

AI解密——体育中的数据科学与人工智能

Lecture 06

机器学习简介

虞思逸

机器学习概览

- **机器学习是什么？为什么使用机器学习？**
- 机器学习的应用示例
- 机器学习系统的类型
- 机器学习的主要挑战
- 从事机器学习工作的准备

什么是机器学习?

- 关于机器学习你想到的关键词有哪一些?
- 下课，抬头看看天边的晚霞
 - >>>>>>>>嗯，今天又是一个好天气
- 走出上体走到水果摊旁，挑了个根底蜷缩、敲起来声音浑浊的青绿西瓜
 - >>>>>>>>嗯，一定是个皮薄肉厚的甜瓜



什么是机器学习?

- 人工智能 (Artificial Intelligence, 简称AI)、机器学习 (Machine Learning, 简称ML) 以及深度学习 (Deep Learning, 简称DL) 是当前热门的三个名词, 这三者之间既有一定联系, 也有明显的区别。



机器学习定义

- 机器学习是一门能够让编程计算机从数据中学习的计算机科学和艺术
- 机器学习(Machine Learning)是计算机科学的子领域，也是人工智能的一个分支和实现方式。Tom Mitchell在他1997年出版的《Machine Learning》一书中指出机器学习这门学科所关注的是计算机程序如何随着经验积累自动提高性能。同时给出了形式化的描述：
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，那么就称这个计算机程序在从经验 E 学习。
- 机器学习主要的理论基础涉及到概率论、数理统计、数值逼近、最优化理论、计算复杂理论等，核心要素是数据、算法和模型。

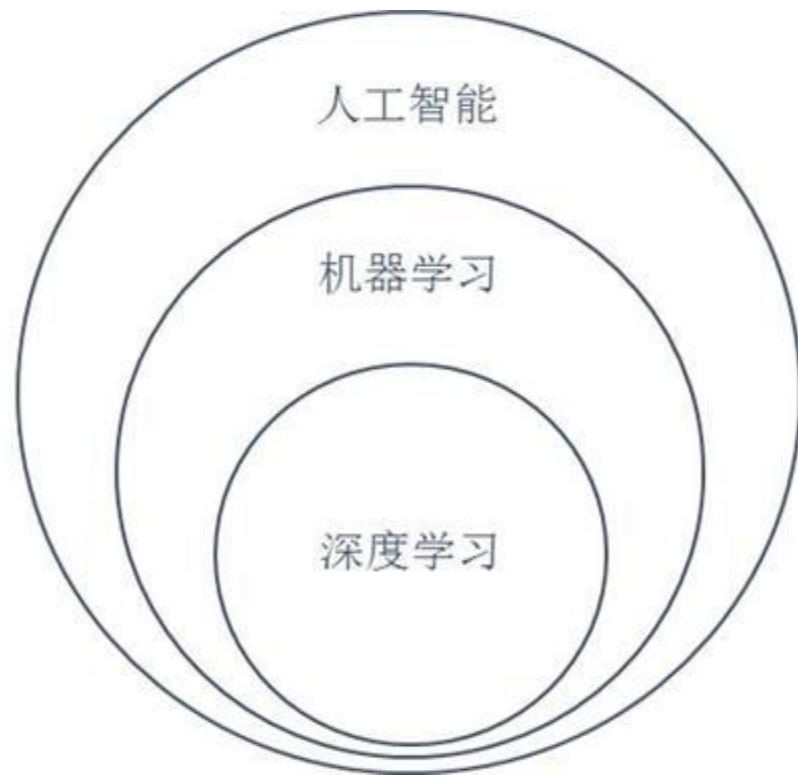
人工智能>>机器学习

- 机器学习是人工智能的一个分支
- 基本思想：基于数据构建统计模型，利用模型对数据进行分析 and 预测



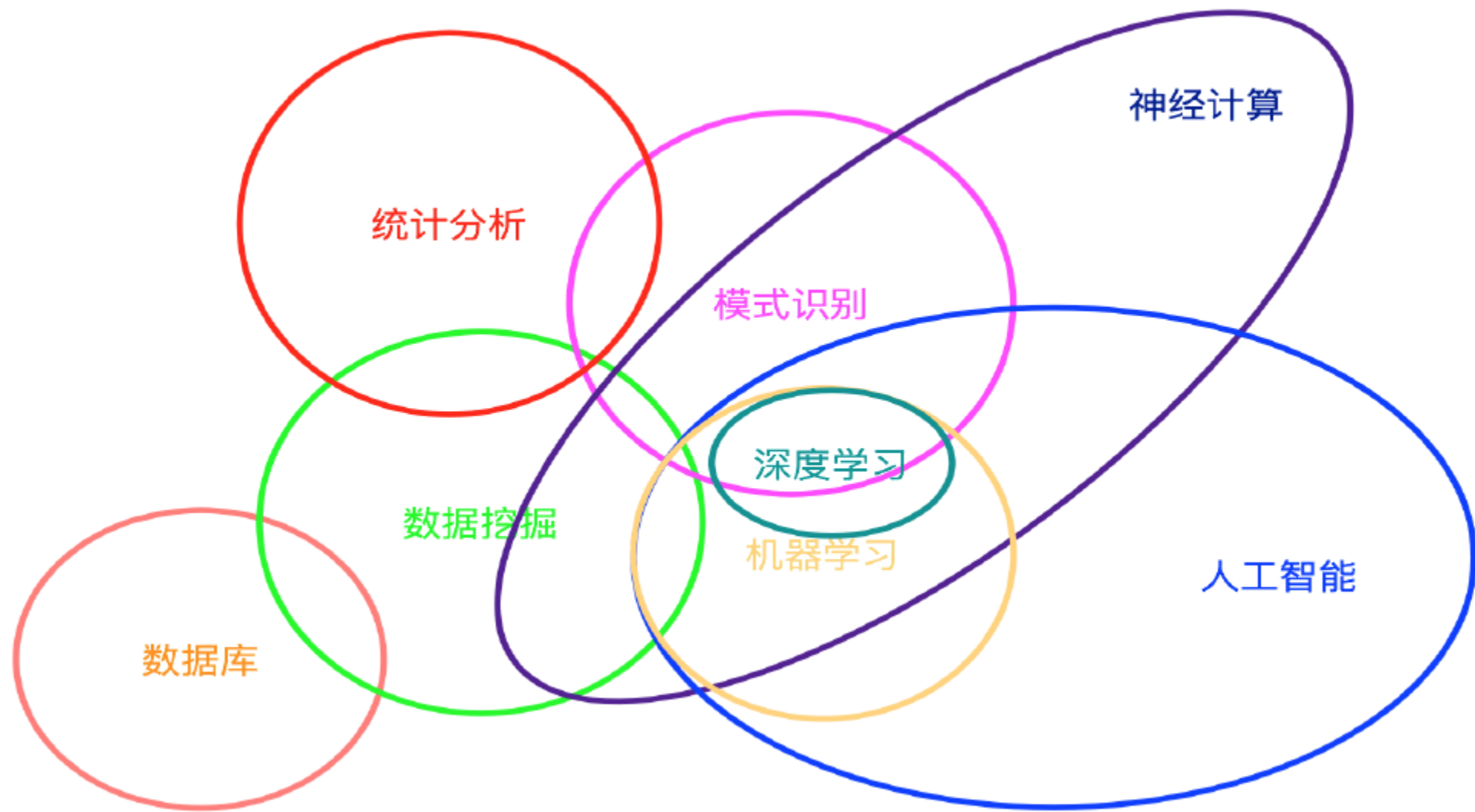
深度学习

- 是基于多层神经网络的机器学习方法
- 是特殊的机器学习实现方法。



人工智能、机器学习和深度学习关系图





机器学习的任务类型

- **回归模型和分类模型。**
 - 回归模型得到的结果是连续值，如预测明天的温度。
 - 分类模型得到的结果是离散值，如预测明天天气（阴、晴、雨）
- 回归模型可以**转换**为分类模型
 - 如根据预测的温度确定是否是高温天气
- 分类问题也可以通过回归模型来**预测**
 - 如某个事件出现的概率



机器学习的过程

- **组织数据**
 - 按要求对数据进行处理，转换成特定的格式
- **建立模型**
 - 利用样本数据进行算法学习，得到预测模型
- **预测结果**
 - 根据模型对输入数据的结果进行预测
- 根据预测结果，还可以**进一步评价和调整模型**



为什么要使用机器学习？

- 通过机器学习算法可以**简化**那些现有解决方案需要大量手动调整或者规则列表超长的问题的代码，并且**提升执行表现**
 - 例如：用传统编程技术编写垃圾邮件过滤器
- 解决传统技术手段根本无法解决的复杂问题
 - 例如：语音识别问题
- 在环境波动中适应新的数据
 - 例如：新写法的垃圾邮件
- 从复杂问题和海量数据中获得洞见（insight）
 - 例如：通过研究训练效果优秀的学习算法的数据，了解人类没有发现的数据中的关联性 or 新趋势

Q:贵专业中机器学习的应用场景？

Syllabus

- 机器学习是什么？为什么使用机器学习？
- **机器学习的应用示例**
- 机器学习系统的类型
- 机器学习的主要挑战
- 从事机器学习工作的准备

机器学习典型应用领域

- 艺术创作
- 金融领域
- 医疗领域
- 自然语言处理
- 网络安全
- 工业领域
- 娱乐行业



机器学习应用

人机大战



- AlphaGo本质上是
 - 深度卷积神经网络CNN、
 - 加强学习RL、
 - 蒙特卡洛树搜索MCTS三者相结合的产物

➤ 多分类问题

- 输入：棋局盘面
- 输出：各个位置的落子概率
- 模型：卷积神经网络CNN
- 数据来源：KGS Go Server上的16万盘6-9段的棋谱，共近3000万步



机器学习应用 趋势预测

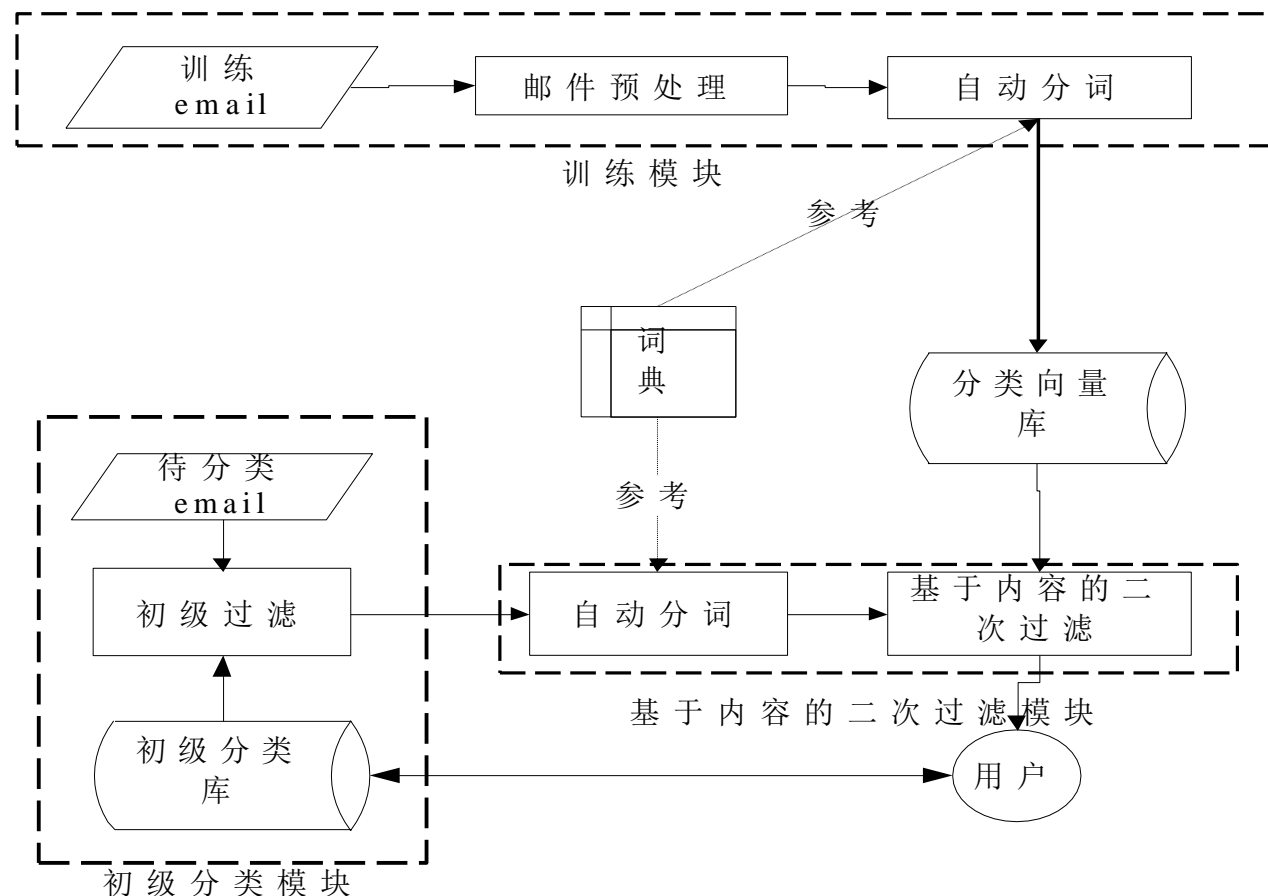


- Google流感趋势预测
- 社保欺诈


```

@ mail.turbomail.org:8080/tmw/9/mailmain?type=vieworg&mbid=0&mbtype=ne...
Return-Path: <runsly@qq.com>
Delivered-To: elaine.li@turbomail.org
X-TM-TID: turbomail.org_0_150599E6FD9
Received: from qq.com ([14.17.43.223])
    by turbomail.org WITH SMTP
    ID S150599E6FD9; Mon, 12 Oct 2015 09:16:07 +0800 (CST)
X-TM-DID: 5ad526c50fcab8f6abb2203944949059
X-TM-CONTENT-DID: 3bf1b081bdfb6c0dc077ee11c4dd85c
X-TM:
14.17.43.223;smtbg331.qq.com 183.60.177.41;turbomail.org;runsly@qq.com;elaine.
li@turbomail.org;remote
DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed; d=qq.com; s=s201307;
    t=1444612872; bh=DYUNGCVvk9kId0Jbdq88xh2DdhnHuJHR1zfkTRFJ6kA=;
    h=In-Reply-To:References:From:To:Subject:Mime-Version:Content-
Type:Content-Transfer-Encoding:Date:Message-ID;
    b=EgyJY1VyjwvT/bvMdEz/pr2VWu07jf6JjS9xDhaH2BeDsLKaUxB/i7ZX3tiZtiH8V
    qFTT/FBpwdZ3cb1NWIQZgk/Rj17ypDr6LUIInwMVLCOIT3G3dnYpPRBCM5jAX5eHYh/
    SZPN5hHlyUYcL31MkXpIg+DRAiUWW48khTP9qA9o=
X-QQ-FEAT: zaIfgOhwV2of/7fEoJtHZrRlkZ47IhHsKakdQHvFKDc=
X-QQ-SSF: 00010000000000F000000000000000Z
X-HAS-ATTACH: no
X-QQ-BUSINESS-ORIGIN: 2
X-Originating-IP: 119.180.199.91
In-Reply-To: <1440468007144@turbomail.org>
References: <1440468007144@turbomail.org>
X-QQ-mid: webmail735t1444612872t2518578
    
```

这封邮件，来源是真实的



机器学习应用

个性化推荐

☆ 与您浏览过的商品相关的推荐

时时Z秒杀

正在抢购 即将开始 已经结束



镇店之宝

¥23.90 - ¥4,888.00

剩余时间 41:55

精品美食美酒超级镇店之宝专场 低至39元



镇店之宝

¥199.00 - ¥899.00

剩余时间 41:55

【闪购全球】大牌男女鞋

★★★★☆ 257

经常一起购买的商品



总价: ¥105.10

立即购买组合



- ☑ 本商品: 清华科技大讲堂: 商务智能(第四版)
- ☑ 商务智能: 数据分析的管理视角(原书第3版)

浏览此商品的顾客也同时浏览



商务智能: 数据分析的管理视角(原书第3版)
拉姆什·沙尔达 (...)



基于
伯特·
平装

排行榜 专属你的购物指南

花卉 计算机与互... 大学教材 文具/耗材 电风扇



1 玫瑰花苗【19.8元任选5件】四季开花大花绿植物盆栽室内外花卉庭院阳台蔷薇



2 7天发芽【三免一】！碗莲种子【20粒已开口碗莲套装】荷花睡莲花种盆栽水培绿



3 买2株送2包肥料 买3株送一株 当年开花 蔷薇花苗爬藤苗攀援花卉观花植物绿植室内

觅me 探索生活

北欧实木床 3免1



会买专辑 甄选优质好物



双手专心拍照，双肩包轻松收纳

双肩包是户外元素与都市元素的结合，无论我们长途跋涉还是短途旅行，双肩包无疑是一款省力便捷的装备，它能

领券中心 前往领券中心



¥300

满2999元可用

【家装节】家具品类券 (...)



¥9.8折

满满10可用元可用

仅可购买智能硬件自营部...



¥100

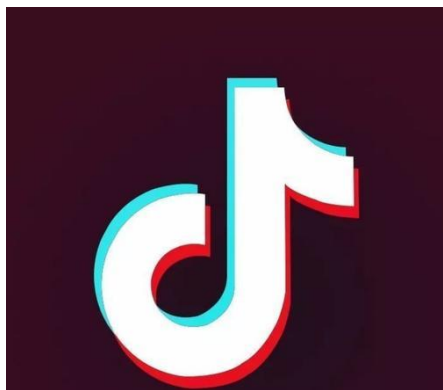
满199元可用

净胤茶叶官方旗舰店

发现好货 发现品质生活

会逛 你想逛的都在这里

信息茧房？





- 新时代的网络营销
- 为什么赢得政治选举与大数据分析联系在一起？

“我们的数据将会指示我们客户该将他们的竞选广告放到哪才能让他们的目标人群最有可能看到。”

机器学习概览

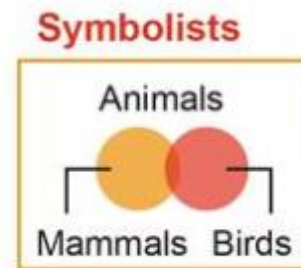
- 机器学习是什么？为什么使用机器学习？
- 机器学习的应用示例
- **机器学习系统的类型**
- 机器学习的主要挑战
- 从事机器学习工作的准备

机器学习的发展

- 机器学习的发展分为知识推理期、知识工程期、浅层学习（Shallow Learning）和深度学习（Deep Learning）几个阶段。
- 在机器学习的发展过程中，随着人们对智能的理解和现实问题的解决方法演变，大致出现了符号主义、贝叶斯、联结主义、进化主义、行为类推主义五大流派。

先熟悉一下各种名词

1、符号主义 (Symbolists)



- 名称：符号主义 (Symbolists)
- 起源：逻辑学、哲学核心思想：认知即计算，通过对符号的演绎和逆演绎进行结果预测
- 问题：知识结构
- 代表算法：逆演绎算法 (Inverse deduction)
- 代表应用：知识图谱 (智能搜索、深度问答、社交网络)
- 代表人物：Tom Mitchell、Steve Muggleton、Ross Quinlan



Tom Mitchell



Steve Muggleton



Ross Quinlan

Likelihood	Prior
Posterior	Margin

2、贝叶斯派 (Bayesians)

- 名称：贝叶斯派 (Bayesians)
- 起源：统计学核心思想：主观概率估计，发生概率修正，最优决策
- 问题：不确定性
- 代表算法：概率推理 (Probabilistic inference)
- 代表应用：反垃圾邮件、概率预测
- 代表人物：David Heckerman、Judea Pearl、Michael Jordan



David Heckerman



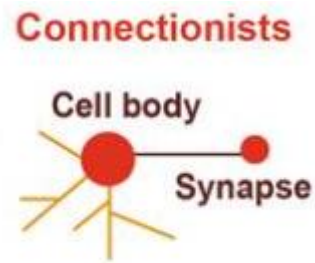
Judea Pearl



Michael Jordan

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)}$$

3、联结主义 (Connectic



- 名称：联结主义 (Connectionist)
- 起源：神经科学核心思想：对大脑进行仿真
- 问题：信度分配
- 代表算法：反向传播算法 (Backpropagation)、深度学习 (Deep learning)
- 代表应用：机器视觉、语音识别
- 代表人物：Yann LeCun、Geoff Hinton、Yoshua Bengio



Yann LeCun



Geoff Hinton



Yoshua Bengio

4、进化主义 (Evolutionaries)

Evolutionaries



- 名称：进化主义 (Evolutionaries)
- 起源：进化生物学
- 核心思想：对进化进行模拟，使用遗传算法和遗传编程
- 问题：结构发现
- 代表算法：基因编程 (Genetic programming) 变异，基因复制，基因删除
- 代表人物：John Koda、John Holland、Hod Lipson



John Koda



John Holland



Hod Lipson

5、行为类比主义 (Analogizer)

Evolutionaries



- 名称：行为类比主义 (Analogizer)
- 起源：心理学
- 核心思想：新旧知识间的相似性
- 问题：相似性
- 代表算法：核机器 (Kernel machines)、近邻算法 (Nearest Neighbor)
- 代表应用：Netflix推荐系统
- 代表人物：Peter Hart、Vladimir Vapnik、Douglas Hofstadter



Peter Hart



Vladimir Vapnik



Douglas Hofstadter


XGBoost —— Kaggle 竞赛王者



陈天奇 CS. PhD 机器学习系统

所在行业 计算机软件

教育经历  华盛顿大学 (University of Washington) · 机器学习

 上海交通大学 · 计算机

个人简介 tqchen.com

机器学习五大流派的演化阶段

1990 年代到 2000 年

- 主导流派：贝叶斯
- 架构：小型服务器集群
- 主导理论：概率论
- 分类：可扩展的比较或对比，对许多任务都足够好了

2010 年代末期

- 主导流派：联结主义+符号主义
- 架构：许多云
- 主导理论：记忆神经网络、大规模集成、基于知识的推理
- 简单的问答：范围狭窄的、领域特定的知识共享

2040 年代+

- 主导流派：算法融合
- 架构：无处不在的服务器
- 主导理论：最佳组合的元学习
- 感知和响应：基于通过多种学习方式获得的知识或经验采取行动或做出回答

1980 年代

- 主导流派：符号主义
- 架构：服务器或大型机
- 主导理论：知识工程
- 基本决策逻辑：决策支持系统，实用性有限

2010 年代早期到中期

- 主导流派：联结主义
- 架构：大型服务器农场
- 主导理论：神经科学和概率
- 识别：更加精准的图像和声音识别、翻译、情绪分析等

2020 年代+

- 主导流派：联结主义+符号主义+贝叶斯+.....
- 架构：云计算和雾计算
- 主导理论：感知的时候有网络，推理和工作的时候有规则
- 简单感知、推理和行动：有限的自动化或人机交互

Ups and downs of Deep Learning

- 1958: Perceptron (linear model)
- 1969: Perceptron has limitation
- 1980s: Multi-layer perceptron
 - Do not have significant difference from DNN today
- 1986: Backpropagation
 - Usually more than 3 hidden layers is not helpful
- 1989: 1 hidden layer is “good enough”, why deep?
- 2006: RBM initialization (breakthrough)
- 2009: GPU
- 2011: Start to be popular in speech recognition
- 2012: win ILSVRC image competition

机器学习的演化

时间	主要成果	代表人物
1943~1956	MP模型、自动机模型、符号演算、逻辑主义	Warren McCulloch / Walter Pitts / Alan Turing / John von Neumann / Shannon
1956~1960s	LISP、框架知识表示	McCarthy/Minsky / Newell & Simon
1960s~1970s	遗传算法、进化策略、模糊集	Rechenberg / Holland / Zadeh
1970s~1980s	专家系统、DENDRAL、MYCIN、PROSPECTOR、PROLOG、EMCIN等	Feigenbaum / Buchanan / Lederberg / Shortliffe
1980s~1990s	Hopfield网络、自组织网络、多层神经网络、知识工程、模糊逻辑、决策树算法等	Hopfield / Kohonen / Feigenbaum / Zadeh / Quinlan
1990s~2000s	Boosting算法、AdaBoost、SVM、随机森林	Schapire / Freund / Vapnik
2000s~至今	深度学习、自我特征学习、无导式学习、增强学习、分布式机器学习	Hinton / LeCun / Bengio / Andrew Ng / Mitchell

• 机器学习的类别

监督式/无监督式学习 标准： 是否在人的监督下训练

- 监督学习

- 比如：

- 分类任务：
垃圾邮件过滤
- 回归任务：
汽车价格预测
房价预测

- 重要的监督学习的算法：

- K-近邻算法(KNN)
- 线性回归(LR)
- 逻辑回归(LR)
- 支持向量机(SVM)
- 决策树和随机森林(DT RF)
- 一部分神经网络(NN)

Supervised learning

In *supervised learning*, the training set you feed to the algorithm includes the desired solutions, called *labels* (Figure 1-5).

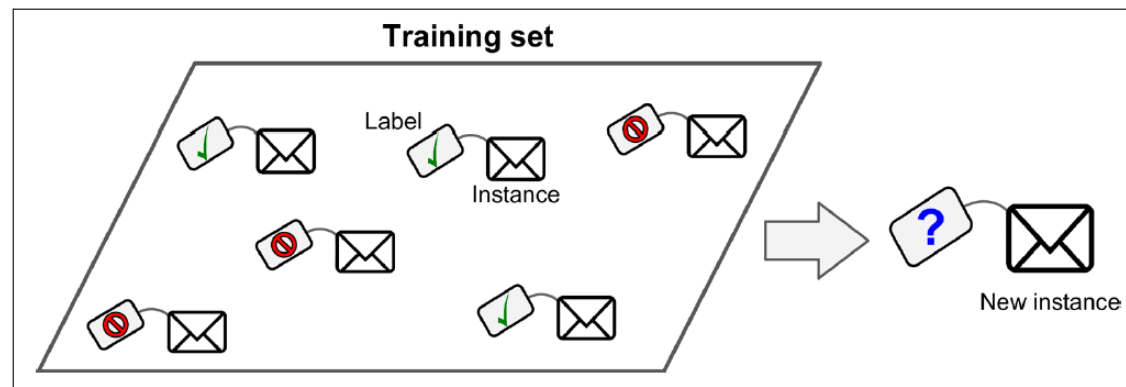


Figure 1-5. A labeled training set for spam classification (an example of supervised learning)

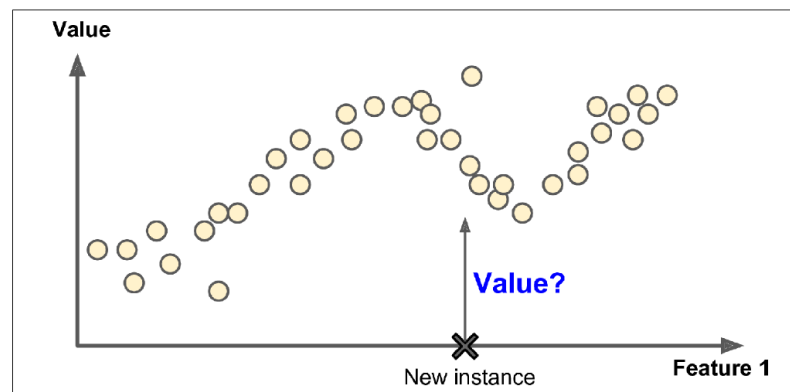
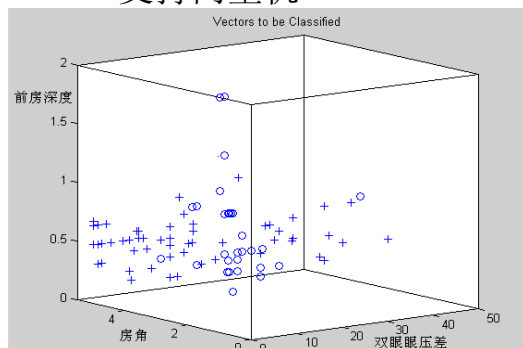


Figure 1-6. A regression problem: predict a value, given an input feature (there are usually multiple input features, and sometimes multiple output values)

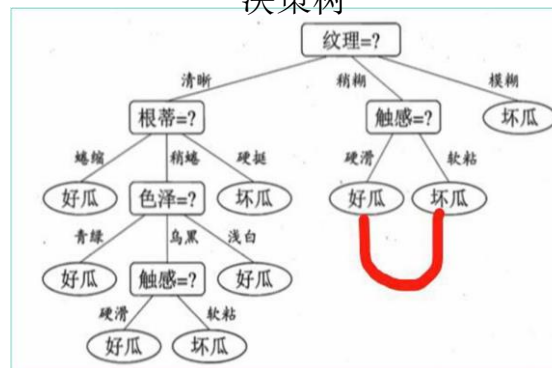
机器学习常用算法

常用分类算法典型应用

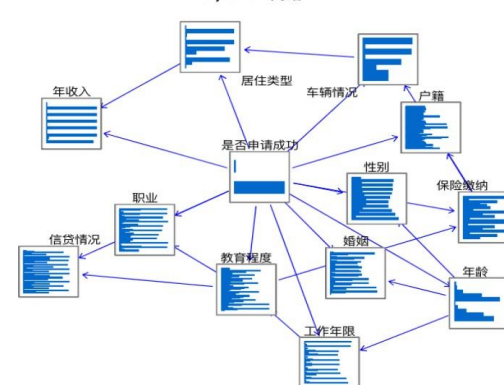
支持向量机



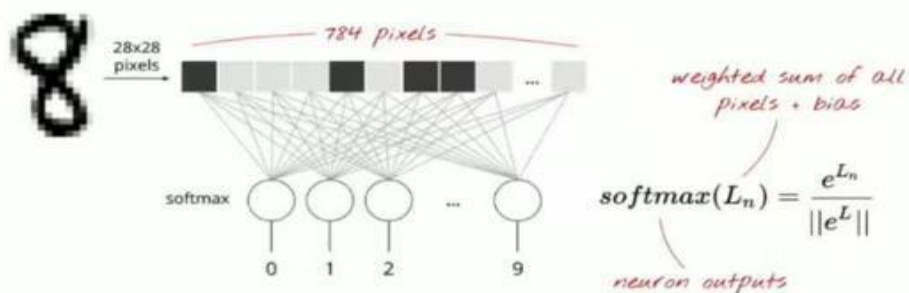
决策树



Bayesian 网络



Very simple model: softmax classification

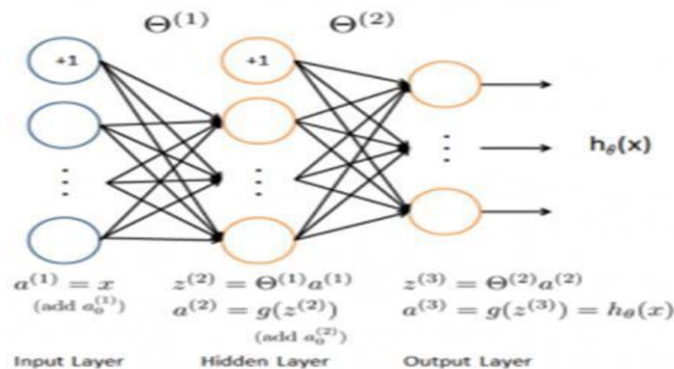
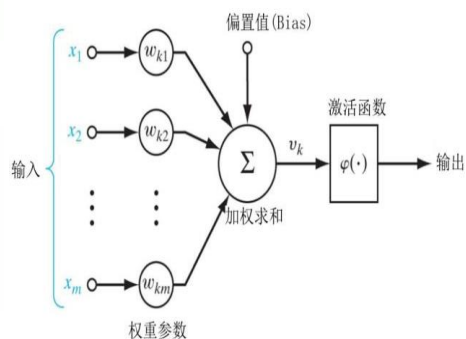
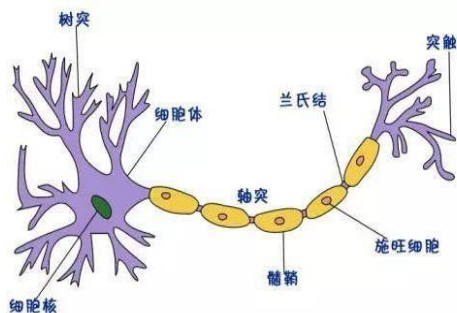


Hello World: handwritten digits classification - MNIST



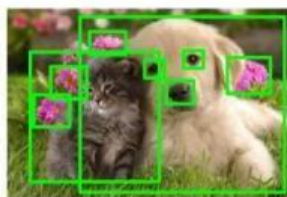
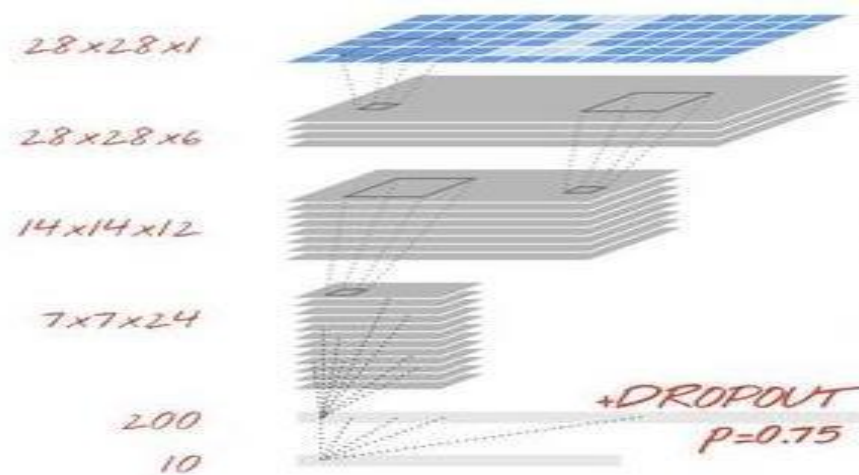
神经网络

- 传统的神经网络为BP神经网络，基本网络结构为输入层、隐藏层和输出层，节点代表神经元，边代表权重值，对输入值按照权重和偏置计算后将结果传给下一层，通过不断的训练修正权重和偏置。递归神经网络（RNN）、卷积神经网络（CNN）都在神经网络在深度学习上的变种。
- 神经网络的训练主要包括前向传输和反向传播。
- 神经网络的结果准确性与训练集的样本数量和分类质量有关。
- 神经网络是基于历史数据构建的分析模型，新数据产生时需要动态优化网络的结构和参数。



数据挖掘常用算法

神经网络和深度学习



- 多层前馈神经网络
- 常见的深度学习神经网络
- 卷积神经网络
- 循环神经网络

airplane

automobile

bird

cat

deer

dog

frog

horse

ship

truck



深度学习

➤ 深度学习是通过构建多个隐藏层和大量数据来学习特征，从而提升分类或预测的准确性。

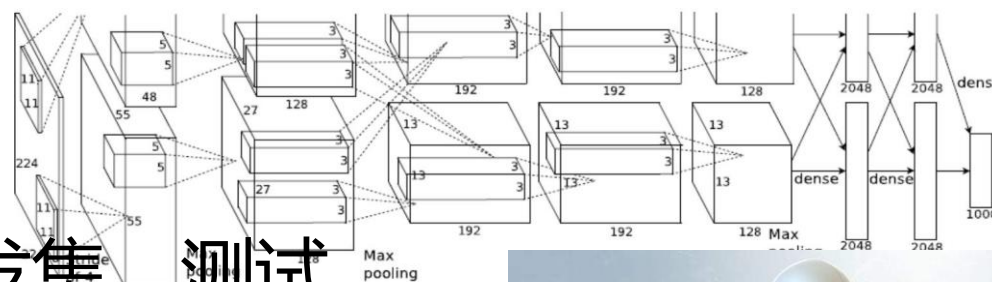
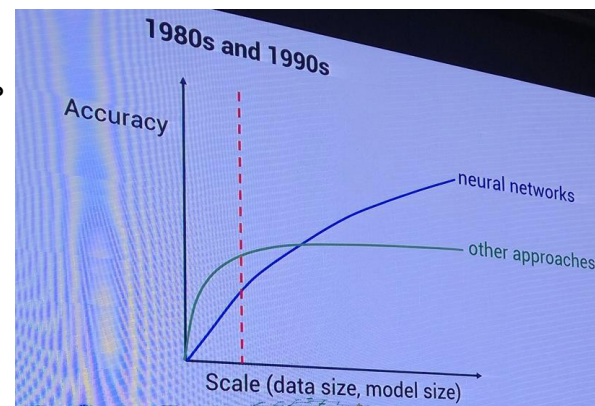
➤ 与神经网络相比，层数更多，而且有逐层训练机制避免梯度扩散

➤ 深度学习包括了

- 卷积神经网络 (CNN)
- 深度神经网络 (DNN)
- 循环神经网络 (RNN)
- 对抗神经网络 (GAN)

➤ 深度学习中训练集、开发集、测试集的样本比例一般为6:2:2。

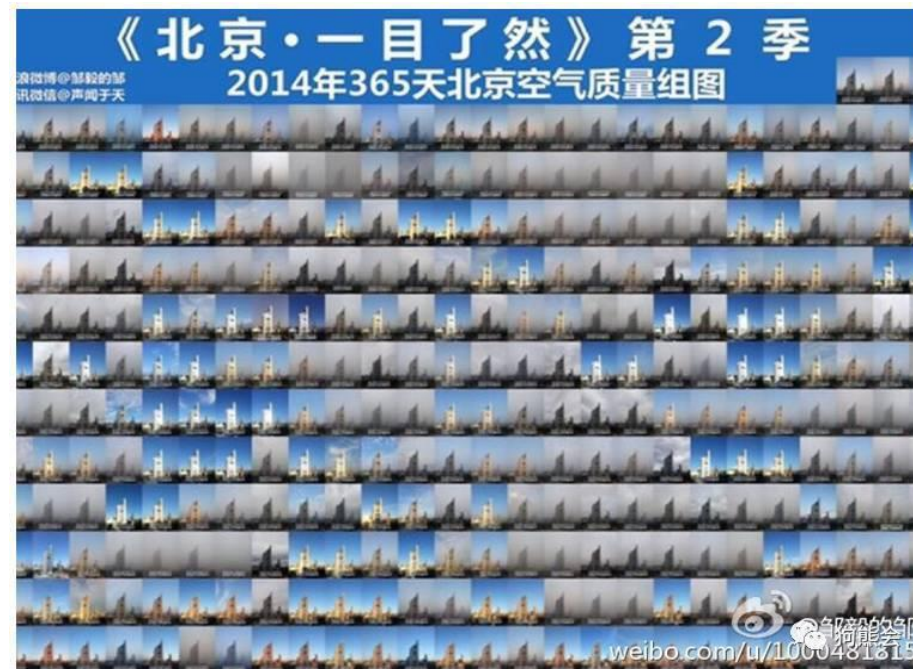
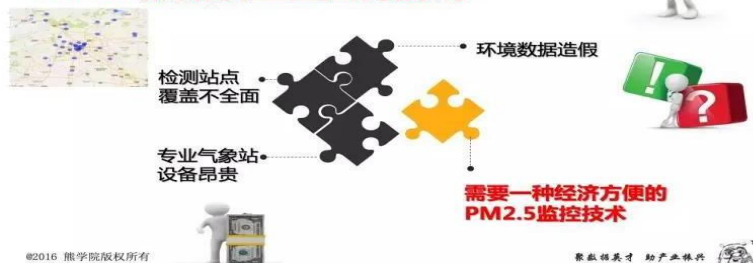
➤ 常见的权重更新方式包括SGD和Momentum。



通过图片识别PM2.5



PM2.5数据质量监控的挑战



通过图片识别PM2.5

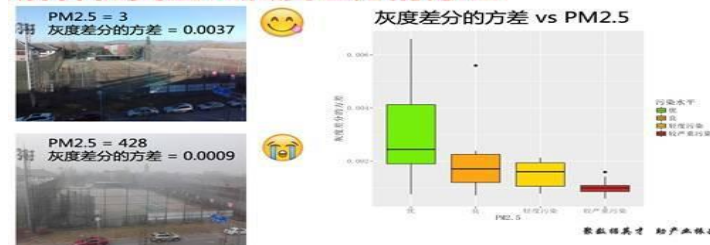
- 从衡量图像清晰程度的角度出发，对图像特征进行观察和分析，得到4个解释性变量：**灰度差分的方差、清晰度、饱和度、高频含量等**



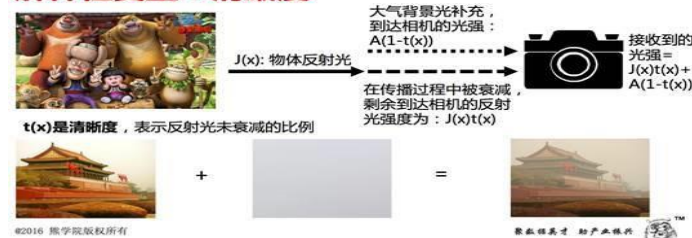
解释性变量：饱和度



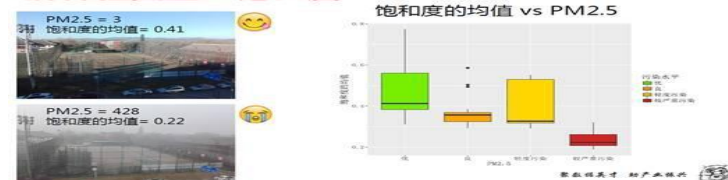
解释性变量：灰度差分的方差



解释性变量：清晰度



解释性变量：饱和度

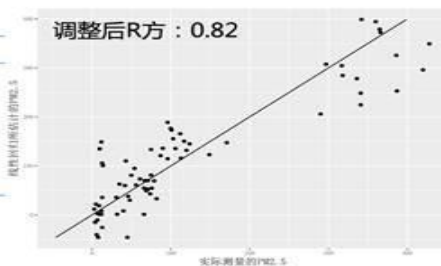


通过图片识别PM2.5

- 多元线性回归的拟合优度为0.82

线性回归：PM2.5

变量	系数估计值	P值
灰度差分的方差	18.273	0.144
清晰度	-70.447	<0.001 ***
饱和度的均值	-45.231	<0.001 ***
高频含量	-40.969	<0.001 ***



©2016 熊学院版权所有

定序回归：污染等级

PM2.5值对应的空气质量等级：中国			
PM2.5值	0-50	50-100	100-150
空气质量等级	一级（优）	二级（良）	三级（轻度污染）
PM2.5值	150-200	200-300	300-500
空气质量等级	四级（中度污染）	五级（重度污染）	六级（严重污染）

预测等级-实际等级	0(完全正确)	1	2	3
百分比 / %	68.1	30.1	0.0	1.4

©2016 熊学院版权所有

预测集与训练集的划分——
留一交叉验证法：
每次提取1个样本作为预测集，剩下的作为训练集
进行对此样本的预测

熊数据英才 助产业振兴

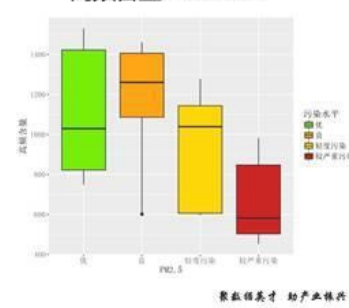
解释性变量：高频含量



解释性变量：高频含量



高频含量 vs PM2.5



- 无监督学习
例如：异常检测：比如信用卡欺诈，又或者从数据集中移除异常值

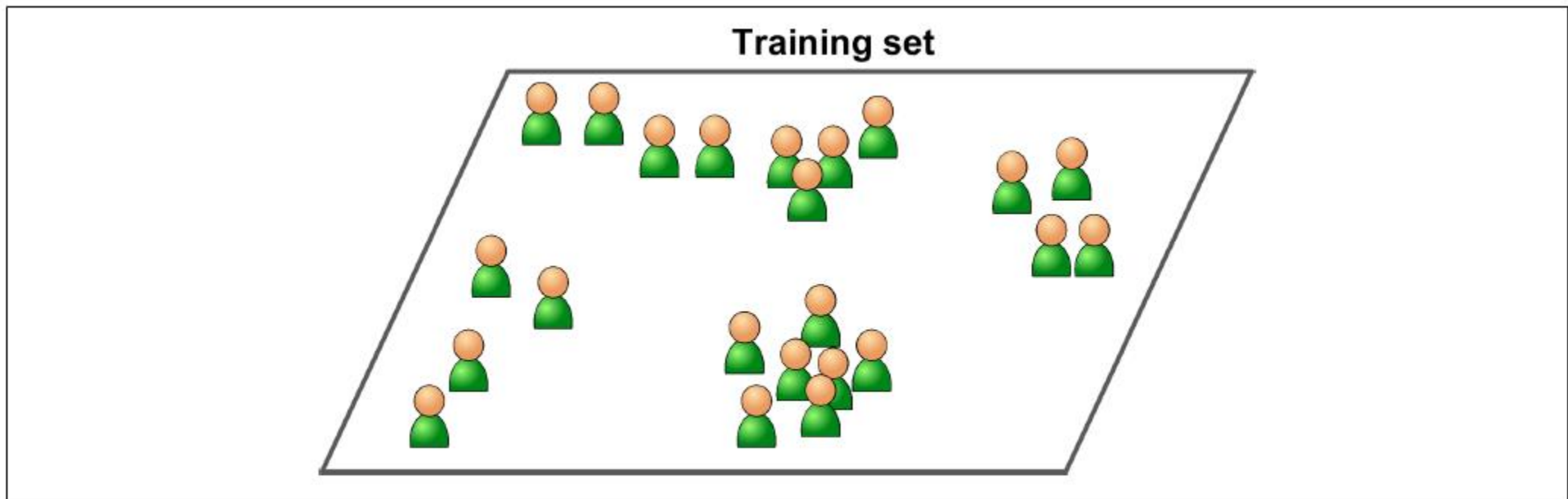


Figure 1-7. An unlabeled training set for unsupervised learning

- 无监督学习（异常检测：比如信用卡欺诈，又或者从数据集中移除异常值）
- 聚类算法（比如，检测相似访客的分组）
-

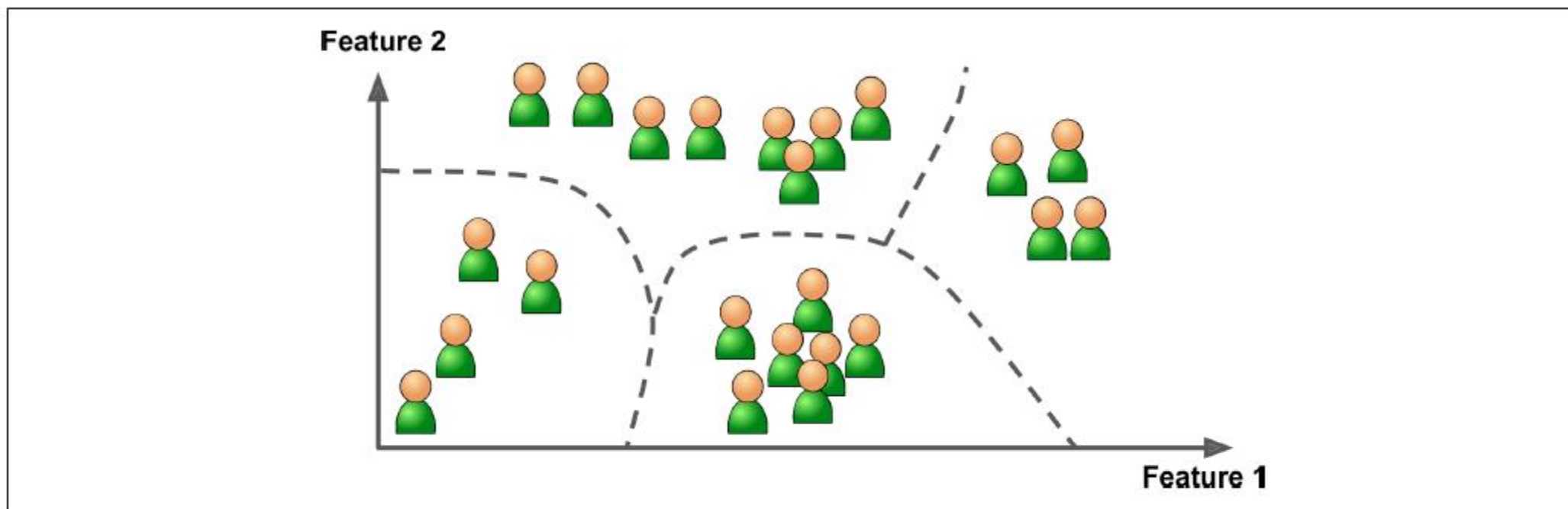
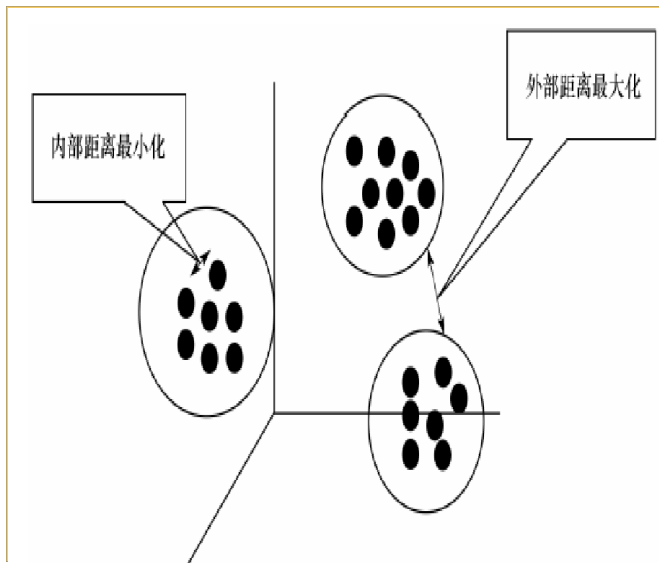


Figure 1-8. Clustering

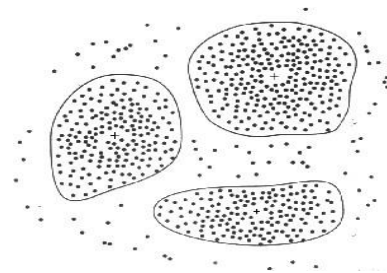
机器学习常用算法

聚类算法

- 淘宝潜在用户分析
- 社交网络用户分析

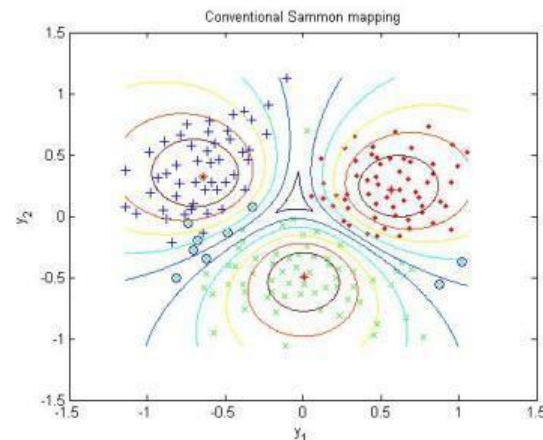
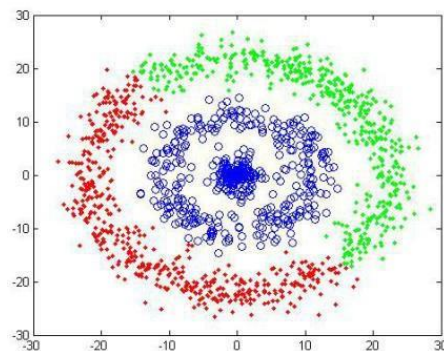


- 聚类分析是把一个给定的数据对象（样本）集合分成不同的簇（组）。
- 聚类就是把整个数据分成不同的组，并使组与组之间的差距尽可能大，组内数据的差异尽可能小。
- K-means是一种常用的聚类算法，用户指定聚类的类别数K，随机地选择K个对象作为K个初始聚类中心。对剩余的每个对象，分别计算与初始聚类中心的距离，根据距离划到不同的簇。然后重新计算每个簇的平均值，求出新的聚类中心，再重新聚类。这个过程不断重复，直到收敛（相邻两次计算的聚类中心相同）。



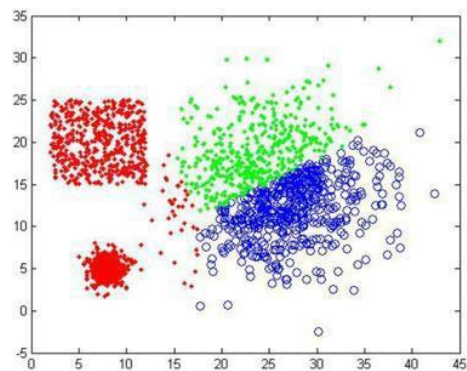
聚类算法

- 聚类是基于无监督学习的分类模型，按照数据内在结构特征进行聚集形成簇群。聚集方法即记录之间的区分规则。
- 聚类与分类的主要区别是其不关心数据的类别。
- 聚类首先选择有效特征向量，然后按照距离函数进行相似度计算。
- 聚类应用广泛
 - 客户群体特征、消费者行为分析、市场细分、交易数据分析
 - 动植物种群分类、医疗领域的疾病诊断、环境质量检测。



常见聚类算法

- 基于层次聚类 (Hierarchical Method)
 - BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)
 - CURE (Clustering Using Representatives)
- 基于划分的聚类
 - K均值 (K-Means)
- 基于密度的聚类
 - DBSCAN (Density-based spatial clustering of applications with noise)
 - OPTICS (Ordering Points To Identify the Clustering Structure)
- 基于机器学习的聚类
- 基于约束的聚类
- 基于网络的聚类

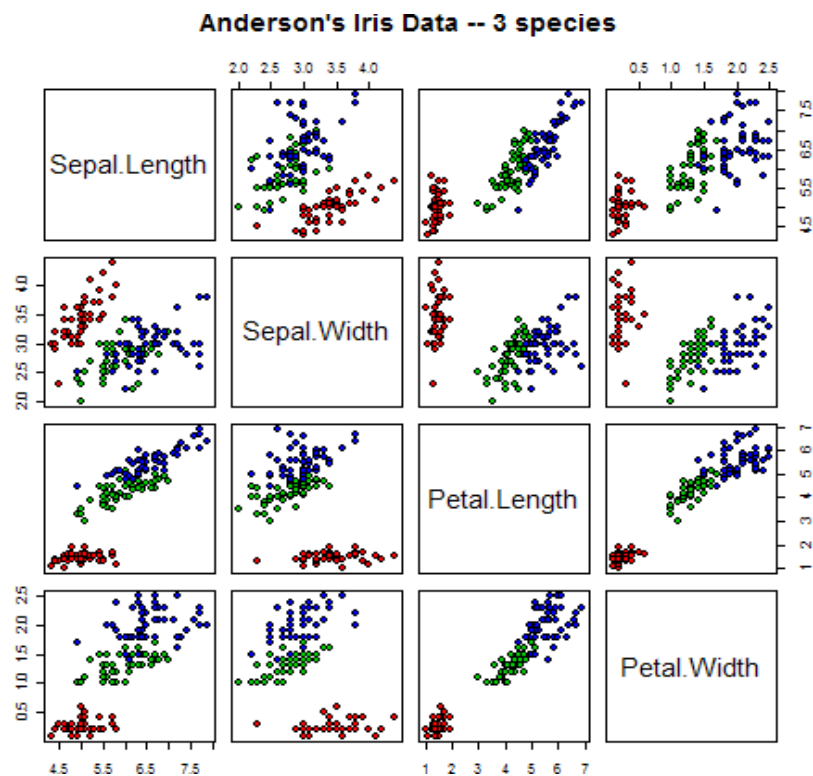


- 无监督学习（异常检测：比如信用卡欺诈，又或者从数据集中移除异常值）
- 可视化和降维
- (可视化，便于人理解数据怎么组织)

数据可视化

➤ 数据可视化在机器学习中的作用：

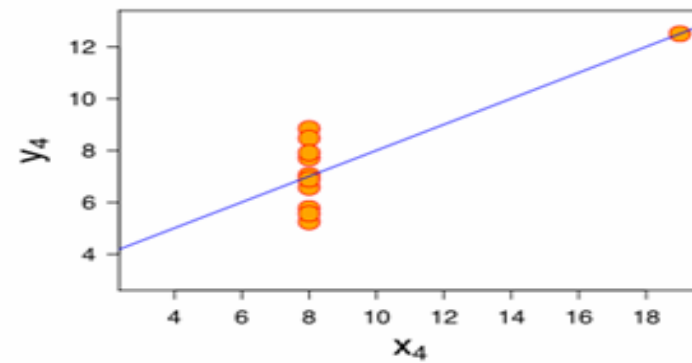
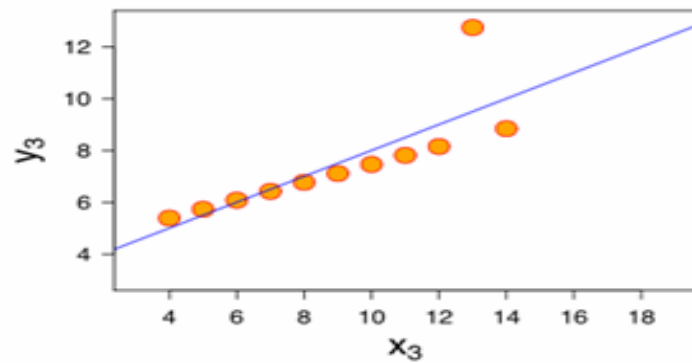
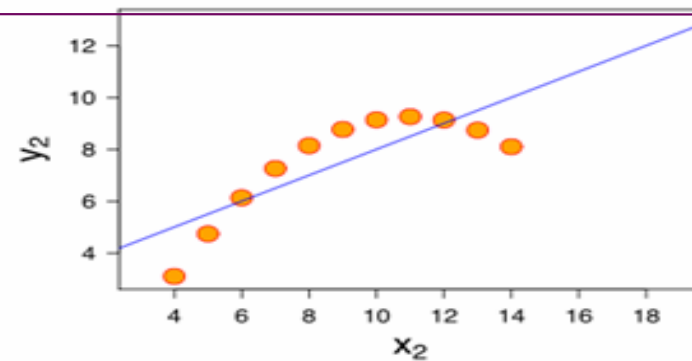
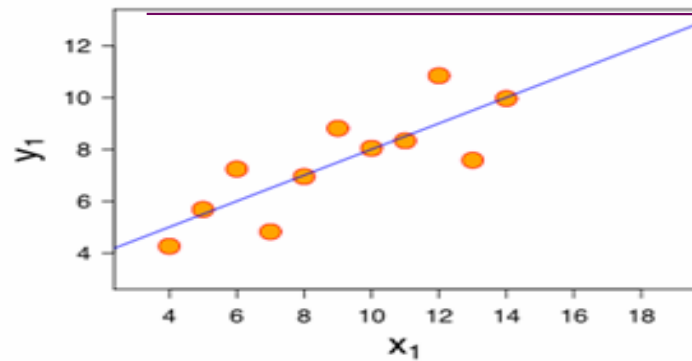
- (1) 视觉是人类获得信息的最主要途径；
- (2) 可视化本身就是一种机器学习方法；
- (3) 可视化可以作为数据预处理的方法或者是机器学习过程的表示方式。
- (4) 机器学习的结果也可以用可视化的形式表示。



可视化技术案例

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

可视化技术案例



- 无监督学习（异常检测：比如信用卡欺诈，又或者从数据集中移除异常值）
- 可视化和降维
- (可视化，便于人理解数据怎么组织)
- (降维，用来做特征提取，减小数据所占空间，提高速度，有可能提高性能)
- 主成分分析(PCA)
 - 核主成分分析(Kernel PCA)
 - 局部线性嵌入(LLE)
 - t-分布随机近嵌入(t-SNE)
- Eg. 里程数和使用年限
-

- 无监督学习（异常检测：比如信用卡欺诈，又或者从数据集中移除异常值）
- 关联规则学习（发现数据属性间的联系：比如超市的商品）
 - （频繁项集）
- Apriori
- Eclat

项集	支持度计数
{牛肉, 鸡肉}	3
{牛肉, 牛奶}	2
{牛肉, 奶酪}	3
{牛肉, 衣服}	1
{鸡肉, 牛奶}	4
{鸡肉, 奶酪}	2
{鸡肉, 衣服}	3
{牛奶, 奶酪}	1
{牛奶, 衣服}	3
{奶酪, 衣服}	1



频繁 2-项集	支持度计数
{牛肉, 鸡肉}	3
{牛肉, 奶酪}	3
{鸡肉, 牛奶}	4
{牛奶, 衣服}	3
{鸡肉, 衣服}	3

- 半监督学习
- 大量未标记数据和少量已标记数据
大多数半监督算法采用无监督和监督算法的结合：
深度信念网络(DBN):基于一种互相堆叠的无监督式组件
(受限玻尔兹曼机)，然后用监督式学习进行微调

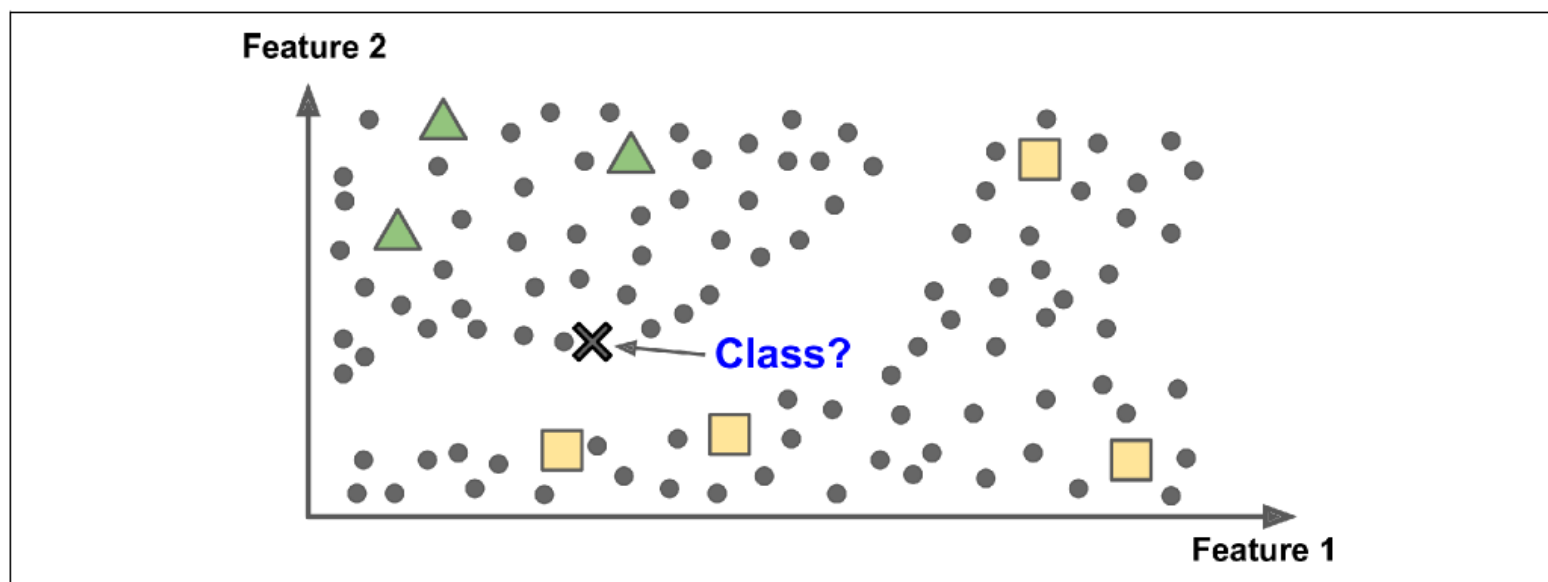


Figure 1-11. Semisupervised learning with two classes (triangles and squares): the unlabeled examples (circles) help classify a new instance (the cross) into the triangle class rather than the square class, even though it is closer to the labeled squares

强化学习

- Reinforcement learning

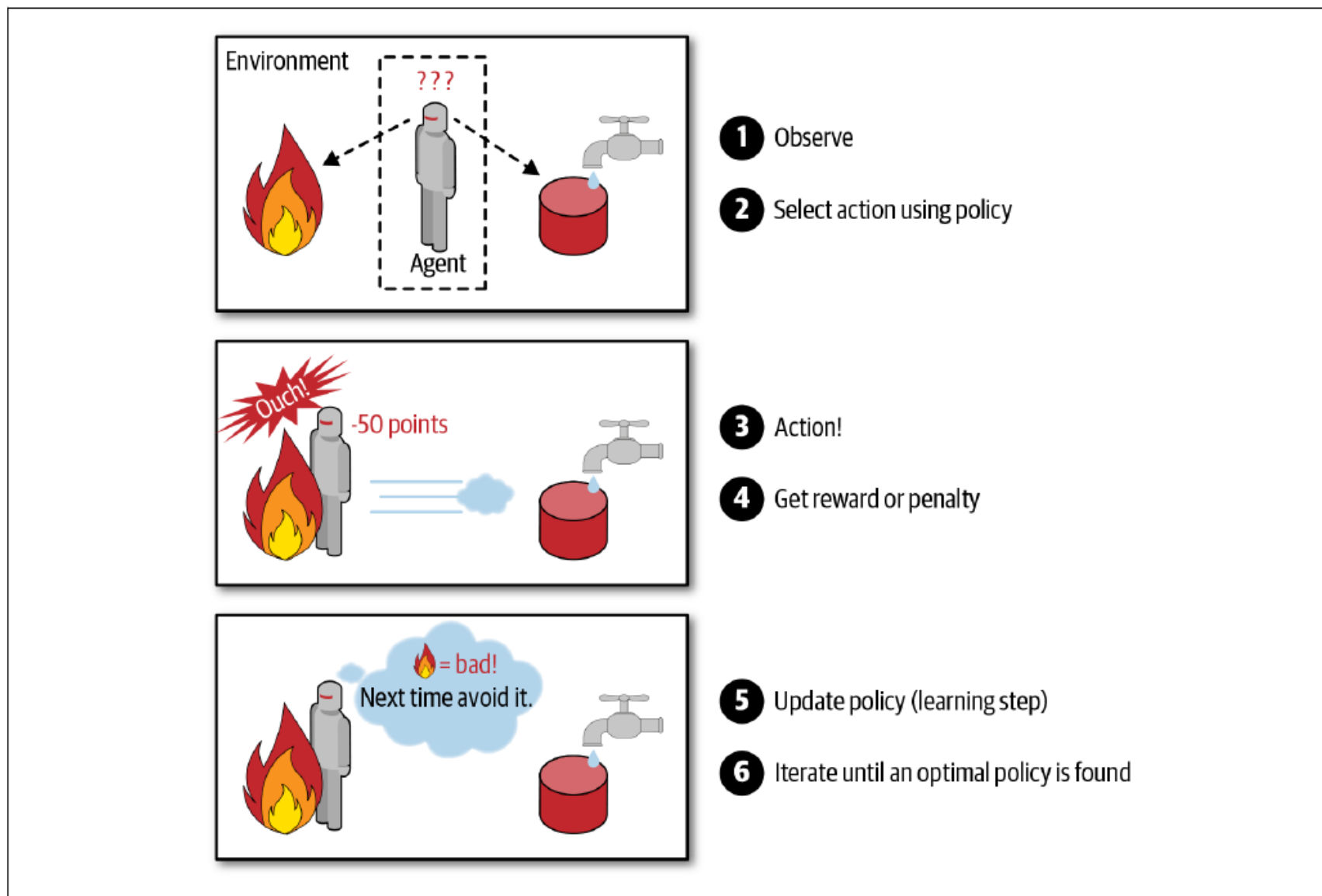
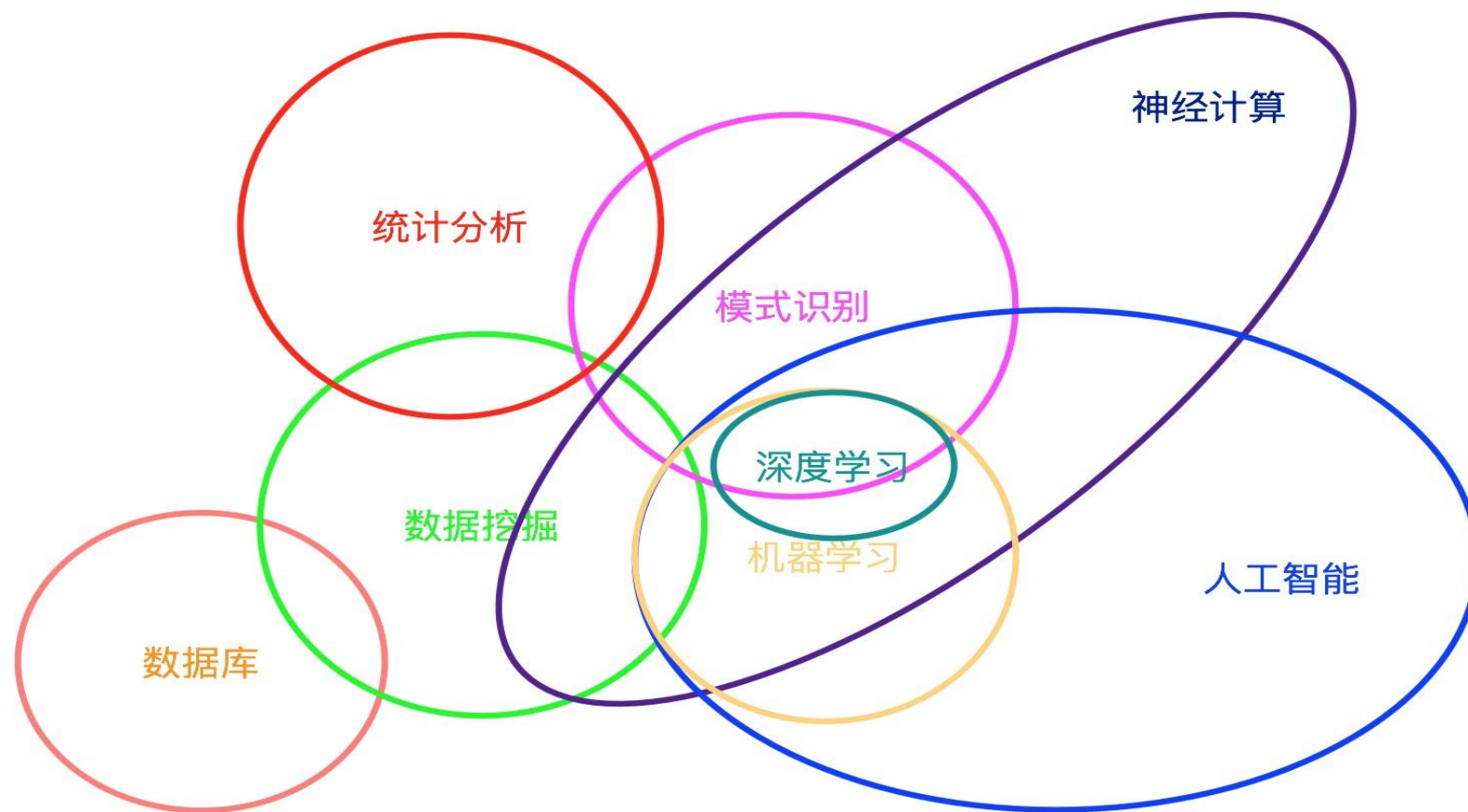


Figure 1-12. Reinforcement Learning

机器学习、人工智能和数据挖掘



机器学习、人工智能和数据挖掘

- **机器学习是人工智能的一个分支，它是实现人工智能的一个核心技术，即以机器学习为手段解决人工智能中的问题。机器学习是通过一些让计算机可以自动“学习”的算法并从数据中分析获得规律，然后利用规律对新样本进行预测。**
- **数据挖掘是从大量的业务数据中挖掘隐藏、有用的、正确的知识促进决策的执行。数据挖掘的很多算法都来自于机器学习，并在实际应用中进行优化。机器学习最近几年也逐渐跳出实验室，解决从实际的数据中学习模式，解决实际问题。数据挖掘和机器学习的交集越来越大。**

第一章 Introduction 机器学习概览

- 机器学习是什么？为什么使用机器学习？
- 机器学习的应用示例
- 机器学习系统的类型
- **机器学习的主要挑战**
- 从事机器学习工作的准备

机器学习的挑战

- 训练数据
 - 数量不足
 - 不具有代表性——采样偏差
 - 低质量数据
- 无关特征
- 过拟合/欠拟合

数据量较少

- 数据挖掘需要一定的数据量作支撑，尽量覆盖领域。
- 数据量增多，其中的规律会越来越明显，也更易发现与分析目标相关的因素
 - 神经网络
 - 深度学习
- 一般来说，数据量是自变量数量的10~20倍为佳。
- 数据样本需要有足够的覆盖范围，需要覆盖与分析目标相关的维



机器学习的挑战

- 训练数据
 - 数量不足
 - 不具有代表性——采样偏差
 - 低质量数据
- 无关特征
- 过拟合/欠拟合

1936 年
富兰克林·迪拉诺·罗斯福
(Franklin Delano Roosevelt)
VS
艾尔弗雷德·兰登
(Alfred Landon)

采样偏差

- 背景：美国失业人数高达九百万，在1929--1935这段期间实际收入下降了约1/3

兰登：“小政府”

口号：

“挥霍浪费的人必须离任”

“我们应该专心致力于自己的事务”

罗斯福：“扩大内需”

口号：

“在我们能够平衡联邦政府的预算之前，必须先平衡美国人民的预算”

《文学文摘》（literary Digest）认为兰登会以57%对43%的优势获胜

- 1、民意调查地址：电话簿、汽车车主登记资料、俱乐部登记 名单中选取。
- 2、仅有23%的人回复了民意调查邮件。

罗斯福以62%对38%的一边倒优势赢得了1936年的选举

机器学习的挑战

- 训练数据
 - 数量不足
 - 不具有代表性——采样偏差
 - 低质量数据
- 无关特征
- 过拟合/欠拟合

数据质量问题与预处理

- 数据质量要求数据是完整的和真实的，并且具有一致性和可靠性
- “垃圾进，垃圾出”
- 数据预处理占用整个机器学习项目60%的工作量
- 问题
 - 数据不完整
 - 异常数据
 - 重复数据
 - 数据不一致



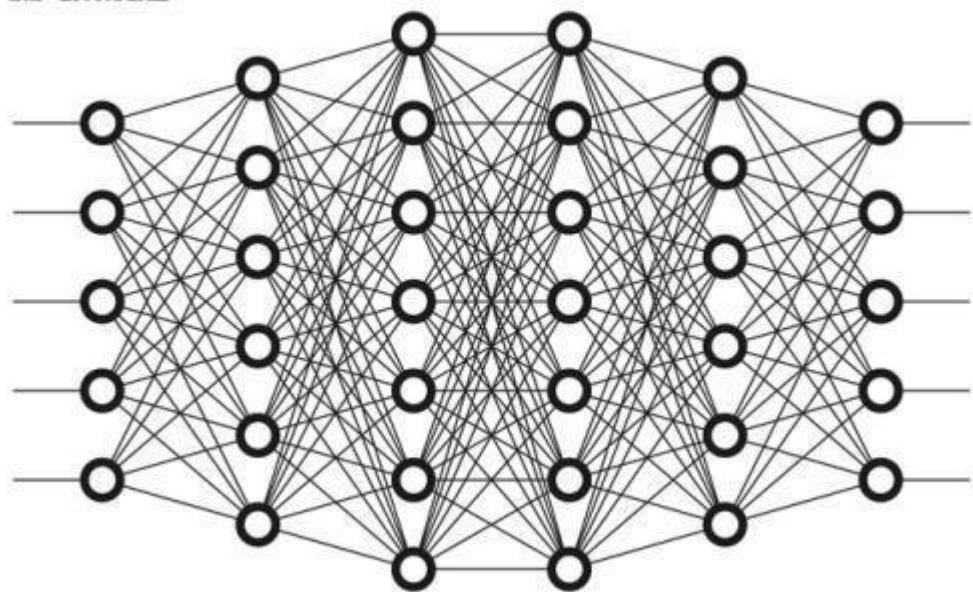
机器学习的挑战

- 训练数据
 - 数量不足
 - 不具有代表性——采样偏差
 - 低质量数据
- 无关特征
- 过拟合/欠拟合

维度灾难

- 当数据中的自变量较多时，会出现维度灾难问题。
- 特别是在矩阵数据中，其中冗余变量占比比较高时，可用数据变成稀疏矩阵，在分类算法处理时就没办法可靠地进行类别划分，在聚类算法中则容易使聚类质量下降。
- 可采用线性代数的相关方法将数据从高维空间影射到低维空间中
 - 主成分分析 (PCA)
 - 奇异值分解 (SVD)

来源：神经网络技术网 tech.16188.com



数据分析常见陷阱 (1)

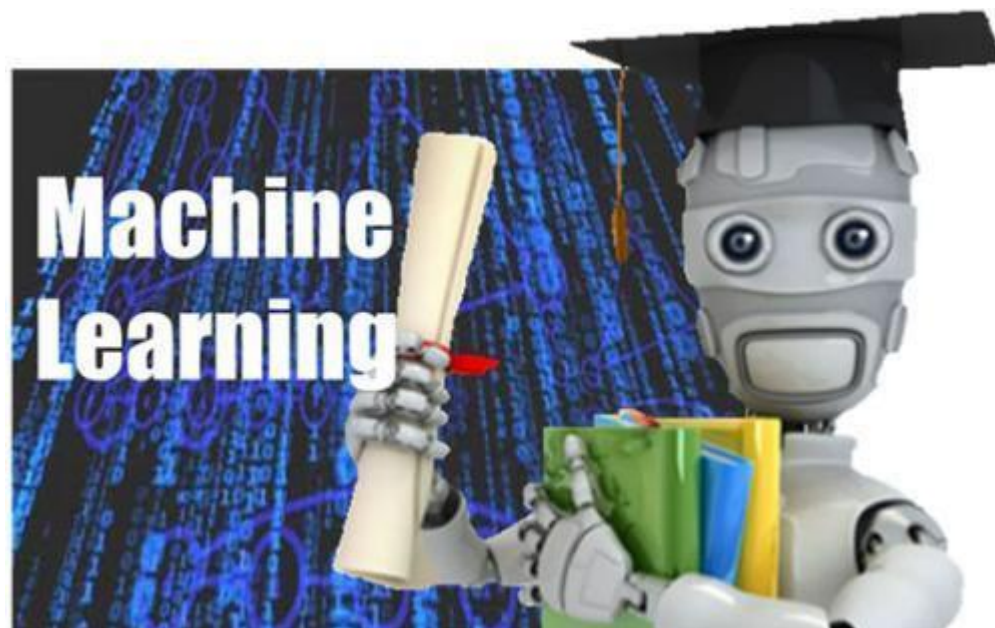
- 错误理解相关关系
 - 事物间的相关性并不意味着存在因果关系，或者有可能其因果关系颠倒了
 - 需要深入理解业务，规避大部分错误
 - 需要分析是否由第三方变量同时引起两种变量的变化，找出其原因
- 错误的比较对象
 - 机器学习中的结果或效果进行比较时，容易将不同样本集进行结果比较，比较对象不合理，其结果自然无效，结论便不能成立
- 数据抽样
 - 数据抽样时如果出现偏差可能会影响分析结果
 - 需要考虑采样标准



- **机器学习概论**
- **机器学习方法及其应用**
- **机器学习常见问题**
- **从事机器学习工作的准备**

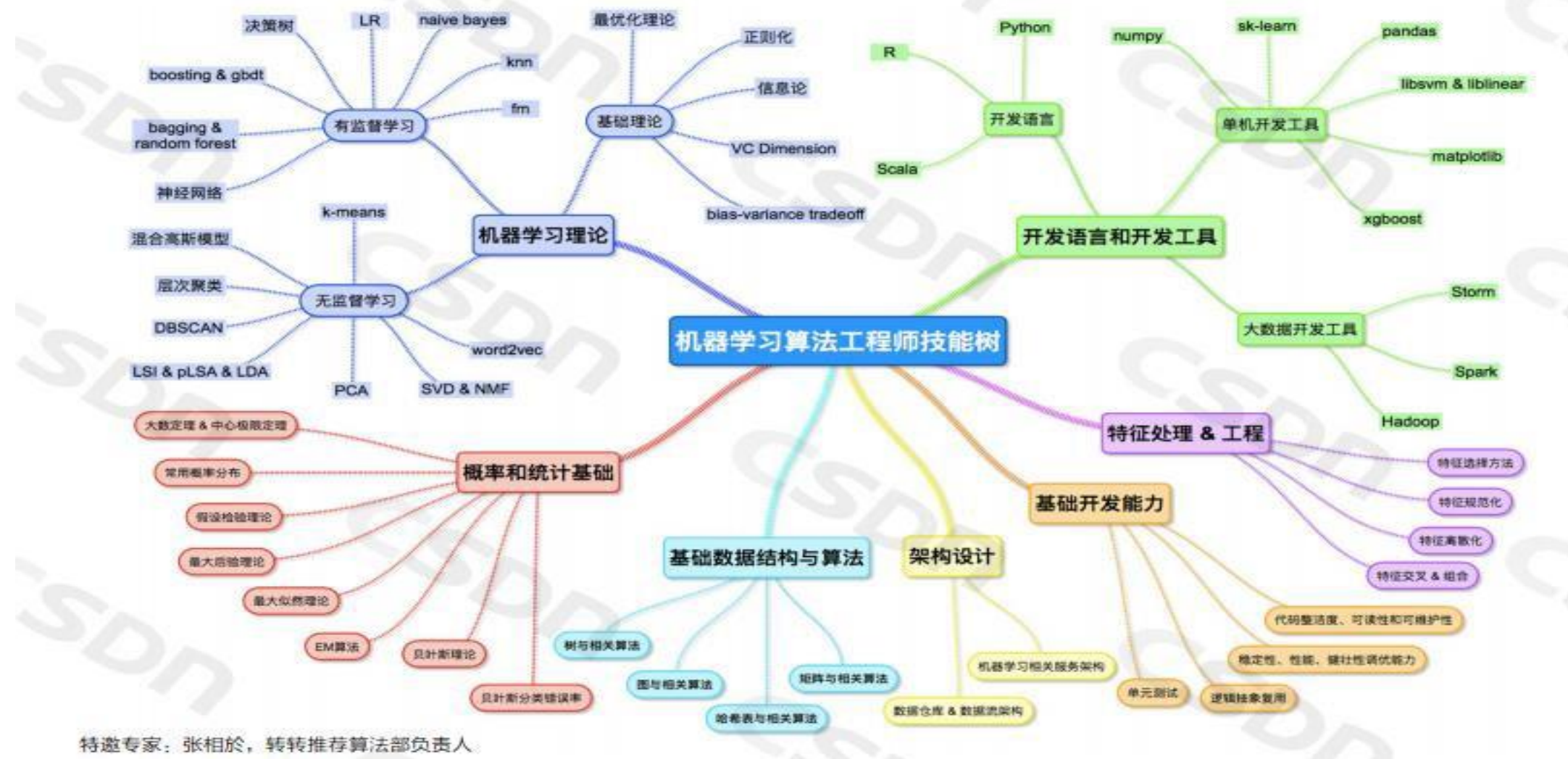
机器学习项目团队的组建

- 职能岗位
 - 项目经理
 - 业务专家
 - 机器学习工程师
 - 数据建模人员
 - 可视化人员
 - 评估人员
 - 其他



我们为什么要学习机器学习?

- 1...
- 2...
- 3...



常见的机器学习语言和平台

- Python
- R
- Pytorch
- Keras
- TensorFlow
- Caffe
- Github
- Kaggle
- Weight and bias
- colab

```
# Basic computational graph
import numpy as np
np.random.seed(0)
import tensorflow as tf

N, D = 3, 4

x = tf.placeholder(tf.float32)
y = tf.placeholder(tf.float32)
z = tf.placeholder(tf.float32)

a = x * y
b = a + z
c = tf.reduce_sum(b)

grad_x, grad_y, grad_z = tf.gradients(c, [x, y, z])

with tf.Session() as sess:
    values = {
        x: np.random.randn(N, D),
        y: np.random.randn(N, D),
        z: np.random.randn(N, D),
    }
    out = sess.run([c, grad_x, grad_y, grad_z],
                    feed_dict=values)
    c_val, grad_x_val, grad_y_val, grad_z_val = out
```

```
import torch
from torch.autograd import Variable

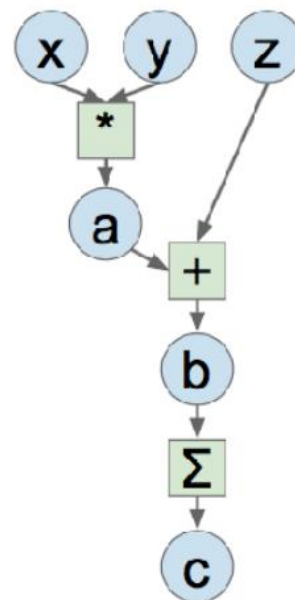
N, D = 3, 4

x = Variable(torch.randn(N, D),
              requires_grad=True)
y = Variable(torch.randn(N, D),
              requires_grad=True)
z = Variable(torch.randn(N, D),
              requires_grad=True)

a = x * y
b = a + z
c = torch.sum(b)

c.backward()

print(x.grad.data)
print(y.grad.data)
print(z.grad.data)
```



代码简洁

keras>pytorch>tensorflow

可读性

pytorch>keras>tensorflow

练习题

1. 如何定义机器学习？
2. 什么是被标记的训练数据集？
3. 最常见的两种监督学习的任务是什么？
4. 列举四种常见的无监督学习任务。
5. 要让一个机器人在各种未知的地形中行走，你会使用什么类型的机器学习算法？
6. 要将顾客分成多个组，你会使用什么类型的算法？
7. 你会将垃圾邮件检测的问题列为监督学习还是无监督学习？