机器学习-概述

万永权
信息技术学院

目录

CONTENTS

什么是机器学习

表 1.1　西瓜数据集

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

对没见过的瓜, 例如“(色泽=浅白) ∧ (根蒂=蜷缩) ∧ (敲声= 浊响)”

机器学习的定义

- 目前为止，尚未有一个公认的机器学习定义。
- Simon认为：如果一个系统能够通过执行某种过程而改进它的性能，这就是学习。
- Minsky认为：学习是在人们头脑中（心理内部）进行有用的变化。
- Tom Mitchell在《机器学习》一书中对学习的定义是：对于某类任务T和性能度量P，如果一个计算机程序在T上以P衡量的性能随着经验E而自我完善，那么，我们称这个计算机程序从经验E中学习。
  - **经验：数据和常识；**
  - **系统：模型或算法；**
  - **性能：准确率或精度等。**

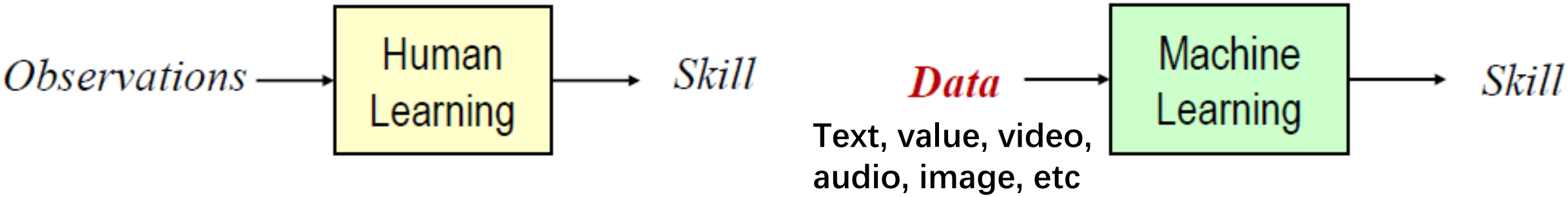
人工智能与机器学习

◆ **人类学习**

人类是从**观察**中积累**经验**来获取**技能**。

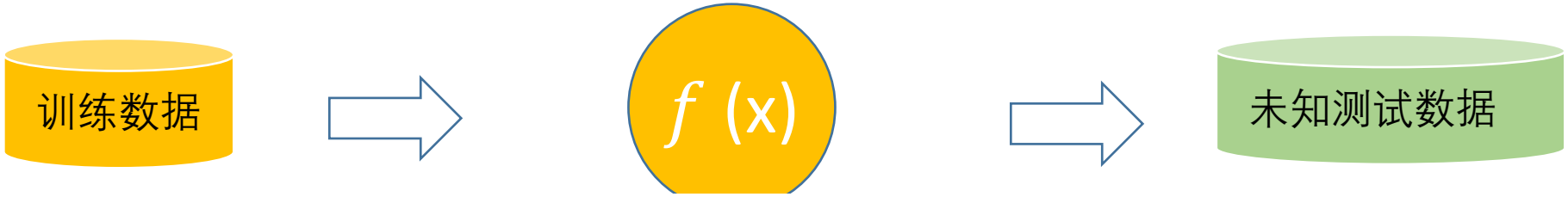
◆ **机器学习**

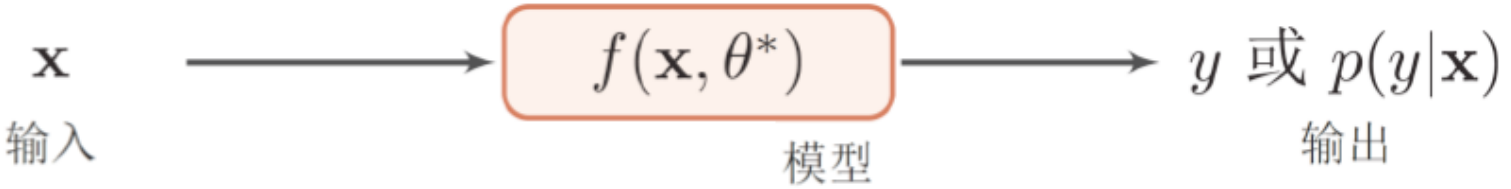
机器是从**数据**中积累或者计算的**经验中**获取技能。



机器模拟人类的学习行为.

机器学习≈构建一个映射函数





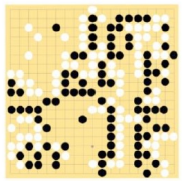
- 语音识别

$f($  $)=$ “你好”

- 图像识别

$f($  $)=$ “9”

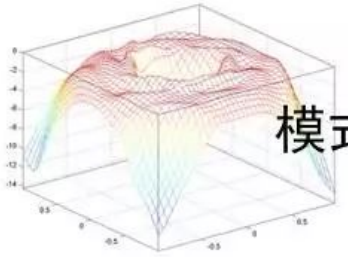
- 围棋

$f($  $)=$ “6-5” （落子位置）

- 机器翻译

$f($ “你好！” $)=$ “Hello!”

机器学习的应用



模式识别

计算机视觉

数据挖掘

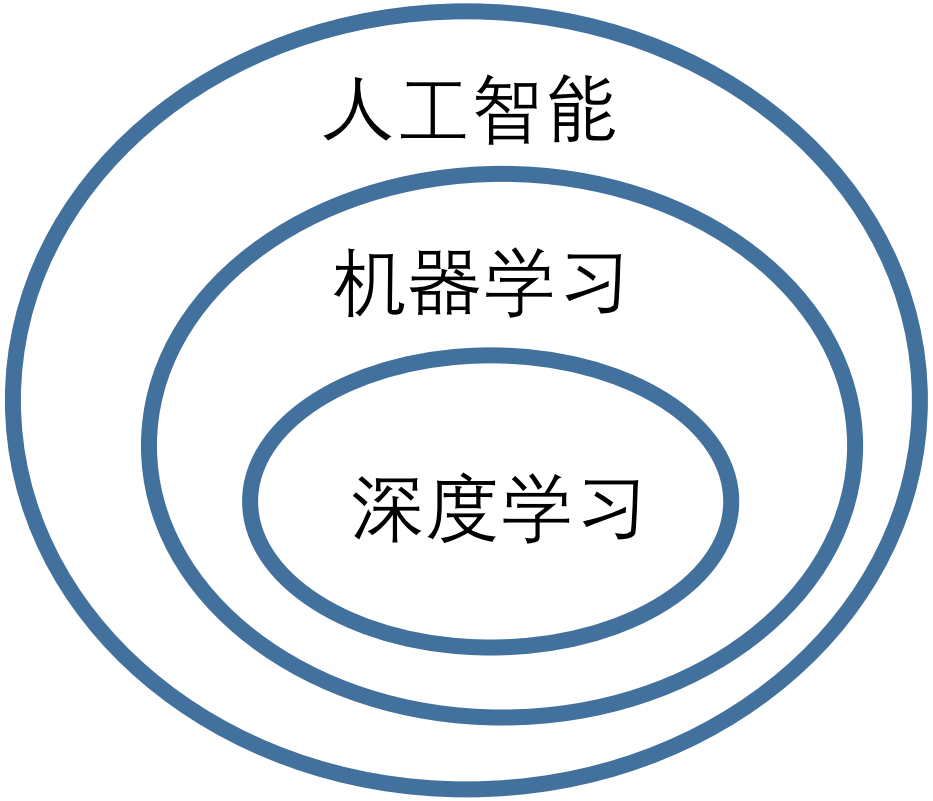
机器学习

语音识别

统计学习

自然语言处理

AI、Machine Learning (ML)、Deep Learning (DL)的关联



成功实现AI应用的三要素:

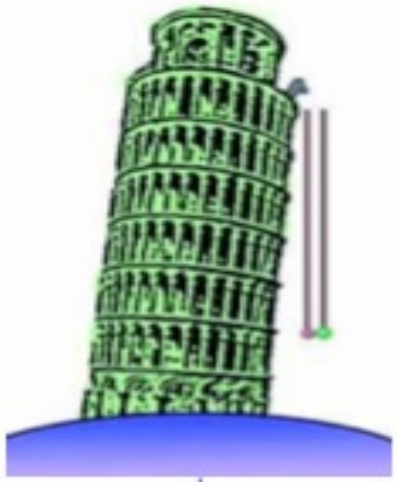
- ◆算法（菜谱）
- ◆算力（厨具）
- ◆数据（食材）

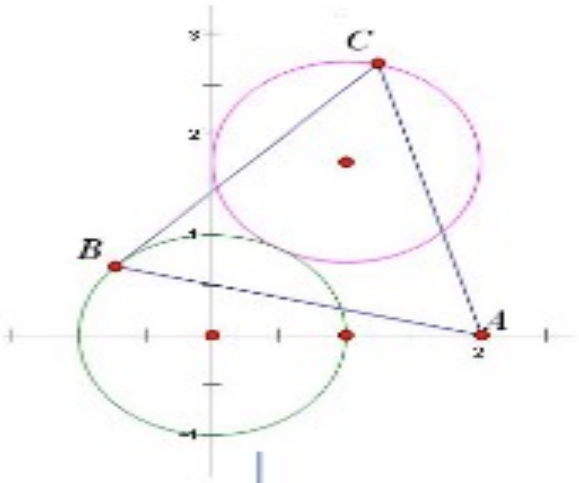
深度学习=大**数据**+高性能**计算**+灵巧的**算法**

大数据与人工智能

大数据对科学研究的影响

图灵奖获得者、著名数据库专家Jim Gray 博士观察并总结人类自古以来，在科学研究上，先后历经了实验、理论、计算和数据四种范式





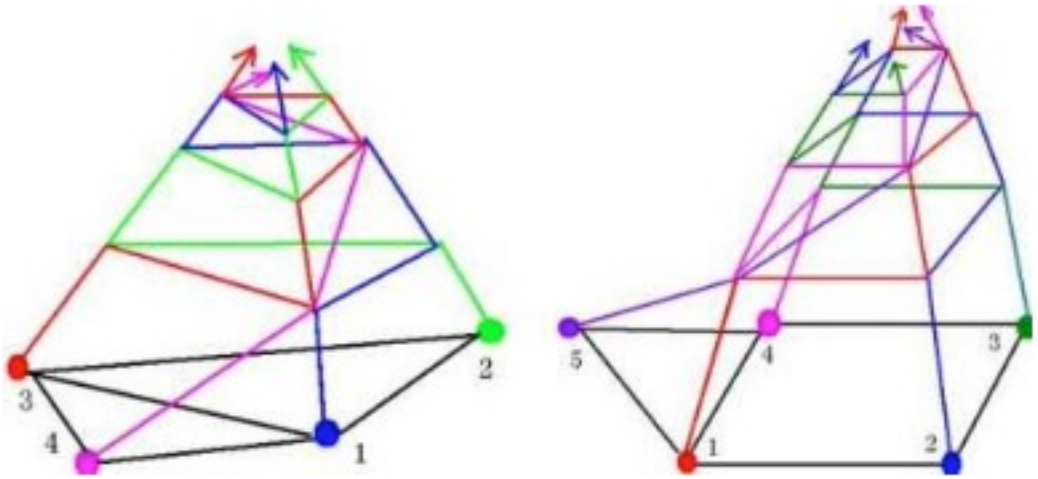
实验　理论　计算　数据

科学研究第一种范式：实验

伽利略

伽利略在比萨斜塔做两个铁球同时落地实验

科学研究第二种范式：理论



几何理论

牛顿三大定律

科学研究第三种范式：计算

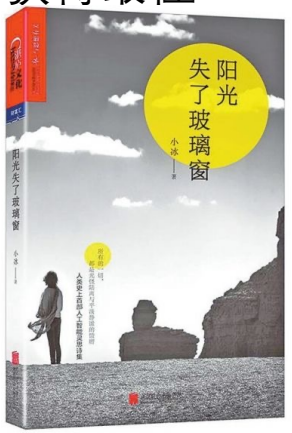
科学研究第四种范式：数据

大数据时代，以数据为中心

案例：微软小冰成功预测奥斯卡奖得主

- 2013年，微软纽约研究院的经济学家大卫·罗斯柴尔德（David Rothschild）利用大数据成功预测24个奥斯卡奖项中的19个，成为人们津津乐道的话题。
- 2016年罗斯柴尔德再接再厉，成功预测第86届奥斯卡金像奖颁奖典礼24个奖项中的21个，继续向人们展示现代科技的神奇魔力。
- 在去年奥斯卡颁奖前夜，微软小冰在微博上发布了自己的四项预测，认为莱昂纳多·迪卡普里奥(又被中国影迷称为“小李子”)有73%的概率会获得最佳男主角。

“*微软小冰*”：一款人工智能机器人



数据产生方式的变革促成大数据时代的来临

图 数据产生方式的变革

2 大数据的定义

- 自2012年以来，“大数据”一词越来越引起人们的关注。
- 在维克托·迈尔-舍恩伯格（Viktor Mayer-Schönberger）编写的《大数据时代》中，大数据指不用随机分析法（抽样调查）这样捷径，而采用所有数据进行分析处理。
- 麦肯锡全球研究所则定义大数据为一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低 （4V）四大特征。
- 通常来说，大数据是指数据量超过一定大小，无法用常规的软件在规定的时间范围内进行抓取、管理和处理的数据集合。

大数据思维

维克托·迈尔-舍恩伯格：

- 需要全部数据样本而不是抽样
- 关注效率而不是精确度
- 关注相关性而不是因果关系

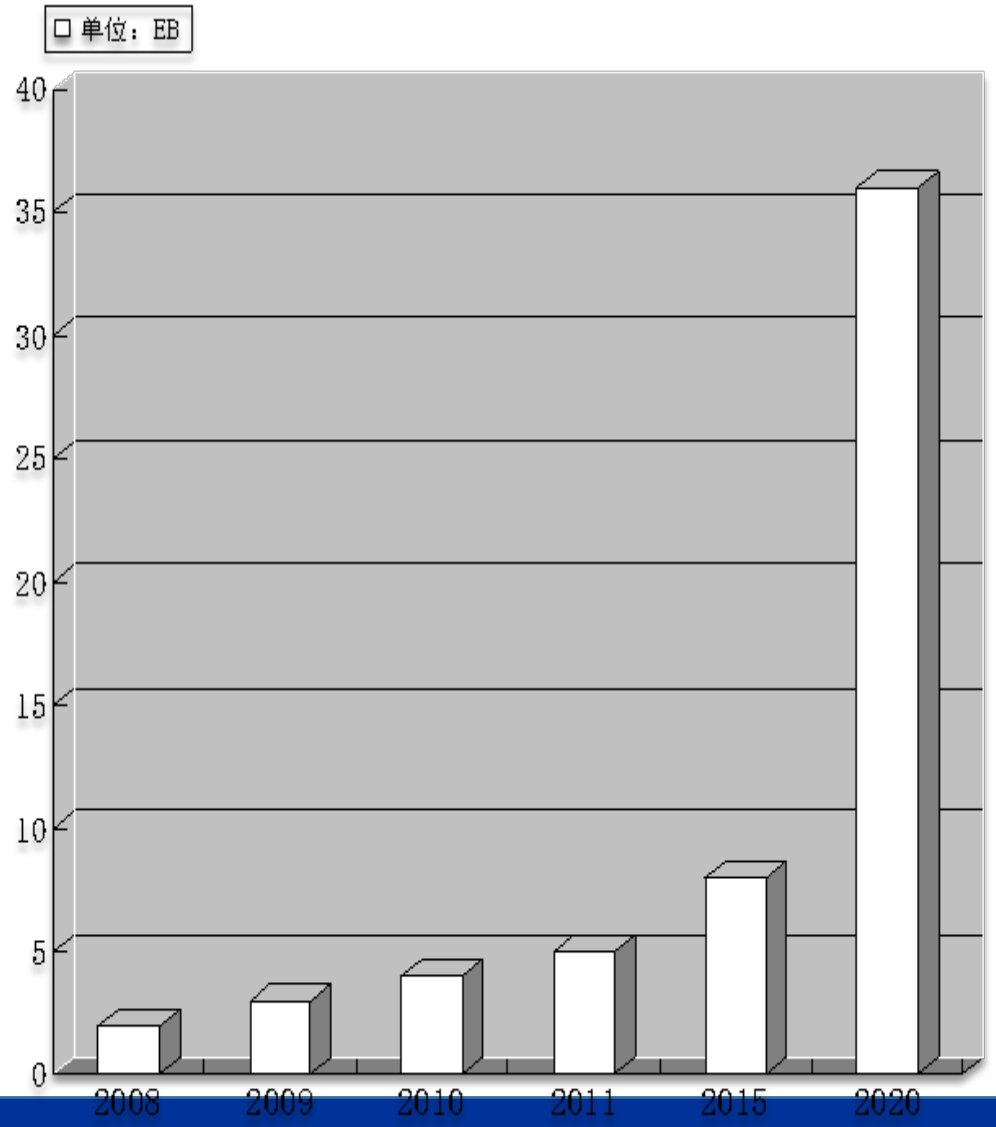
数据科学的技术权威

大数据的特征

- 大数据的主要特征可用“5V+1C”来进行概括:
  - 数据量大（Volume）
  - 数据类型多（Variety）
  - 数据时效性强（Velocity）
  - 价值密度低（Value）
  - 准确性高（Veracity）
  - 复杂性高（Complexity）

特征1：数据量大

- 大数据的起始计量单位最少是PB级以上。1024MB=1GB；1024GB=1TB；1024TB=1PB；1024PB=1EB；1024EB=1ZB。
- 根据国际数据公司（IDC）的《数据宇宙》报告显示：2020年全球数据总量将超过40ZB（相当于4万亿GB），这一数据量是2011年的22倍。



特征2：数据类型多

- 从数据组织形式的角度来看，数据类型可以简单地被分为：
  - 结构化数据：能够用数据或统一的结构加以表示，如数字、符号等，是传统的关系数据模型、行数据，存储于数据库。包括银行交易数据、商品购买信息数据等格式严谨的数据库数据。
  - 非结构化数据：是指那些无法通过事先定义的数据模型表达或无法存入关系型数据库表中的数据，例如办公文档、图片、音频和视频等。
  - 半结构化数据：介于完全结构化数据和非结构化数据之间，XML、HTML文档就属于半结构化数据。它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。

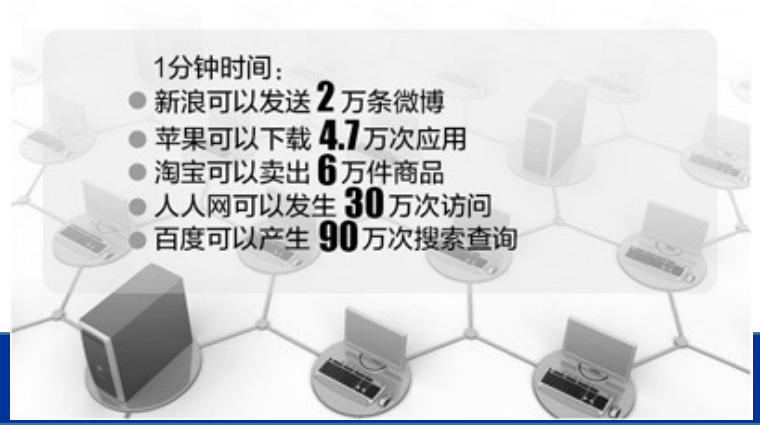
大数据环境下的数据类型繁多。在早期，绝大部分的数据信息是以结构化的表形式存放在数据库中。这些数据处理起来比较方便，但是，随着计算机技术的快速发展，大数据环境下，半结构化数据和非结构化数据在整个数据量中所占的比例大幅度上升。据统计，在企业数据中，目前已有超过80%的数据是以非数据结构化的形式存在的，结构化数据仅仅占20%不到。多类型的数据对数据的处理能力提出了更高的要求。值得注意的是，由于非结构数据占据了大数据的统治地位，而其所蕴含了无尽的知识和能量，这就要求现代数据处理技术提出了更高的要求，从算法到架构，以应对非结构化数据增加带来的挑战。

数据类型

数据类型	数据类型的描述
结构化数据	结构化的数据是指可以使用关系型数据库表示和存储，如MySQL、Oracle、SQL Server等，表现为二维形式的数据。数据以行为单位，一行数据表示一个实体的信息，每一行数据的属性是相同的。
半结构化数据	半结构化数据属于同一类实体可以有不同的属性，这些属性可能是数值型的，也可能是文本型的，还可能是字典或者列表。常见的半结构数据有XML和JSON。
非结构化数据	就是没有固定结构的数据。各种文档、图片、视频/音频等都属于非结构化数据。对于这类数据，一般直接整体进行存储，且存储为二进制的数据格式。

特征3：数据时效性强

- 数据时效性高：数据增长速度快，处理速度也快，获取数据的速度也要快。
- 在大数据环境下，随着数据量的剧增和数据类型逐渐多样化，数据中所隐藏的高时效性特征显得越来越突出。
- 在这样的背景下，企业必须要实时分析所拥有的最新数据，并提取其中有价值的信息，以产生对未来的生产具有指导意义的分析结果。
- 例如，在台风天气中，气象部门应实时汇报台风过境前后的路径走向。这就需要相关技术部门随时收集某一刻最新的台风路径数据进行分析，并及时做好应对措施。



特征4：价值密度低

- 随着物联网的广泛应用，信息感知无处不在，产生海量数据，但这些数据价值密度较低，如何通过强大的机器学习算法迅速地完成数据的价值“提纯”，是大数据时代亟待解决的难题。
- 另一方面，在Value这个层面，大数据要求我们处理的数据集是有巨大商业价值或社会价值的。
  - 阿里巴巴愿意花巨大代价提高推荐系统的准确性，就是在于其推荐系统的准确率的提高，能大大提高平台的交易量，从而具有非常巨大的商业价值。
  - 我们在全国部署“天眼”系统，提高大数据技术在天眼系统的分量，就是因为天眼系统分析能力的一小步提升，都能在降低犯罪率、打击犯罪、保障人民群众安全、信用取证等方面都有巨大的社会价值。

特征5：准确性高

- 准确性是指数据的可信赖度，即数据的质量。
- 大数据中的内容是与真实世界中的发生息息相关的，研究大数据就是从庞大的网络数据中提取出能够解释和预测现实事件的过程，通过大数据的分析处理，最后能够解释结果和预测未来。
- 在小数据时代，由于小数据集搜集数据比较困难，因而在分析数据时往往更着重于分析方法，这会不可避免地产生一些主观偏差，准确性不高。
- 大数据时代，通过技术手段分析全部数据，准确性大大提高。

特征6：复杂性高

- 复杂性是指数据本身的复杂性、计算的复杂性和信息系统的复杂性。
  - 数据本身的复杂性：表现在图文检索、主题发现、语义分析、情感分析等数据分析工作十分困难，其原因是大数据涉及复杂的类型、复杂的结构和复杂的模式，数据本身具有很高的复杂性。
  - 计算机的复杂性：表现在大数据计算不能像处理小样本数据集那样做全局数据的统计分析和迭代计算，在分析大数据时，需要重新审视和研究它的可计算性、计算复杂性和求解算法。
  - 系统的复杂性：表现在大数据对计算机系统的运行效率和能耗提出了苛刻要求，大数据处理系统的效能评价与优化问题具有挑战性。

大数据与人工智能

- 人工智能实现最大的飞跃是大规模并行处理器的出现，特别是GPU，它是具有数千个内核的大规模并行处理单元，而不是CPU中的几十个并行处理单元。这大大加快了现有的人工智能算法的速度。
- 人工智能应用的数据越多，其获得的结果就越准确。在过去，人工智能由于处理器速度慢、数据量小而不能很好地工作。也没有先进的传感器，并且当时互联网还没有广泛使用，所以很难提供实时数据。如今，人们拥有所需要的一切：快速的处理器、输入设备、网络和大量的数据集。
- 没有大数据就没有人工智能。

机器学习的基本术语

机器学习的基本术语

1. 数据集（Dataset）

数据集是指数据的集合。例如（20301001，张三，175cm，70kg）。

2. 样本（Sample）

- 样本也称为实例（Instance），指待研究对象的个体，包括属性已知或未知的个体。
- 例如，每个学生所对应的一条记录就是一个“样本”。 数据集即为若干样本的集合。

3. 标签（Label）

- 标签是为样本指定的数值或类别。
- 在**分类**问题中，标签是样本被指定的特定类别;
- 在**回归**问题中，标签是样本所对应的实数值。
- **已知样本**是指标签已知的样本，**未知样本**是指标签未知的样本。

机器学习的基本术语

4. 特征（Feature）

- 特征是指样本的一个独立可观测的属性或特性。
- 它反映样本在某方面的表现或性质，
- 例如“姓名”“身高”是“特征”或“属性”。
- 特征的取值，例如“张三”“175cm”是“特征值”或“属性值”。

5. 特征向量（Feature Vector）

- 特征向量是由样本的n个属性组成的n维向量，第i个样本X_i表示为：

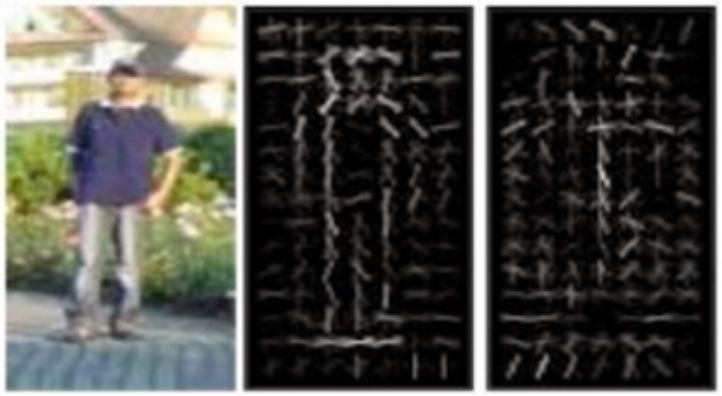
$$(x_{i1} \quad x_{i2} \quad \dots \quad x_{in})$$

- **特征分为**
  - **手工式特征**也称为**设计式特征**,是指由学者构思或设计出来的特征, 如SIFT、HOG等。
  - **学习式特征**是指由机器从原始数据中自动生成的特征。
    例如，通过卷积神经网络获得的特征就属于学习式特征。

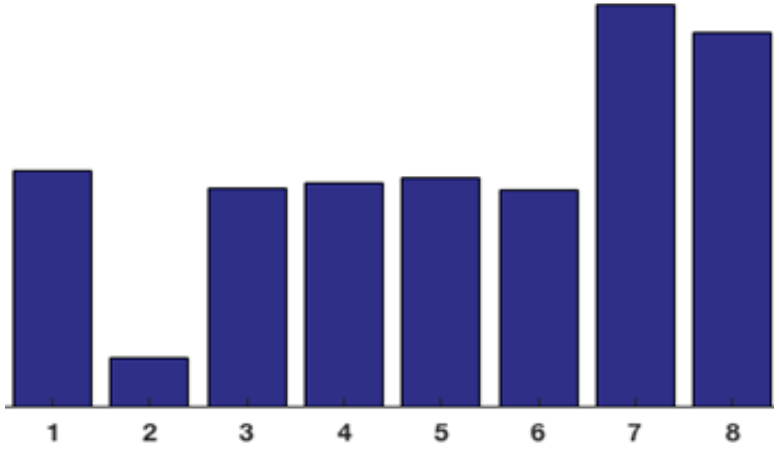
手工式特征 / 设计的特征

例如:

- 用边缘检测算子提取的边缘特征。
- HOG (Histogram of Oriented Gradients, 定向梯度直方图)

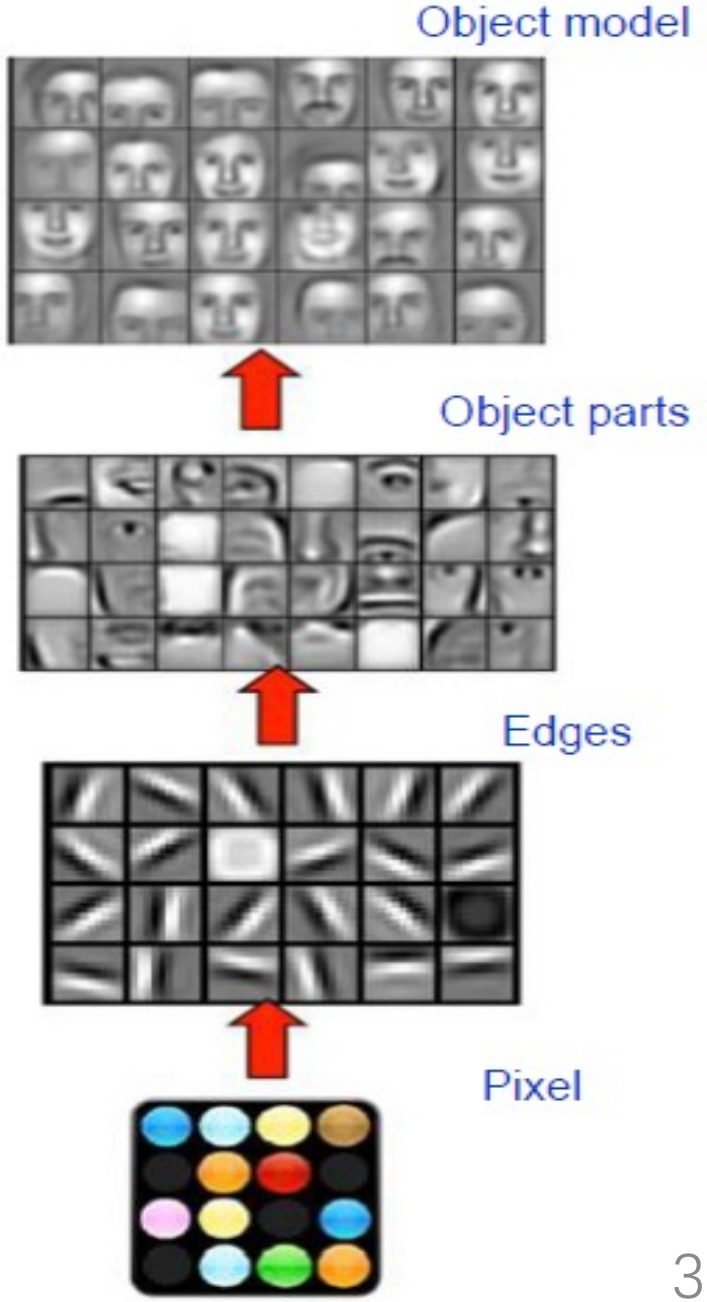


边缘特征



HOG特征

学习式特征



4.1.2 机器学习的基本术语

通常将数据集分成训练集、验证集和测试集，需要保证这三个集合是不相交的。

（1）训练集（Training Dataset）

- 训练过程中使用的数据称为“**训练数据**”，其中每个训练数据称为一个“**训练样本**”；
- 每个训练样本都有一个已知标签，由所有训练样本及其标签组成的集合称为“**训练集**”。
- 训练集包括**一个样本集**和**一个对应的标签集**，用于学习得到拟合样本的模型。
- 一般地，训练集中的标签都是正确的，称为**真实标签**（Ground-Truth）。
- 例如，在图像分类任务中，训练集包括
  - 一个由特定图像组成的样本集合
  - 一组由语义概念（如山、水、楼等）组成的标签集合，标签即为Ground-Truth。

4.1.2 机器学习的基本术语

（2）验证集（Validation Dataset）

- 在实际训练中，有时模型在训练集上的结果很好，但对于训练集之外的数据的结果并不好。
- 此时，可单独留出一部分样本，不参加训练，而是用于调整模型的超参数，并对模型的能力进行初步评估，这部分数据称为**验证集**。
- **超参数**（hyperparameter）是指模型中人为设定的、无法通过训练得到的参数，如NN的层数、卷积的尺寸、滤波器的个数、KNN和K-Means算法中的K值等。

（3）测试集（Test Dataset）：

- 测试过程中使用的数据称为“**测试数据**”，被预测的样本称为“**测试样本**”，测试样本的集合称为“**测试集**”。
- 测试集不参与模型的训练过程，仅用于评估最终模型的**泛化能力**。

4.1.2 机器学习的基本术语

7. 泛化能力（Generalization Ability）

泛化能力是指训练得到的模型对未知样本正确处理的能力，即模型对新样本的**适应能力**，亦称为**推广能力**或**预测能力**。

8. 模型参数

- 给定训练集，希望能够拟合一个函数 $f(x,\theta)$来完成从输入的特征向量到标签的映射。
- 对于连续的标签或非概率模型，通常会采用拟合函数来表示从输入空间（样本集X）到输出空间（标签集Y）的映射： $Y' = f(x,\theta)$

  其中，Y' 是样本x的**预测标签**，θ为模型中可训练得到的参数，即**模型参数**，也称为**学习参数**，**并非是由人为设置的超参数**。

4.1.2 机器学习的基本术语

9.学习算法（learning algorithm）

- 希望为每个样本x**预测的标签**与其所对应的**真实标签**都相同，这就需要有一组好的模型参数θ。
- 为了获得这样的参数θ，则需要有一套学习算法来优化函数f，此优化过程称为**学习**(Learning)或者**训练**(Training)，拟合函数f称为**模型**（Model）。

10. 假设空间（hypothesis space）

- 从输入空间至输出空间的映射可以有多个，它们组成的映射集合称为**假设空间**。
- **学习的目的：**在此假设空间中选取最好的映射，即**最优的模型**。
- 用训练好的最优模型对测试样本进行预测的过程称为**测试**。

4.1.2 机器学习的基本术语

11. **损失函数（Loss function），**也称为**代价函数**（Cost Function），用于度量**预测标签**和**真实标签**之间差异或损失。

真实标签集表示为Y，**预测标签集**表示为$Y'=f(X)$，则损失函数记为$L(Y,f(X))$，是一个**非负**的实值函数。常用的损失函数包括:

- **0-1**损失函数公式为: $L(Y,f(X)) = \begin{cases} 0, & \text{if } Y = f(X) \\ 1, & \text{if } Y \neq f(X) \end{cases}$
- **平方**损失函数公式为: $L(Y,f(X)) = \frac{1}{2}(Y-f(X))^2$ **平方误差损失也称为L2损失。**
- **绝对**损失函数公式为: $L(Y,f(X)) = |Y-f(X)|$, **绝对误差也称为L1损失。**
- **对数**损失函数公式为: $L(Y,f(X)) = -logP(Y|X)$
- **交叉熵**损失函数公式为: $L(Y,f(X)) = -\sum_{c=1}^{C} Y_c \, log f(X_c)$

4.1.2 机器学习的基本术语

12.风险函数（Risk Function）

- 风险函数又称**期望损失**（Expected Loss）或**期望风险**（Expected Risk），是所有数据集（包括训练集和预测集）上损失函数的期望值，用于度量平均意义下模型预测的好坏。
- **机器学习的目标**是选择风险函数最小的模型。

4.1.2 机器学习的基本术语

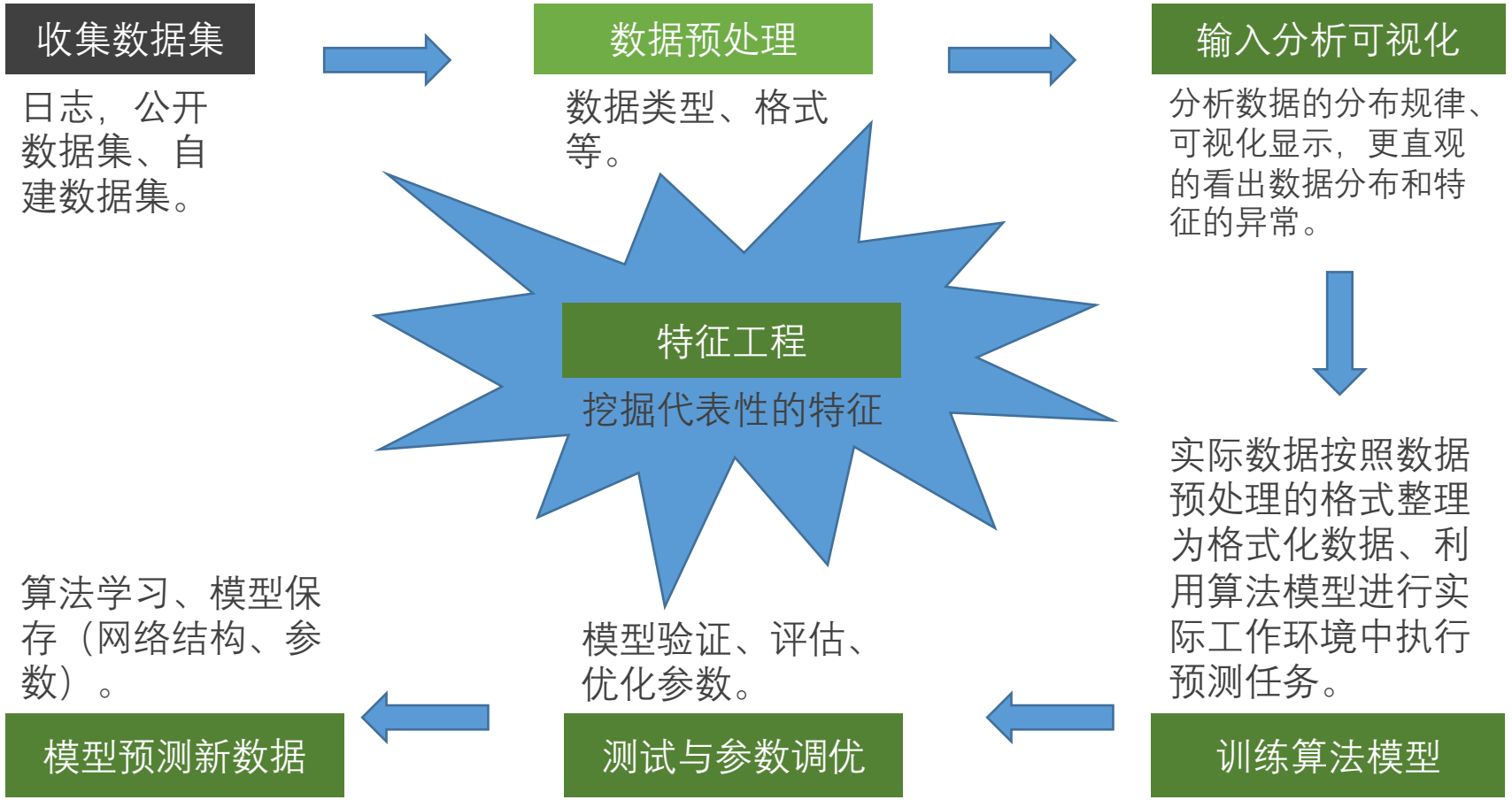
13.优化算法

- 在获得了数据集、确定了假设空间以及选定了损失函数之后，需要解决**最优化**问题。
- 机器学习的训练和学习的过程，就是求解最优化问题的过程，寻找**全局最优解**。
- 若最优化问题存在显式的解析解，则可以很容易求得它的解；
- 但通常不存在解析解，则只能通过数值计算的方法来不断逼近它的解。
- 最简单也最常用的优化算法是**梯度下降法**（Gradient Descent，GD）。
- 梯度下降法通过不断迭代的方式来降低风险函数的值，公式为：$\theta_{t+1} = \theta_t - \eta \frac{\partial R(\theta)}{\partial \theta}$

  其中，θ_t为第t次选代时的参数值，R(θ)为风险函数，η为优化的步长，又称为**学习率**。
  - 学习率过小，会导致学习速度太慢，还可能导致陷入局部最优；
  - 学习率过大，又会出现震荡，严重时会导致发散。

机器学习的一般流程

一般流程


收集数据集
日志，公开数据集、自建数据集。
数据预处理
数据类型、格式等。
输入分析可视化
分析数据的分布规律、可视化显示，更直观的看出数据分布和特征的异常。
特征工程
挖掘代表性的特征
实际数据按照数据预处理的格式整理为格式化数据、利用算法模型进行实际工作环境中执行预测任务。
训练算法模型
测试与参数调优
模型验证、评估、优化参数。
模型预测新数据
算法学习、模型保存（网络结构、参数）。

4.1.2 机器学习的基本术语

14. **机器学习的基本流程**是：**数据预处理→模型学习→模型评估→新样本预测**。

① 数据预处理

收集并处理数据，有时还需要完成数据增强、裁剪等工作，划分训练集、验证集、测试集。

② 模型学习，即模型训练

- 在训练集上运行学习算法，利用损失函数和优化算法求解一组模型参数，得到风险函数最小的最优模型。
- 一般在训练集上会反复训练多轮，即训练样本被多次利用。

数据清洗

- 对各种脏数据进行对应方式的处理，得到标准、干净、连续的数据，提供给数据统计、数据挖掘等使用。

数据的完整性

例如人的属性中缺少性别、籍贯、年龄等；
解决方法：信息补全（使用身份证件号码推算性别、籍贯、出生日期、年龄等）；剔除；

数据的唯一性

例如不同来源的数据出现重复的情况等；
解决方法：按主键去重（用sql或者excel“去除重复记录”）/ 按规则去重（如不同渠道来的客户数据，可以通过相同的关键信息进行匹配，合并去重）

数据的合法性

例如获取数据与常识不符，年龄大于150岁；
解决方法：**设置字段内容**（日期字段格式为“2010-10-10”）；**类型的合法规则**（性别 in [男、 女、未知]）

数据的权威性

例如出现多个来源的数据，且数值不一样；解决方法：为不同渠道设置权威级别，如：在家里，首先得相信媳妇说的。。。

数据的一致性

例如不同来源的不同指标，实际内涵是一样的，或是同一指标内涵不一致；
解决方法：建立数据体系，包含但不限于指标体系、维度、单位、频度等

4.1.2 机器学习的基本术语

③ 模型评估

- 将验证集样本输入到学习获得的模型中，用以评估模型性能，还可以进一步**调节模型的超参数**，找到最合适的模型配置。
- 常用的模型评估方法为K折交叉验证。
- 通常所说的“模型调参”一般指的是调节超参数，而不是模型参数。

④ 新样本预测

- 将测试集中的样本输入到训练好的模型中，对比预测的结果与真实值，计算出各种评价指标，以此来**评价模型的泛化能力**。
- 例如，图像分类任务有精确率（Precision）和召回率（Recall）等评价指标。

机器学习的基本要素

从机器学习的基本流程可知，**学习算法有三个基本要素**：

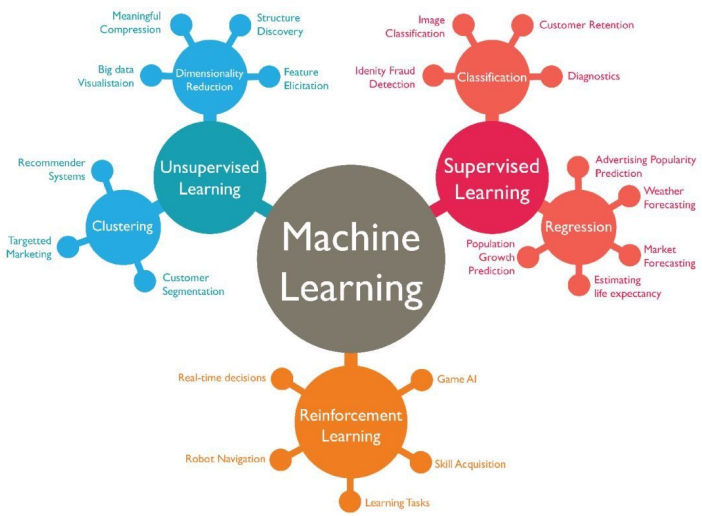
- **模型**（哪一类模型：线性模型、概率模型、非线性模型、网络模型）
- **损失函数（学习准则、学习策略）**：选出什么损失函数来衡量错误的代价，才能找到**最优的模型参数**。
- **优化算法，也称为优化器，**最简单也最常用的优化算法是**梯度下降法。**

这三个要素都需要学者根据经验人为确定。

机器学习的分类

ML的分类

- 监督学习 (supervised learning)
- 无监督学习 (unsupervised learning)
- 强化学习 (reinforcement learning)



ML的分类

1、监督学习

- 监督学习是从给定的训练数据集（简称训练集）中学习一个函数（模型），新的数据（测试集）到来时，可以根据这个函数（模型）进行预测；

- 在监督式学习下，输入数据被称为“训练数据”，每组训练数据有一个明确的类别标签或结果。

- 反垃圾邮件系统的训练数据，带有该邮件的类别，或者是垃圾邮件，或者是非垃圾邮件。

ML的分类

1、监督学习

- 在建立模型时，监督式学习建立一个学习过程，将模型的预测结果与“训练集”的真实结果进行比较，计算误差，不断调整优化模型，直到模型的预测结果达到一个预期的准确率。

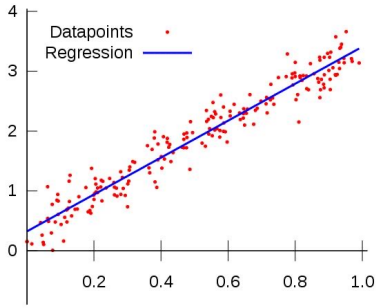
- 常见的监督学习任务包括回归分析和分类。

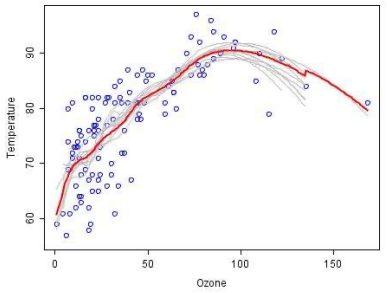
ML的分类

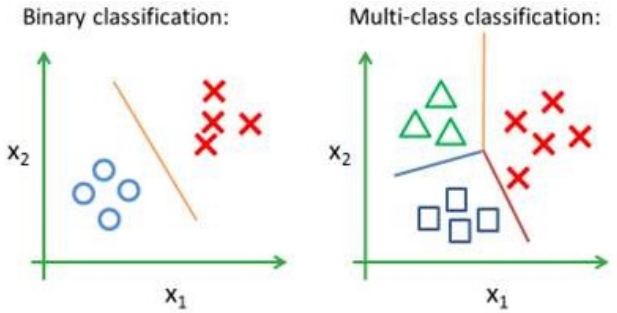
1、监督学习

常见的监督学习任务包括回归分析和分类。

- 回归（regression）： 特征输入 ⟹ 输出是连续值
- 分类（classification）：特征输入⟹输出是离散值







ML的分类

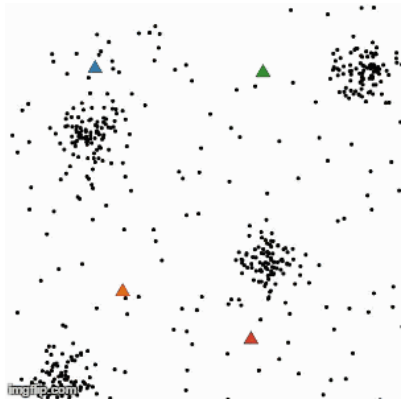
2、无监督学习

- 在无监督式学习中，数据并不被特别标识，学习模型是为了推断出数据的一些内在结构；
- 常见的应用场景包括关联规则的学习以及聚类等。常见算法包括 Apriori 算法和 K-Means 算法。
- 监督学习和无监督学习的区别：训练集中的数据是否已被标注。

ML的分类

2、无监督学习

- 聚类（clustering）
  输入一批样本数据 ⟶ 划分为若干簇
- 关联规则分析
  给定一批记录 ⟶ 记录中各项的关联关系



交易号	产品
T01	啤酒
T01	尿布
T02	啤酒
T02	尿布
T03	尿布

ML的分类

3、强化学习

- 强化学习通过观察来学习动作的完成，每个动作都会对环境有所影响，学习对象根据观察到的周围环境的反馈来做出判断；
- 在强化学习下，输入数据直接反馈到模型，模型必须对此立刻做出调整；
- 常见的应用场景包括动态系统以及机器人控制等。常见算法包括 Q-Learning 以及时间差学习。

机器学习方法

- 例： Open AI公司让AI玩躲猫猫
  - 研究人员设计的虚拟环境包括一个封闭的空间，里面有各种各样的物体，比如积木、坡道、移动障碍物和固定障碍物。这些智能体本身由强化学习算法控制。在每一场比赛中，这些智能体被分成两组：隐藏者（蓝色）和搜寻者（红色）。隐藏者成功躲避搜寻者即接受奖励，反之则进行惩罚；搜寻者找到隐藏者即为奖励，反之惩罚。和人类的捉迷藏游戏一样，隐藏者有几秒钟的时间藏起来。除此以外，研究人员没有给这些智能体任何其他指示。
  - 参考视频： AI人工智能玩捉迷藏.mp4