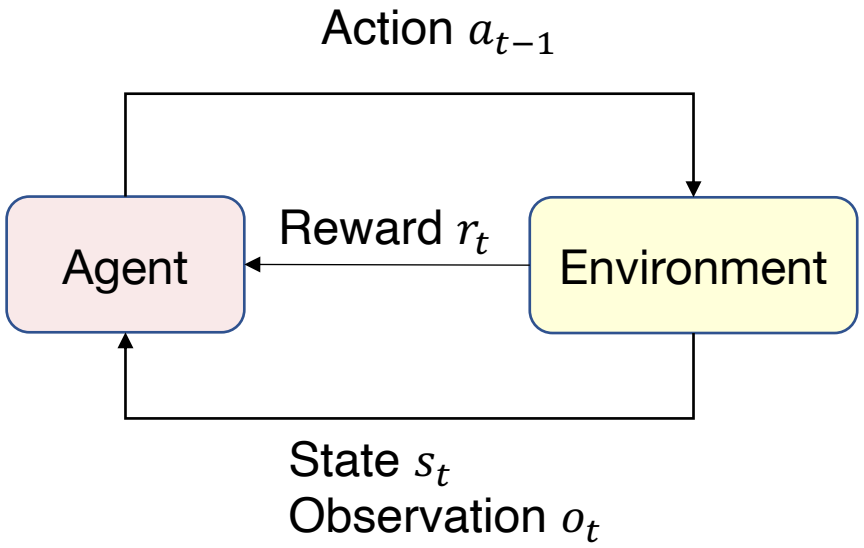
典型任务: 聚类、降维、概率密度估计等

强化学习

- 通过与环境交互来学习

pain

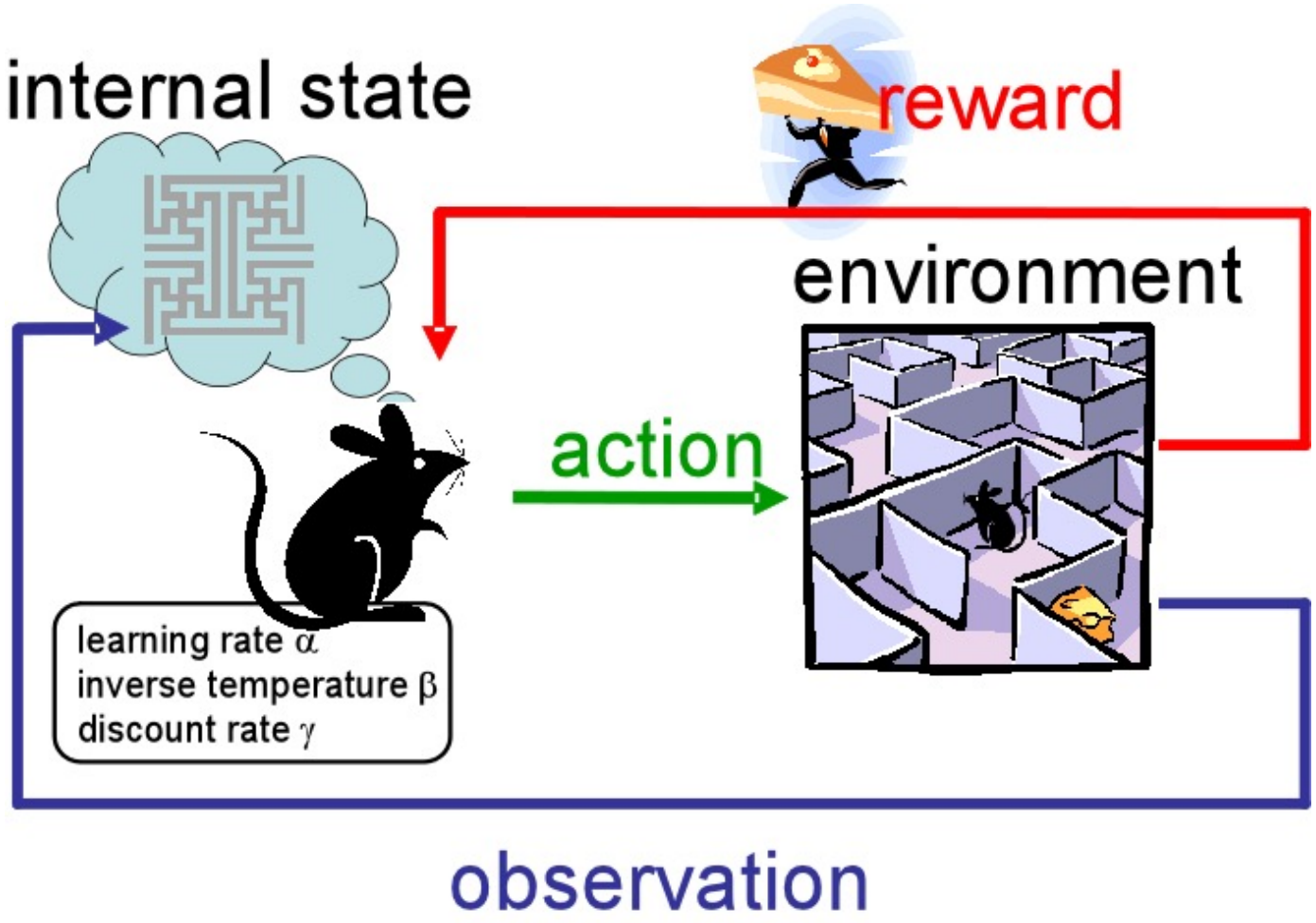
compliment



学习状态(states)到动作(actions)的映射，以最大化长期奖励(reward)。

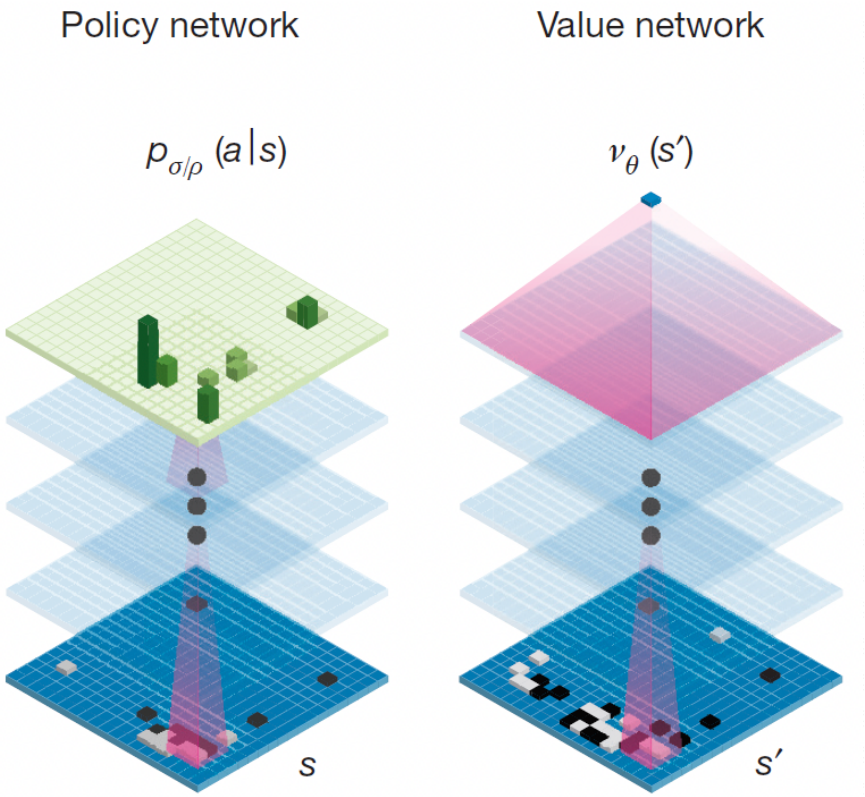
举例: 练习题只给总分不给参考答案(刷题模式)

举例


internal state
reward
environment
action
learning rate α
inverse temperature β
discount rate γ
observation

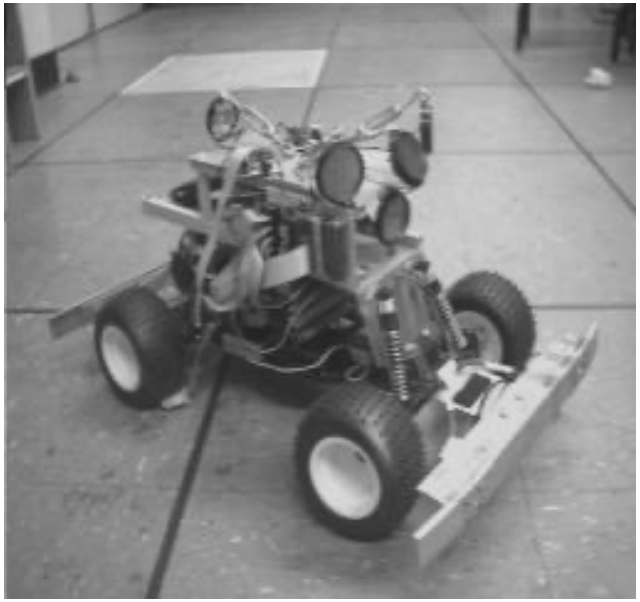
强化学习

举例: AlphaGo



强化学习

举例: 机器人



- **输入**: 传感器信号和增强信号
- **输出**: 避免负反馈争取正的奖励 (避免障碍物, wall following, etc.)

学习 or 记忆？

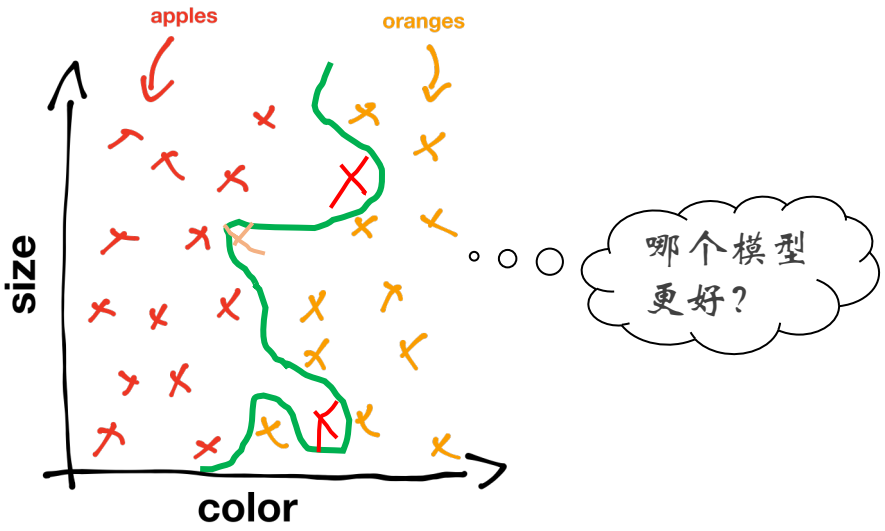
浅谈泛化性与过拟合

泛化性

学习后的分类器能否处理没有见过的水果？

这是苹果吗？

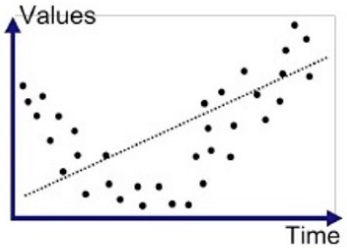
学习的泛化性问题: 模型应该学习普遍的规律还是记住特定细节?



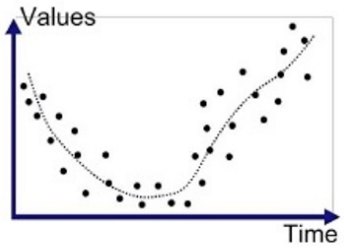
欠拟合 vs 过拟合

欠拟合 – 因模型不够复杂无法刻画真实规律，模型假设可能不成立

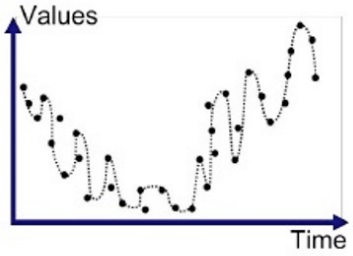
过拟合 – 因模型过于复杂导致记住数据的细节(如随机噪声)而不是潜在的规律



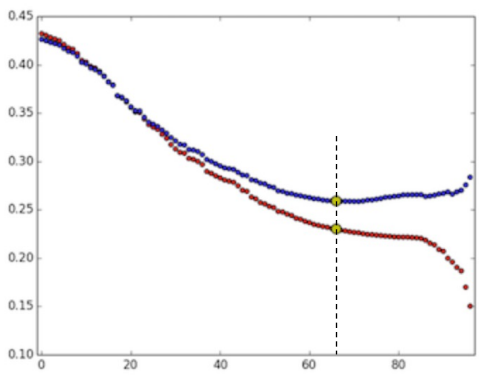
Underfitted



Good Fit/Robust



Overfitted



过拟合

现实案例: (驾考宝典)

1. 机动车遇到行人时应 A. 快速通过 B. **减速通过** C. 正常行驶
2. 在如图所示的场景中，驾驶员应 A. 停车等待 B **减速让行** C.
3. 遇到下图所示的情况，下面哪个操作正确？ A.
4. …

三短一长选最长

非正常行驶的场合，应选择谨慎的选项

- 题中有"**应**"字样的选B,
- 有"**操作**"字样的选C,
- 最长的题目选D,
- 看图题有女孩的选A,没人的选C

避免过拟合

选择合理复杂度的模型

常见的避免过拟合方法

- 增加训练数据
- 正则化（惩罚模型复杂度）
- 数据划分 & 交叉验证（训练时用未知数据测试确保泛化能力）
- 早停
- 引入先验知识（如贝叶斯先验）
- …

措施一：数据划分

模型拟合的是总体数据的分布而不仅仅是训练集的分布.

- 为了评估泛化误差，需要在训练的时候保留一部分未知数据。
- 数据划分 (Hold-Out):
  - 训练集 (e.g., 50%) 平时练习
  - 验证集 (e.g., 25%) 模拟考试
  - 测试集 (e.g., 25%) 高考!

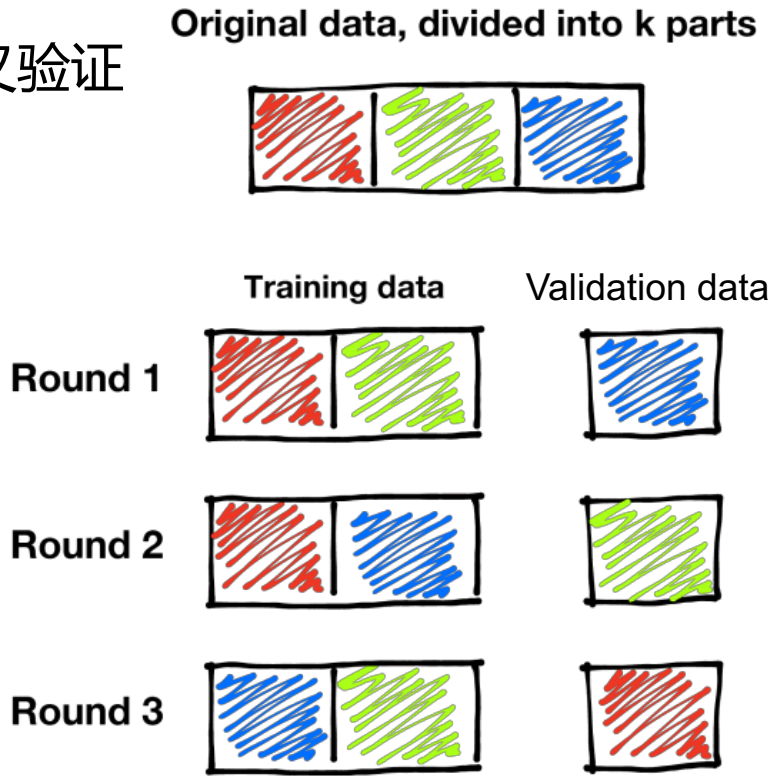
In general, design a good loss function to indicate the performance of the predictive model over the whole-data distribution instead of the training data can help achieve compromise between simplicity and complexity of the model structure.

措施二：交叉验证 (Cross Validation)

如何既划分数据，又能充分利用数据训练？

(特别是当数据规模较小的时候)

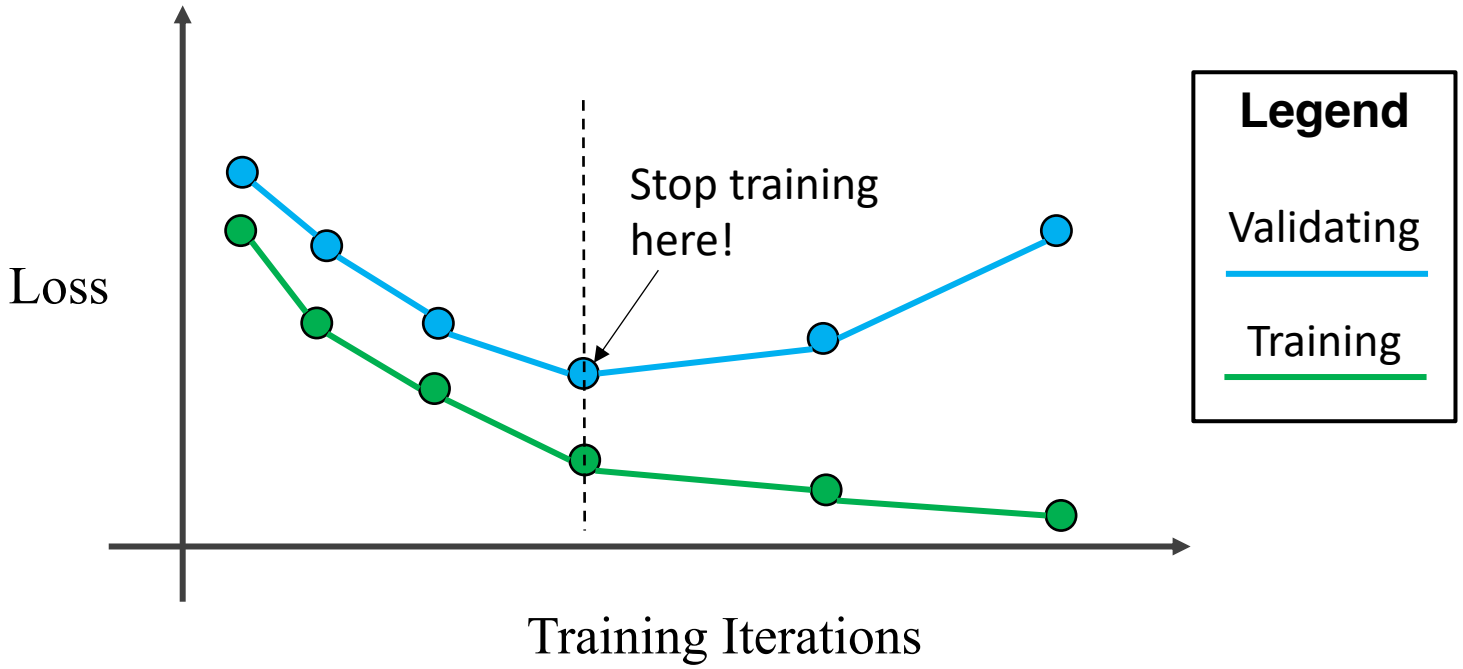
- K-折交叉验证



- Allow for more train-test splits
- A better indication of how well your model performs on unseen data
- Takes more computational power and time

措施三: 早停 (Early Stop)

- 在有过拟合迹象时及时停止训练

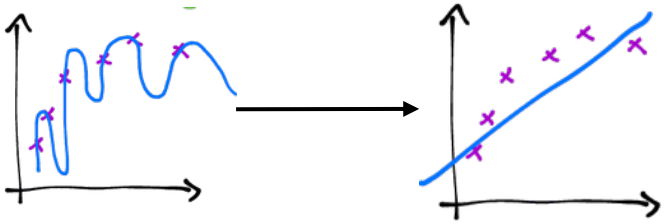


措施四: 正则化 (Regularization)

- 对参数增加惩罚项防止模型过拟合训练数据。

$$L(\mathbf{W}, \lambda_1, \lambda_2) = ||y - \mathbf{W}X||^2 + \lambda_2||w||_2^2 + \lambda_1||w||_1$$

在损失函数中惩罚模型复杂度



No free lunch 理论

- 没有通用的最好的模型
- 不同模型需要适配数据特征和应用

下节预告

线性回归

- 简单而典型的机器学习算法
- 类似最小二乘法