

机器学习与数据挖掘

2024年春季学期

方厚章 田隆

计算机科学与技术学院

机器学习参考教材与资料（“学在西电”课程“资料”栏下载）

● 参考教材（推荐）

1. 《**机器学习**》(1-10章), 周志华, 清华大学出版社, 2016

（书中所有公式推导见南瓜书：<https://datawhalechina.github.io/pumpkin-book/#/>）

2. 《**数据挖掘导论**》(第2版)(1-7,13,14, 16章), 陈封能, 迈克尔·斯坦巴赫, 阿努吉·卡帕坦, 维平·库玛尔, 机械工业出版社, 2019

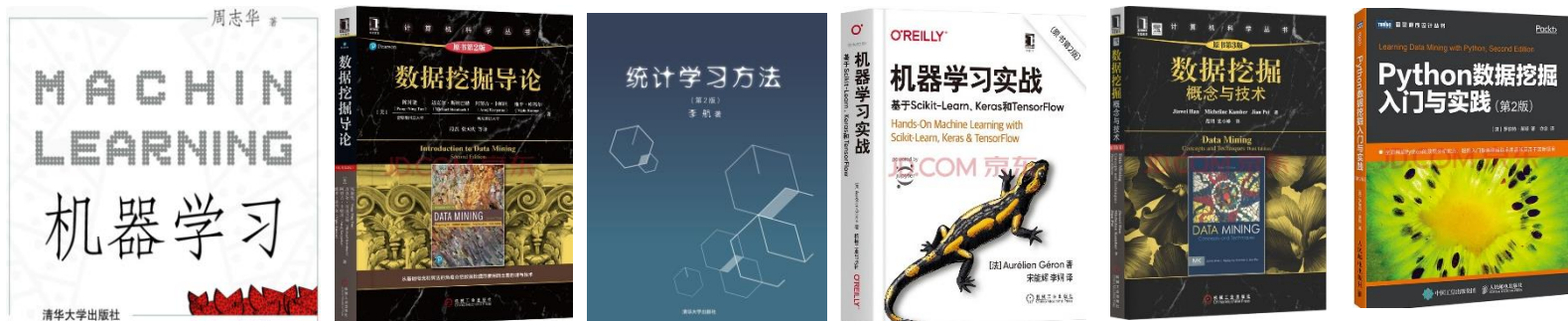
3. 《**统计学习方法**》(第2版), 李航, 清华大学出版社, 2019

● 参考资料

1. 《数据挖掘-概念与技术（原书第3版）》，Jiawei Han, Micheling Kamber, Jian Pei 等著, 范明, 孟小峰译, 机械工业出版社, 2012

2. 《**机器学习实战：基于Scikit-Learn、Keras和TensorFlow**(原书第2版)》，Aurelien Geron著, 王静源, 贾玮, 边蕤, 邱俊涛译, 机械工业出版社, 2020

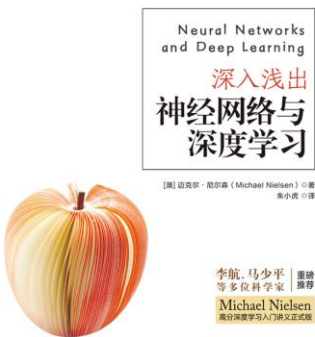
3. 《Python数据挖掘入门与实践》，Robert Layton著, 杜春晓译, 人民邮电出版社, 2016



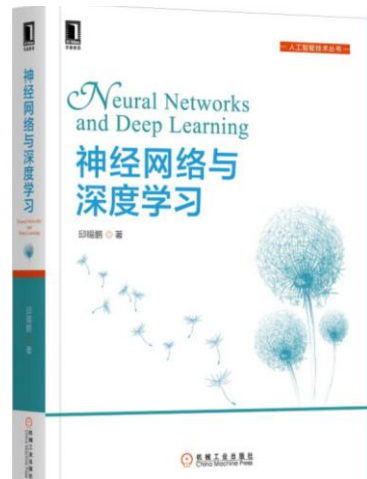
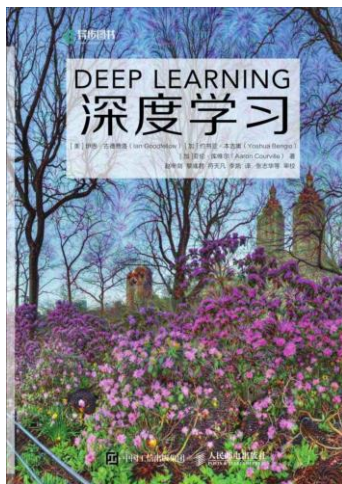
英文参考资料深度学习参考资料

1. 《深入浅出神经网络与深度学习》，迈克尔·尼尔森（Michael Nielsen）著，朱小虎译，人民邮电出版社，2020（入门）
2. 《深度学习》，Ian Goodfellow, Yoshua Bengio, Aron Courville 著，赵申剑，黎劼君等译，人民邮电出版社，2017（进阶）
3. 《神经网络和深度学习》，邱锡鹏著(复旦大学，计算机科学与技术学院)，机械工业出版社，2020（进阶）

图灵 图灵程序设计丛书



中国工信出版集团 人民邮电出版社
POST & TELECOM PRESS



Python 编程

- 安装编译环境 (Anaconda + PyCharm)

(安装流程参考: https://blog.csdn.net/weixin_44337883/article/details/123687561)

- 推荐使用 Anaconda 软件自带Python3.5及以上解释器和其它工具包
- 推荐使用程序编辑软件PyCharm (大型程序、远程调用服务器调试)

社区版 (<https://www.jetbrains.com/pycharm/>)

- 基于浏览器 Jupyter Notebook (练习)
- 云上编程 Google Colab (有限显存的GPU支持)

苦逼学生党的Google Colab使用心得 <https://zhuanlan.zhihu.com/p/54389036>

- PyTorch 框架 (对初学者友好)

- 参考书:

1. 《Python编程从入门到实践》(第3版)(Python基本语法: 1-9章), 埃里克·马瑟斯著, 袁国忠译, 人民邮电出版社, 2023, 中文版

2. 《Deep Learning with PyTorch》, Eli Stevens, Luca Antiga, Thomas Viehmann, Manning, 2020. PyTorch官方推荐. (Eli Stevens, 软件工程师, 硅谷工作15年), 中文版

(解读 https://www.bilibili.com/video/BV1g44y1L7MU?p=4&vd_source=e2a659e4f2341f3776669e0d2d2ec443)

数学基础要求（参考补充资料和参考书）

- 概率论与数理统计
 - 常见概率分布
 - 贝叶斯法则
 - 期望，方差，协方差，中值，最大似然
- 基本线性代数和微积分
 - 向量和矩阵计算
 - 单/多变量/向量值偏导数, 梯度
- 基本算法, 数据结构

- 参考书:
 - 《**Mathematics for Machine Learning**》, Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong, Cambridge University Press, 2020

- [中英字幕]吴恩达机器学习系列课程
https://www.bilibili.com/video/BV164411b7dx?from=search&seid=8554616197240266978&spm_id_from=333.337.0.0
- [双语字幕]吴恩达深度学习
deeplearning.aihttps://www.bilibili.com/video/BV1FT4y1E74V/?spm_id_from=333.788.recommend_more_video.2

课程考核

- 课程内容:

- 《机器学习》（周志华著）(1-10章)
- 《数据挖掘导论》(第2版) (1-9章)
- 其它机器学习或深度学习相关参考资料

- 最终成绩:

- 平时成绩（考勤、大作业、实验报告）40%
- 考试成绩 60%，考试卷面成绩不满50分直接判不及格。
- 作业和实验报告根据完成的实际质量打分。

- 作业提交: 电子版作业发送课代表，随堂作业纸质版课堂收取。作业原则上不接受补交，如果有特殊情况，确需推迟提交作业，请提交请假条

- 作业抄袭处罚: 作业要自己独立完成，布置作业后尽早完成。发现任何一次作业抄袭/被抄袭者，按照本科生院规定双方平时分将严重扣分！

- 课件与课程辅助资料:

(1) 西电课堂“资料”栏，(2) QQ群 (277264431)， “群文件”

第六条 课程应加强过程性评价，平时成绩在总成绩中应占一定的比例，一般应高于 20%，低于 60%。单纯的课堂考勤不作为记分项。期末考核成绩低于 50 分(不含 50 分)的，课程总评成绩判定为不合格，期末成绩即为该课程的总评成绩。

第九条 学生在一学期内无故缺课累计超过某门课程教学时数的 25%，或在任课教师课堂上采用各种方式随机抽查中，一学期有三次无故未到，取消该生正常考试资格，该课程正考的成绩以不合格记载。

第十六条 成绩一经提交，原则上不得更改。确因特殊原因需要更改的，由任课教师提交申请，说明更改理由并附相关证明材料，依次经开课单位审核、本科生院审批通过后，方可变更成绩。属于教师疏忽等主观因素导致错误成绩，按照学校相关规定认定教学事故。原则上成绩变更期限是开课学期后的下一学期两周内，逾期不予受理。

- 绪论 (Introduction)
- 数据和特征工程 (Data and Feature Engineering)
- 线性模型 (Linear Model)
- 决策树 (Decision Tree)
- 神经网络 (Neural Network)
- 支持向量机 (Support Vector Machine)
- 关联分析 (Association Analysis)
- 聚类分析 (Cluster Analysis)

机器学习 概述

什么是机器学习?

- 如果你是一个科学家



- 如果你是一个工程师/企业家
 - Get lots of data
 - Machine Learning
 - ???
 - Profit!

什么是机器学习?

- **“机器学习是对能通过经验自动改进的计算机算法的研究”。**

-Tom Mitchell 《机器学习》(1997)

- **“机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。**

”

- (土)埃塞姆 阿培丁(Ethem Alpaydin)
《机器学习导论》第3版

为什么研究机器学习？-设计更好的计算系统

● 开发系统

- 手动构建太困难/昂贵
- 需要特定的详细技能/知识
- *知识工程瓶颈*

● 开发系统

- 根据个人用户进行调整和定制
- 个性化的新闻或邮件过滤器
- 个性化辅导

● 从大型数据库中发现新知识

- 医学文本挖掘（例如，偏头痛到钙通道阻滞剂到镁）
- *数据挖掘*

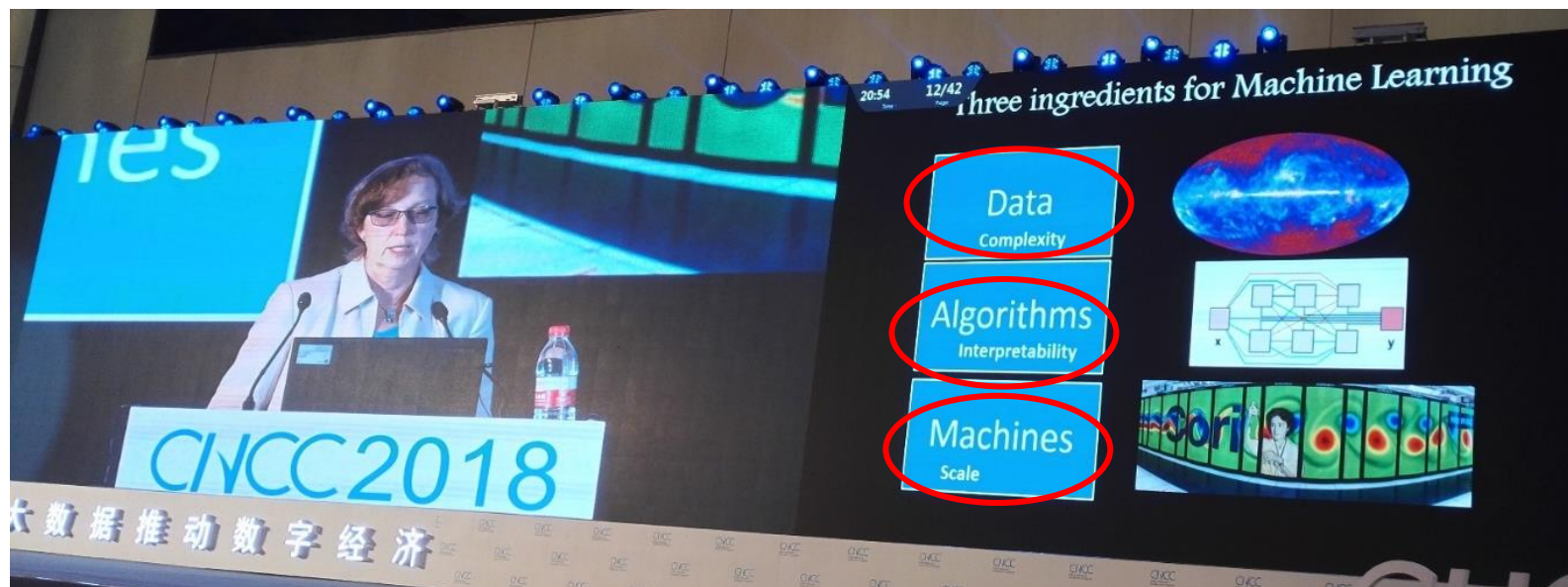
为什么研究机器学习？-认知科学（Cognitive Science）

- 学习的计算研究可以帮助我们理解人类的学习
 - 以及其他生物有机体
- 从大型数据库中发现新知识
 - 医学文本挖掘（例如，偏头痛到钙通道阻滞剂到镁）
 - **数据挖掘**

为什么研究机器学习？-时机已成熟

- 数据（Data）
 - 大量的数据可获得
- 算法（Algorithms）
 - 许多高性能和高效的算法可以获得
- 计算资源（Computing）
 - 大量的计算资源可获得（GPU）

机器学习三要素-大数据、算法、计算资源



2018中国计算机大会 (CNCC) (杭州萧山)

Katherine Yelick-美国工程院院士，计算机科学家，加州大学伯克利分校电子工程和计算机科学教授，劳伦斯伯克利国家实验室计算科学实验室副主任

人工智能技术进步的核心推动力-机器学习技术 (周志华, 南京大学)



南大周志华 V

2017-12-7 09:05 来自 HUAWEI Mate 9 Pro

最近有个说法, 认为人工智能技术的进步, 例如AlphaGo系列的成功, 主要是由计算能力带来的。这个说法绝对是错误的! 最重要的进步是由机器学习技术的进展带来的, 计算能力起到了促进作用而不是根本作用。举个例子: 深蓝下国际象棋略超过卡斯帕罗夫, 其技术是以专家规则为主机器学习为辅, 每秒需评估6亿个位置, 所以必须使用IBM的专用设备; 今天以机器学习为主的国际象棋程序, 例如Pocket Fritz 4 实力超过卡斯帕罗夫曾达到的最高等级分, 仅需每秒评估2万个位置, 所以在手机上就能运行。即便AlphaGo面对的是复杂得多的围棋, 也仅需每秒评估6万个位置。从6亿减少到2万, 这是机器学习算法带来的提高, 更不用说计算过程的目标方向已经有了根本的改变 收起全文 ^

以模型为中心到以数据为中心 (From Model-Centric to Data-Centric)

Andrew Ng 12:43 AM · Apr 18, 2021 · Twitter Web App

This Sunday is my birthday! The best gift 🎁 to me would be if you can watch this video and let me know what you think. [youtube.com/watch?v=06-AZX...](https://www.youtube.com/watch?v=06-AZX...)

Lets you and I work to shift AI from Model-Centric toward Data-Centric AI development, which will help many teams.

Data is Food for AI

PREP	ACTION
Source and prepare high quality ingredients	Cook a meal
Source and prepare high quality data	Train a model

- MLOps' most important task: Ensuring consistently high-quality data in all phases of the ML project lifecycle
- 用一个公式总结:
- **Better AI = Data (80%) + Code(20%)**

- “If 80 percent of our work is data preparation, then ensuring data quality is the important work of a machine learning team.”
----- Andrew Ng

机器学习的分类

监督学习
(Supervised Learning)

Labeled data
Direct feedback
Predict outcome/future

无监督学习
(Unsupervised Learning)

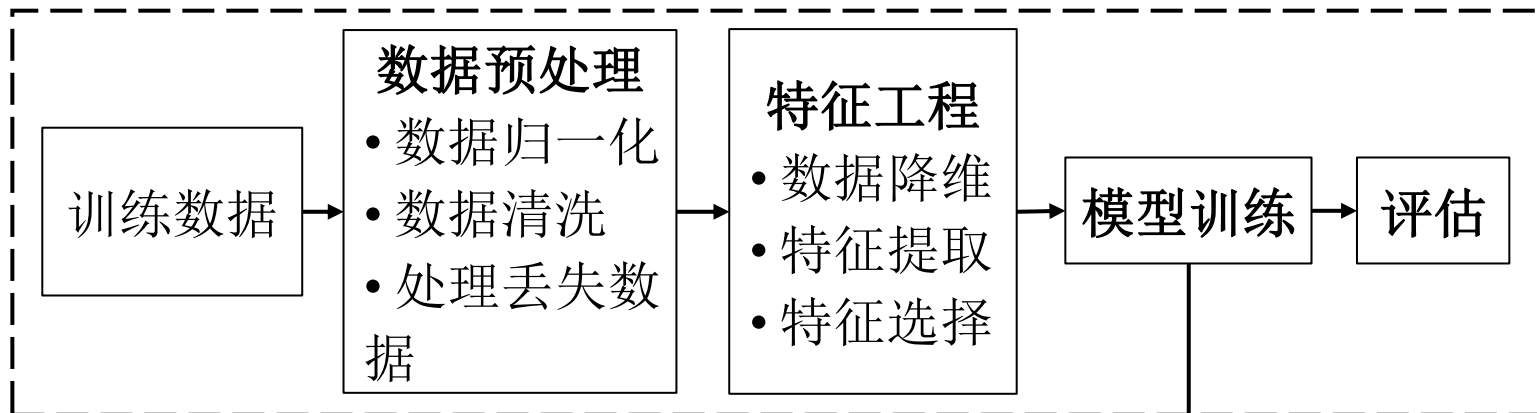
No labels/targets
No feedback
Find hidden structure in data

强化学习
(Reinforcement Learning)

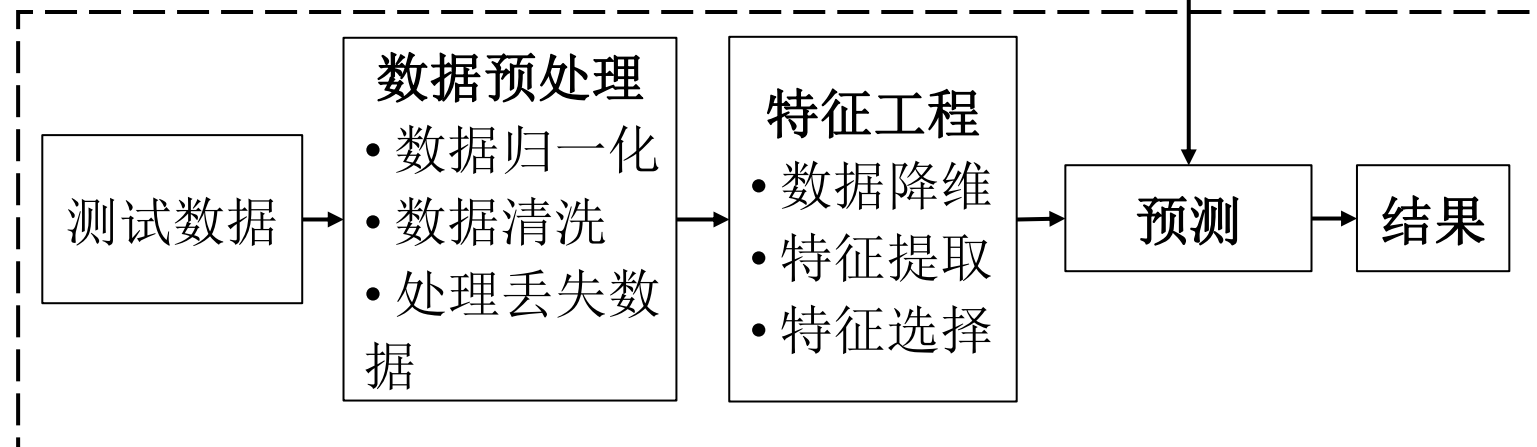
Decision process
Reward system
Learn series of actions

机器学习基本处理流程

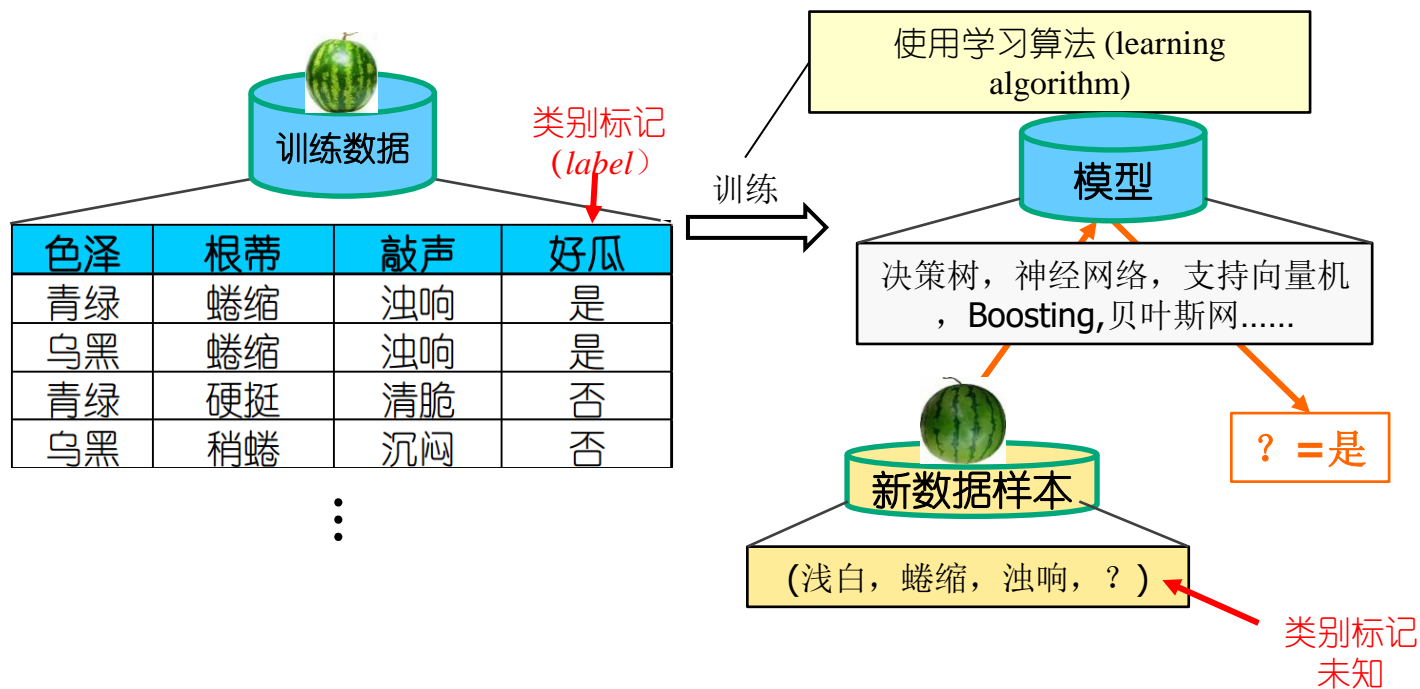
训练/学习阶段



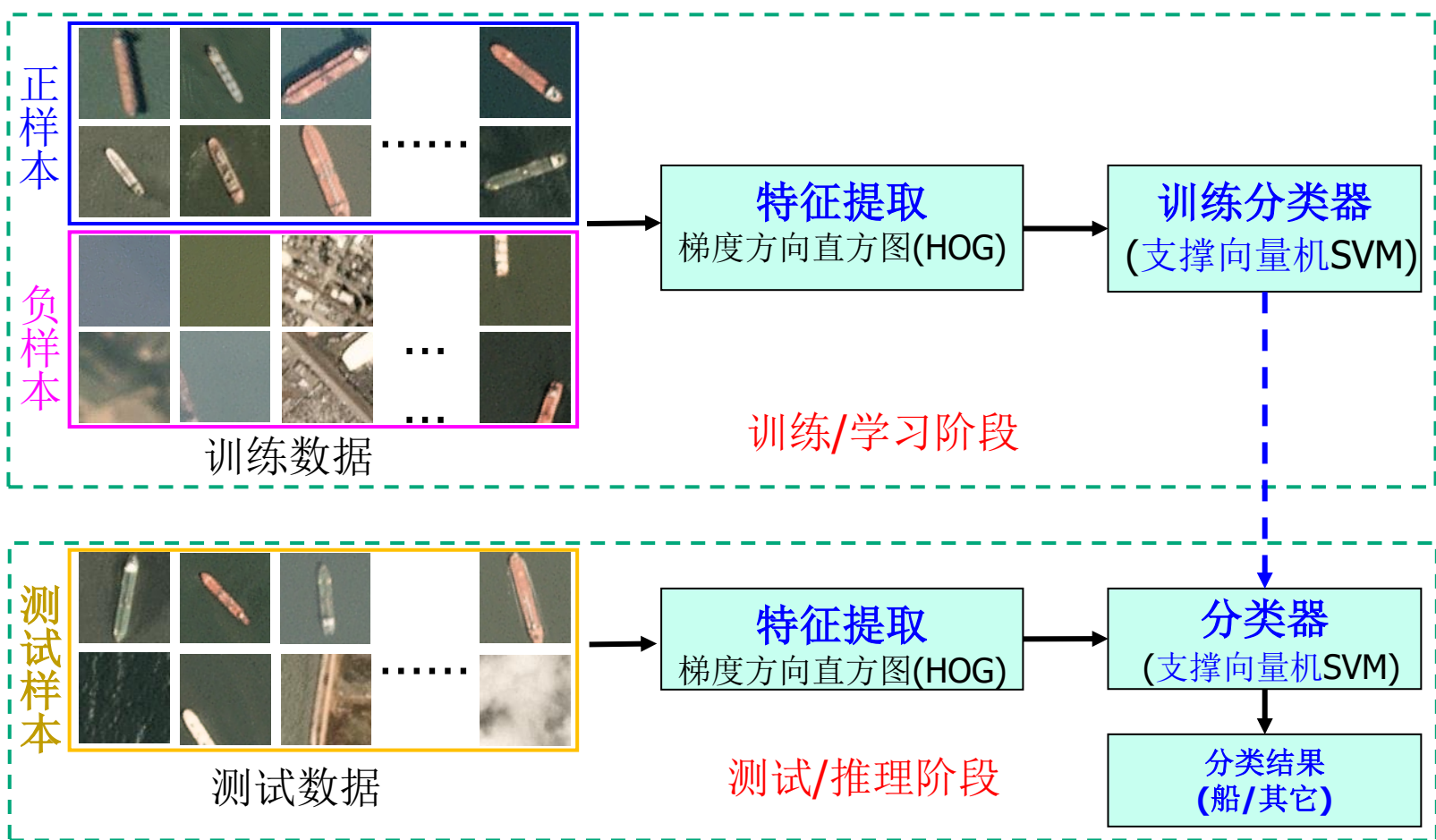
测试/应用阶段



典型的机器学习过程-西瓜的二分类



传统机器学习实例-卫星图像中的船舶分类算法

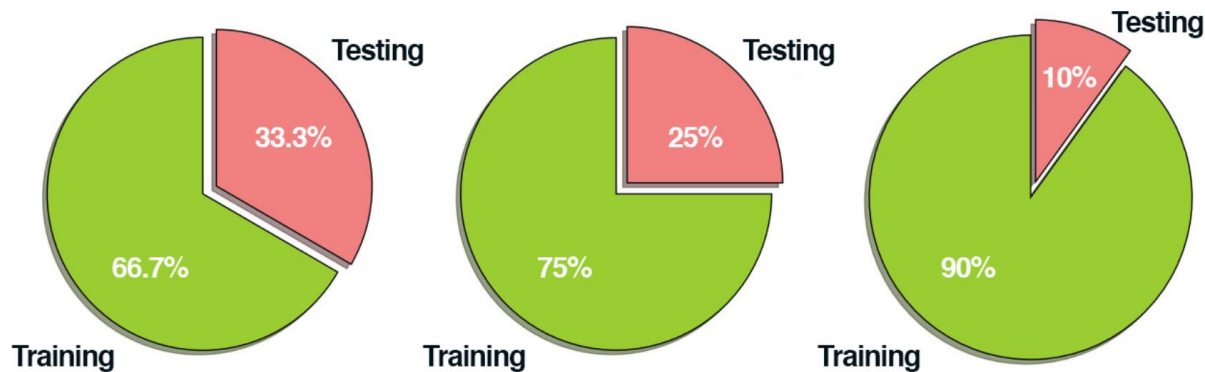


机器学习的具体流程:

1. 训练和测试集划分
2. 模型的训练、测试和评估

机器学习的流程-数据集划分

- 第 1 步: 收集数据集。标记数据。
- 第 2 步: 拆分数数据集。
- 训练集 (training set)-训练模型，确定模型/网络参数
- 验证/开发集 (validation/ development set)-调整超参数(学习率，正则化参数 等)，选择特征，以及对学习算法作出其它决定
- 测试集 (test set)-评估算法的性能



训练集、开发集、测试集比例划分

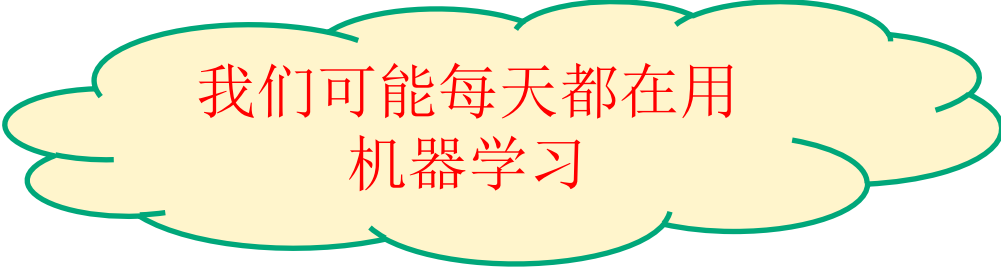
- 没有一般规则, 指明训练集、验证集和测试集比例各占多少是合适的 (依赖于训练样本的容量和数据的信噪比)
 - 当样本数量不多 (小于1万) 的时候, 通常将训练集/验证集/测试集的比例设为60%:20%:20%
 - 在没有验证集的情况下, 训练集/测试集的比例设为70%:30%
 - 当样本数量很大 (百万级别) 的时候, 通常将相应的训练集/验证集/测试集比例设为98%:1%:1%或者99%:1%(训练集/测试集)
- 没有一般规则, 指明多少数据是足够的 (依赖于基础函数的信噪比和拟合数据的模型的复杂度)
 - 验证集的规模应该尽可能大, 至少要能够区分出你所尝试的不同算法之间的性能差异。通常来说, 验证集的规模应该在1000 到 10000 个样本数据之间

机器学习的流程-训练和测试

- 第 3 步: 训练模型。（对于深度学习，随机梯度降）。
- 第 4 步: 评价。(查准率(精度)，查全率(召回率)，F1-度量)

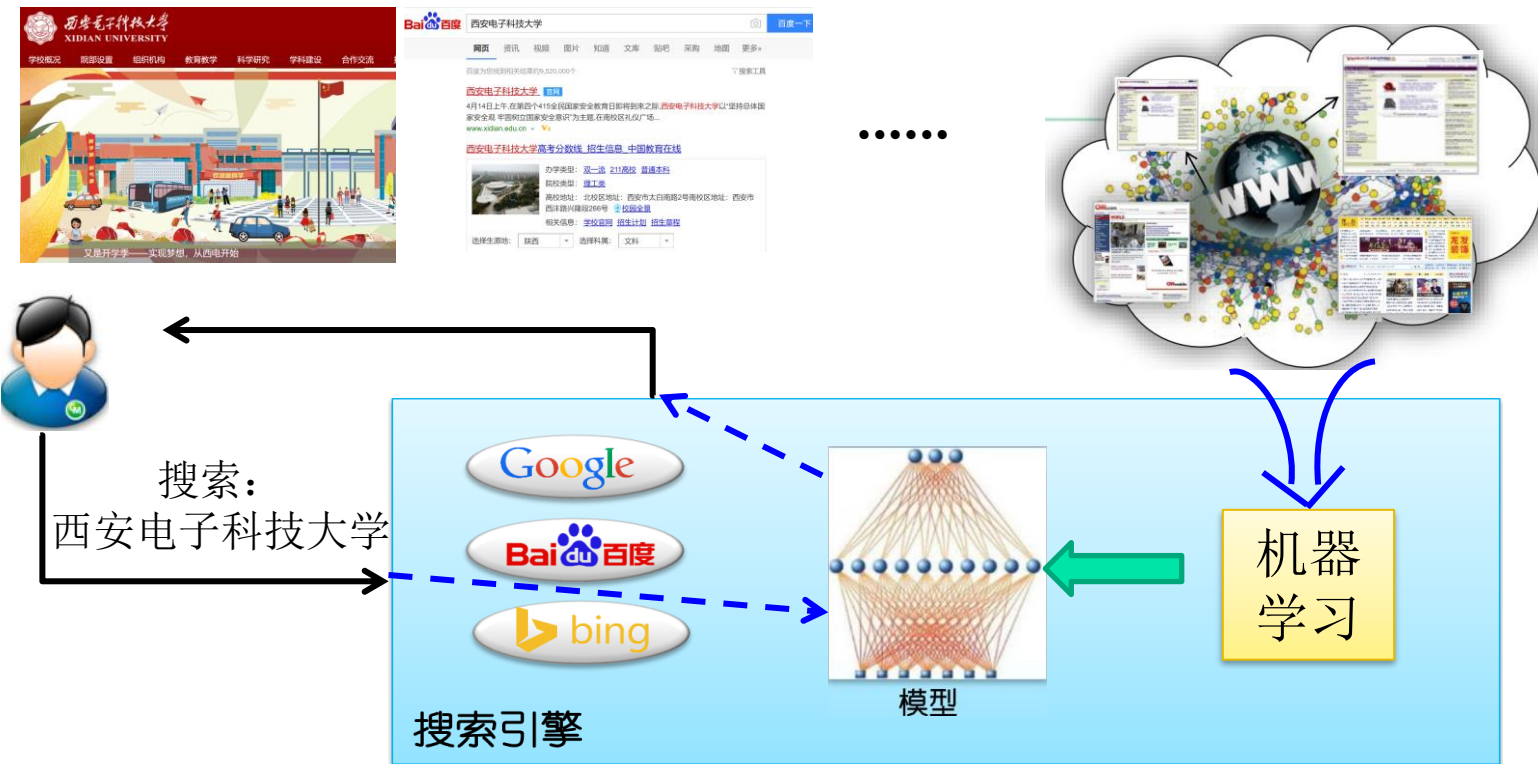
注意：训练集 \cap 验证/开发集 \cap 测试集 = \emptyset （空集）

机器学习能做什么？



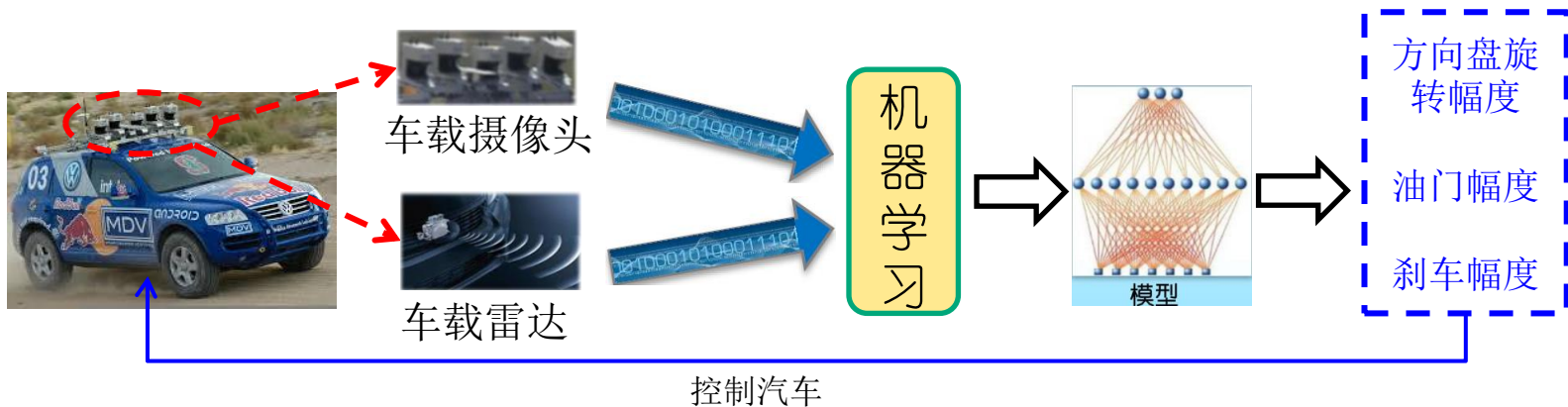
我们可能每天都在用
机器学习

例如：互联网搜索

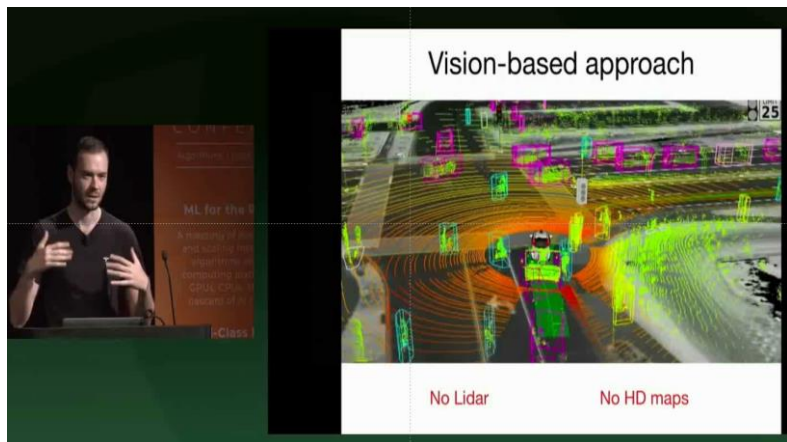


机器学习/深度学习技术正在支撑着各种搜索引擎

例如：自动驾驶（即将改变人类生活）



Google Waymo's fully self-driving cars are here—
<https://www.youtube.com/watch?v=aaOB-ErYq6Y>



Tesla Autopilot and Multi-Task Learning for Perception and Prediction-前特斯拉AI总监 Andrej Karpathy

视觉 (Vision)

- Object detection, instance segmentation

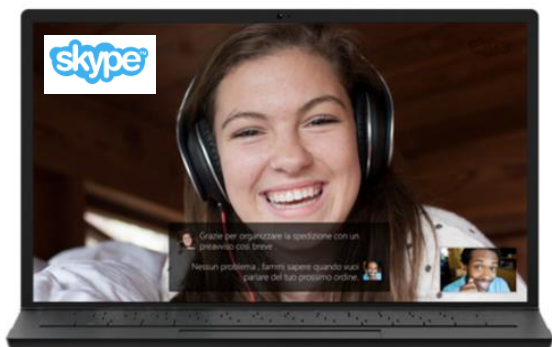


Object detection
(无人机对地车辆识别与跟踪)



K. He, G. Gkioxari, P. Dollar, and
R. Girshick, [Mask R-CNN](#),
ICCV 2017 (Best Paper Award)

语音和自然语言处理 (Speech and natural language)



Skype Translator

Break down the language barrier with your friends, family and colleagues.

Our online translator can help you communicate in 7 languages for voice calls, and in more than 50 languages while instant messaging.

Skype Translator uses machine learning. So the more you use it, the better it gets. Thanks for being patient as the technology graduates from Preview mode.

<https://www.skype.com/en/features/skype-translator/>



Google Translate App

- Translate between 103 languages by typing
- Offline: Translate 52 languages when you have no Internet
- Instant camera translation: Use your camera to translate text instantly in 30 languages
- Camera Mode: Take pictures of text for higher-quality translations in 37 languages
- Conversation Mode: Two-way instant speech translation in 32 languages

<https://play.google.com/store/apps/details?id=com.google.android.apps.translate&hl=en>

See also: [The Great AI Awakening](#) (New York Times Magazine)

数据挖掘 概述

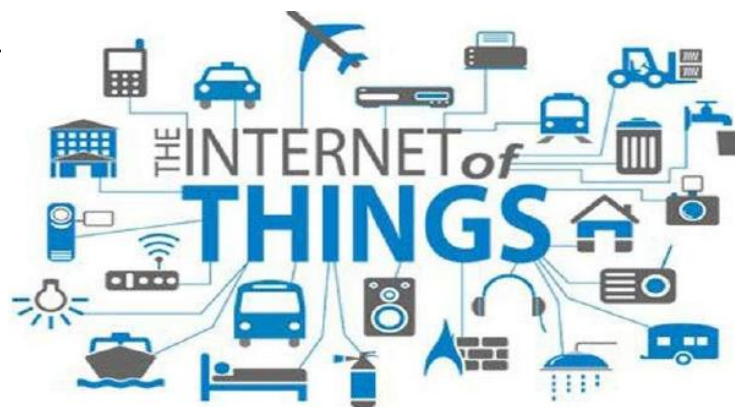
为什么进行数据挖掘？-商业角度

- 企业收集和存储海量数据
 - 百货公司/杂货店购买数据
 - 银行/信用卡交易
 - 网络和社交媒体数据
 - 移动和物联网
- 计算机更便宜，功能更强大
- 商业竞争以提供更好的服务
 - 大规模定制和推荐系统
 - 有针对性的广告
 - 改善物流



2015年全球移动终端
产生的数据量

6300PB



物联网（IoT）

无所不在的数据收集

信息基础设施



GPS



一卡通



手机

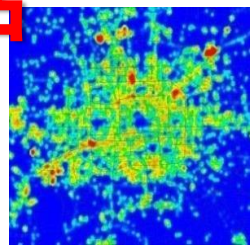
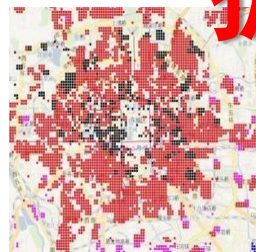
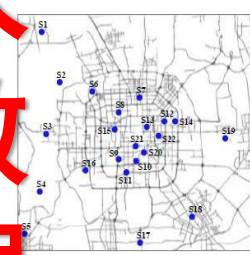
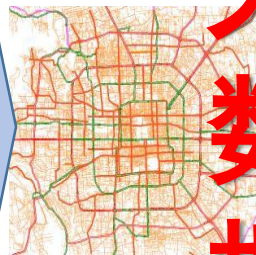
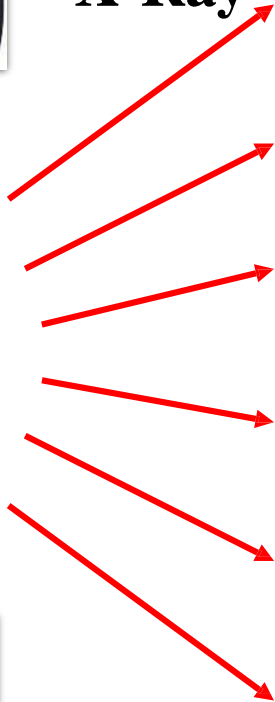


传感器



视频监控

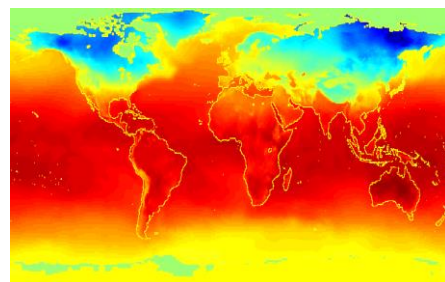
“X-Ray”



大数据

为什么进行数据挖掘？-科学研究角度

- 以极快的速度收集和存储数据（GB/小时）
 - 卫星上的远程传感器
 - 扫描天空的望远镜
 - 微阵列生成基因表达数据
 - 科学仿真生成 TB 级数据
- 数据挖掘可以帮助科学家
 - 识别模式和关系
 - 对数据进行分类和分段
 - 提出假设



什么是数据挖掘？

One of many definitions:

*"Data mining is the science **of extracting useful knowledge** from huge data repositories."*

ACM SIGKDD, Data Mining Curriculum: A Proposal

<http://www.kdd.org/curriculum>

什么是数据挖掘

- 在大型数据存储中，自动地发现有**有用**信息的过程
 - 探查大型数据集，发现先前**未知**的**有用**信息
 - 或是预测未来观测结果

如：预测某新客户是否会在一家商场消费**100**元以上？

- 更严谨的表述
 - **数据挖掘**就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取**隐含在其中的、人们事先不知道的、但又是潜在有用**的信息和知识的过程。

什么（不）是数据挖掘

- **非数据挖掘**

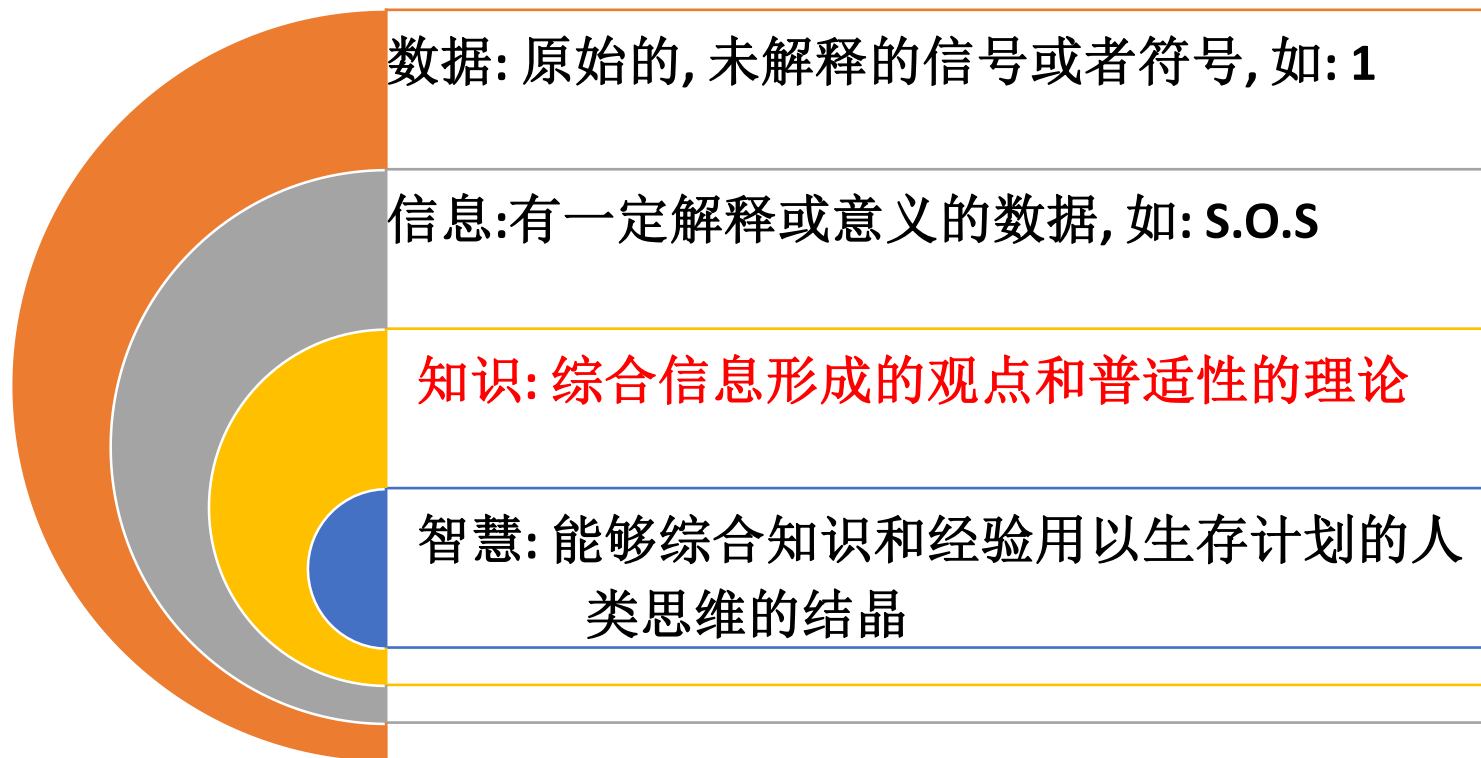
- **从电话簿查找电话号码**
- **从Web中查找信息 “数据挖掘”**
- **获得职工的平均资薪**
-

- **数据挖掘**

- **某插班生应该读几年级？**
- **买哪只股票更可能挣钱？**
- **怎么才能多卖化妆品？**
- **海量文档该如何归类？**
- **行驶车辆如何预警？**
- **广告如何派送更好？**
-

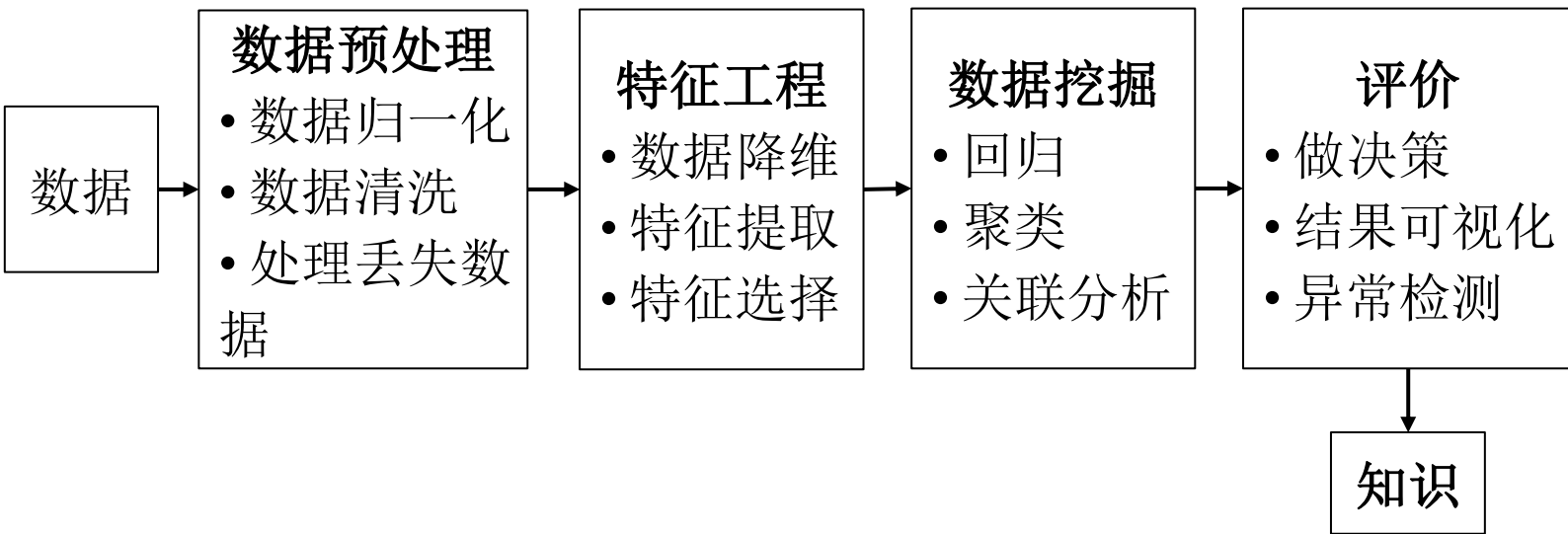
数据挖掘的核心任务是知识发现

- Knowledge Discovery in Database (KDD)



从数据中获取知识！
为决策提供支持！

数据挖掘基本流程



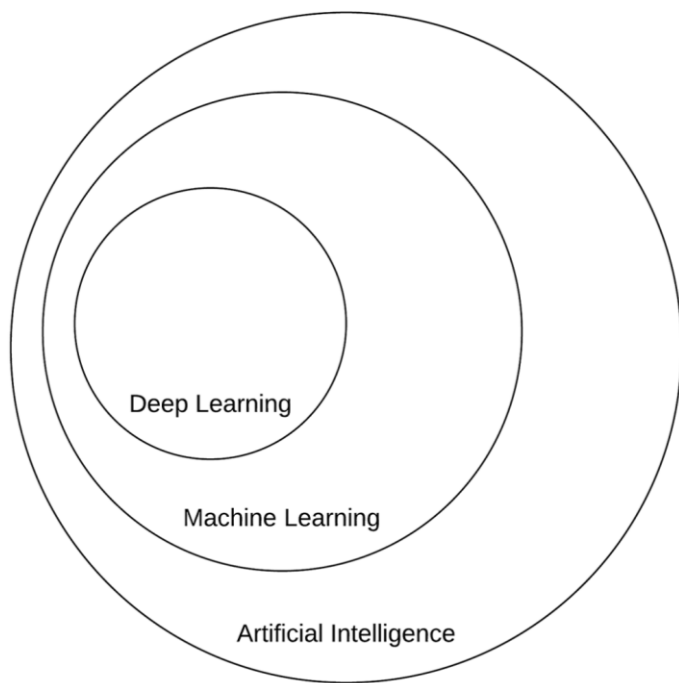
机器学习与数据挖掘的区别

- **机器学习**是人工智能的一个分支，旨在使系统从提供的数据中自动学习，并随着时间的推移改进它们的学习，而无需明确编程。它被用作一种数据挖掘技术。
- **数据挖掘（Data Mining or Data Science）**侧重于分析数据并从中提取知识和/或未知的有趣模式。目标是了解数据中的模式以解释某些现象，而不是开发一个可以预测未知/新数据结果的复杂模型。

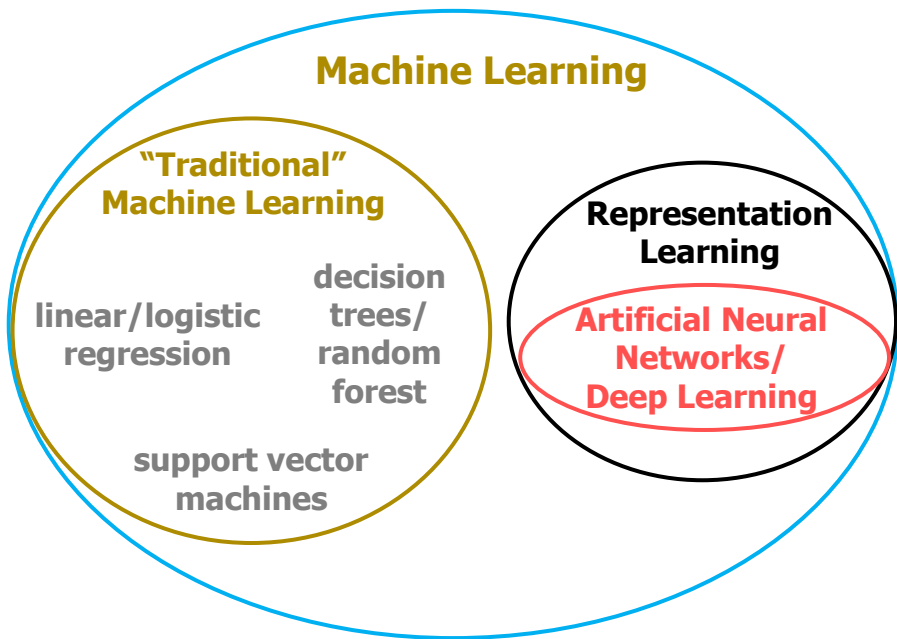
提取的知识可以进一步用于商业应用，例如，可以对现有数据使用数据挖掘来了解公司的销售趋势，然后构建机器学习模型以从该数据中学习，找到相关性并适应新数据。

- **两者的相似性**
 - **机器学习**通常被视为更接近人工智能。
 - **数据挖掘**通常被视为更接近软件工程。
- **深度学习**是机器学习的一个子领域，其中模型是神经网络

机器学习、深度学习与人工智能



机器学习，深度学习与人工智能

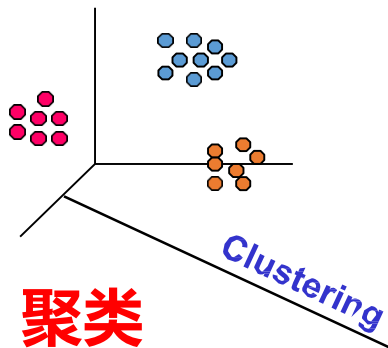


“传统”机器学习与表示学习

数据挖掘的任务类型

- **预测问题**：预测对象的未知特性 → 预测任务（分类和回归）
 - **聚类问题**：获取数据中未知模式
 - **关联分析**：获取未知的关联关系
 - **异常检测**：获取未知的数据异常
- 描述任务

数据挖掘的任务类型



Data

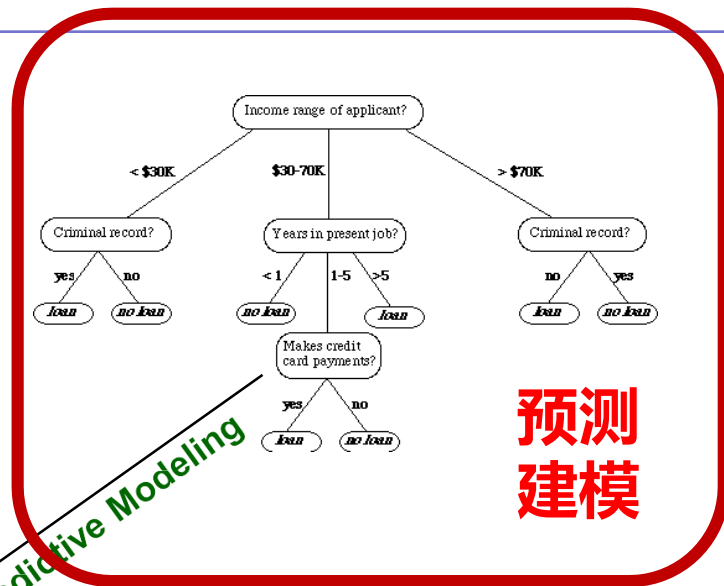
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

关联
分析



Association
Analysis

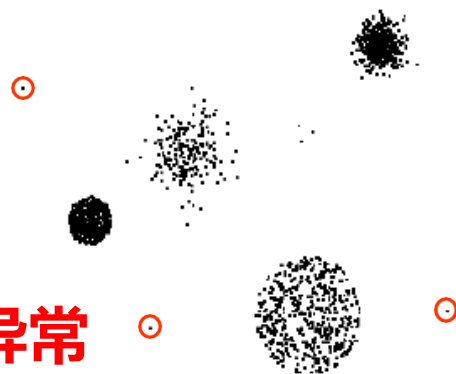
Predictive Modeling



预测
建模

Anomaly
Detection

异常
检测

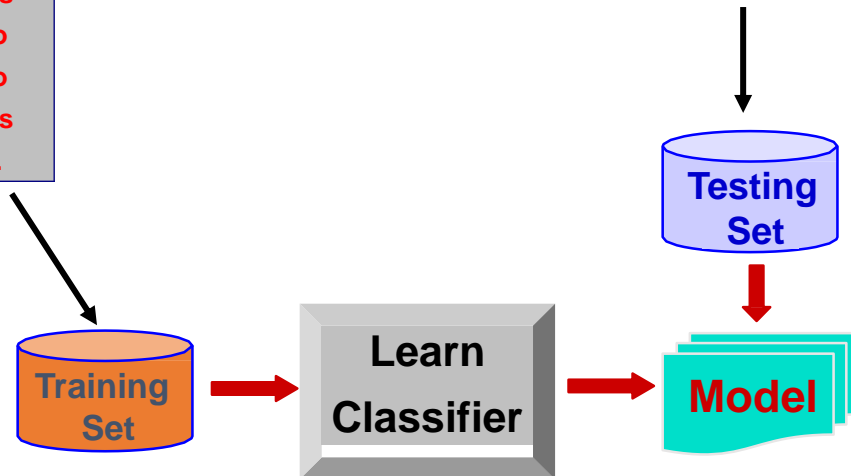


分类问题: Classification

categorical categorical quantitative class

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...



问题的定义

- 训练集 (**training set**)
 - 给定一组对象，该对象可以用一组属性(**attributes**)和一个类别(**class**)标记进行描述。
- 分类模型
 - 寻找一个能够描述**attributes**和**class**关系的函数，该函数以**attributes**为输入，以**class**为输出。
- 目标：对只知道**attributes**而不知道**class**的对象分配尽可能准确的**class**估计。
 - 需要引入一个专门的测试数据集**test set**对模型的准确性进行评估。
 - 在分类问题中，通常需要将数据分为两个部分，一部分用作训练集，另一部分用作测试集。

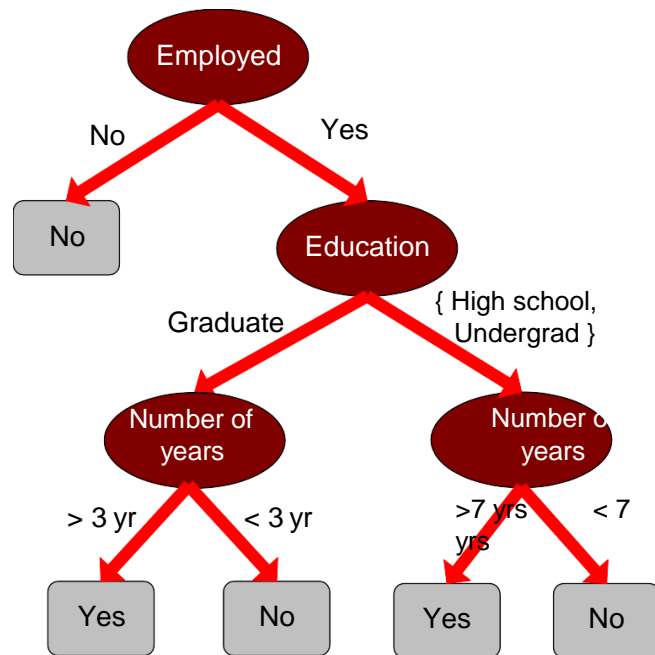
分类问题：Classification

• 核心思想

- 寻找一个模型使其能够对**特征与类别标记之间的关系**进行描述，并用该模型对没有类别标记的特征对象进行分类。

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

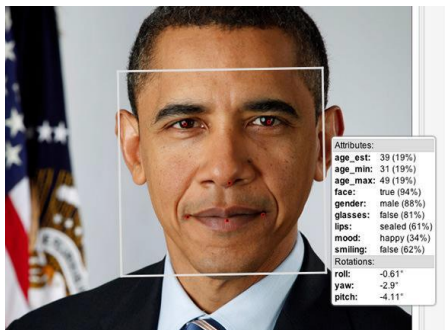


垃圾邮件过滤

- 对象：邮件
- 特征：邮件中的单词
 - 文本向量化 \rightarrow {单词1的数量, 单词2的数量,}
- 训练标注：{是,否}为垃圾邮件，人工标注的训练集
- 输出：一封新邮件是否是垃圾邮件的**概率**

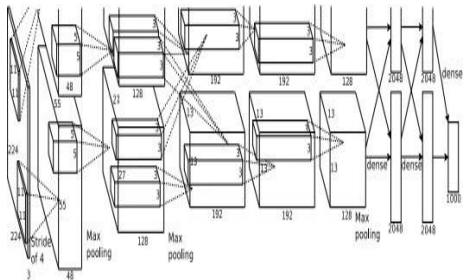
分类问题的应用

图
像
识
别
问
题



人脸照片

+



深度学习模型

输出结果

Obama

Hillary

Trump

Bush

.....

ImageNet

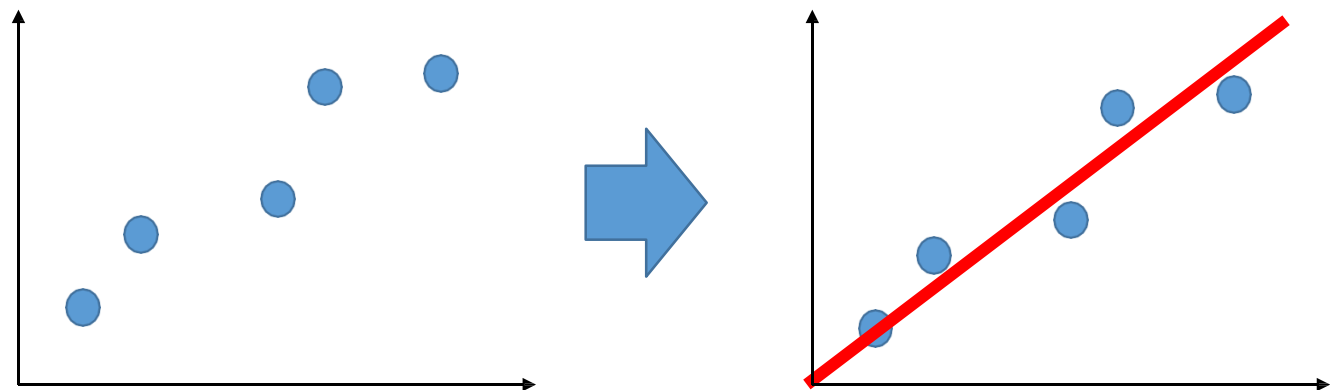
世界纪录

9x.x%

分类技术的分类

- 基本分类模型
 - 决策树 Decision Tree based Methods
 - 规则学习 Rule-based Methods
 - 最近邻 Nearest-neighbor
 - 神经网络 Neural Networks
 - 贝叶斯方法 Naïve Bayes and Bayesian Belief Networks
 - 支持向量机 Support Vector Machines
- 集成方法
 - 提升方法 Boosting, Bagging, 随机森林 Random Forests

回归问题 Regression



regression



英 [rɪˈɡreʃn]  美 [rɪˈɡreʃən] 

n. 回归；衰退：（尤指因催眠或精神疾患，或为逃避目前忧虑）回到从前；

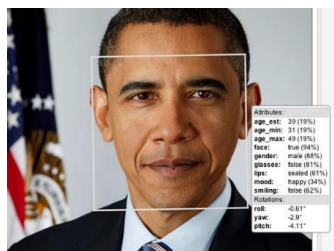
（统计学）回归

回归问题 Regression

- 训练集 **training set**
 - 给定一组对象，该对象可以用一组**特征属性**和一个**被预测属性**（**连续变量**）进行描述。
- 回归模型
 - 寻找一个能够描述**特征属性**和**被预测属性**关系的函数，该函数以**特征属性**为输入，以**被预测属性**为输出。
- 目标：最小化模型预测值和真实被预测属性之间的差，例如用均方误差度量。
- Regression和Classification之间核心区别是被预测属性**是否连续**。

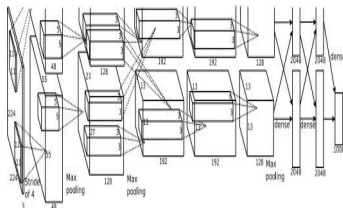
图像识别 vs 房价预测

图像识别问题



人脸照片

+



深度学习模型

输出结果

Obama

Hillary

Trump

Bush

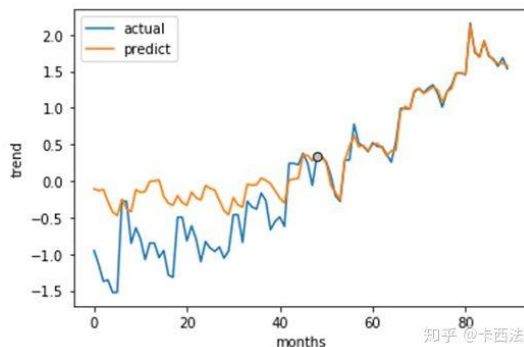
.....

ImageNet

世界纪录

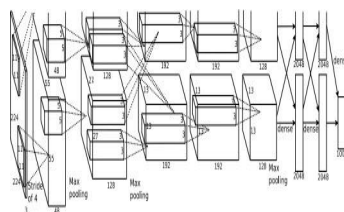
96.5%

房价预测问题



过去90个月全国平均房价变化情况
(橙线是预测值, 蓝线是真实值,
灰点后48个月是训练集, 灰点前
48个月是测试集)

+



深度学习模型

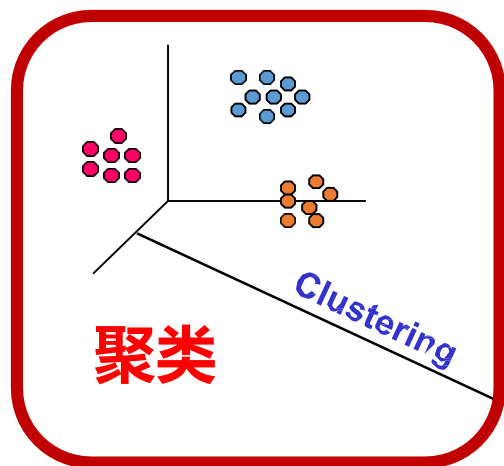
输出结果

高



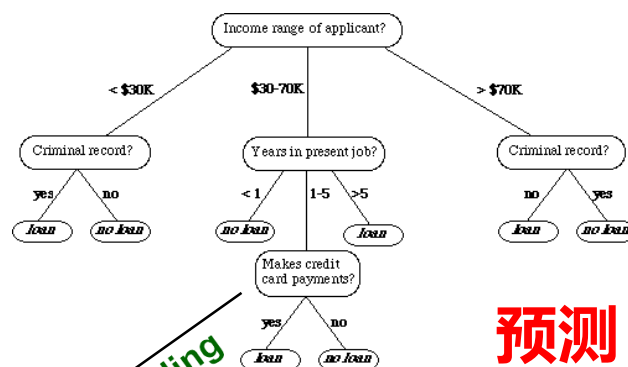
低

数据挖掘的任务类型



Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



预测
建模

关联
分析

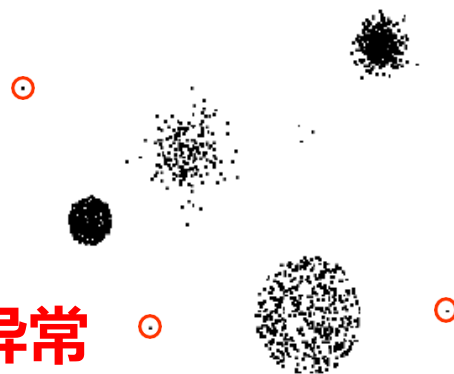


Association
Analysis



Anomaly
Detection

异常
检测



聚类问题 Clustering



数据集



聚类结果

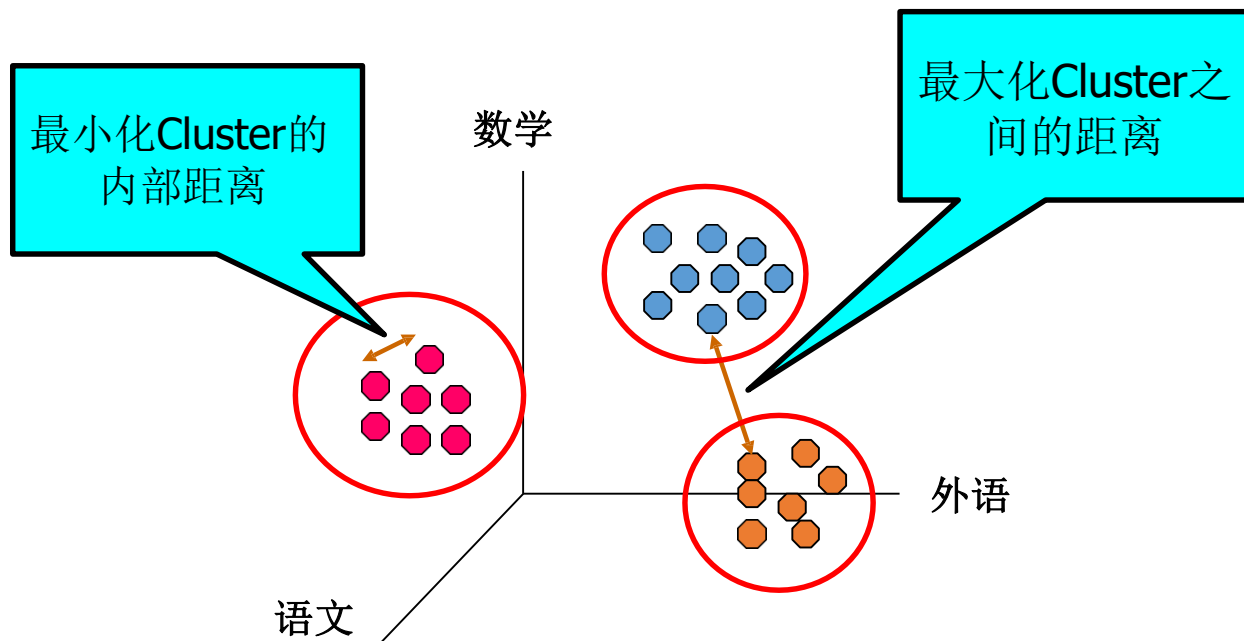
聚类问题的定义

- 给定一组数据点
 - 两个数据点之间可以用一种距离(distance)进行度量。
- 目标：寻找一组数据点的组合(cluster, 簇)使得
 - 同一Cluster内部的点之间的距离尽可能的小；
 - 不同Cluster之间的点之间的距离尽可能的大。
- 度量距离的设计
 - 每个点均可以使用一组属性(attributes)/特征进行描述；
 - 如果描述数据点的属性是连续值，那么可以使用属性空间的欧式距离进行度量；
 - 针对问题场景设计不同的距离度量方法。

聚类问题 Clustering

• 核心思想

- 对给定对象集寻找一种分组方式，使得**组内**的各个对象尽可能的相似，**组间**的对象差异尽可能的大。



聚类分析的功能

- 帮助理解数据的特征
 - **物以类聚人以群分**，用聚类算法可以 **“看一看” 数据中到底有什么。**
 - 将基因和蛋白质根据功能的相似性进行聚类
 - 将股票根据相似的价格波动幅度进行聚类
- 降低数据分析的难度
 - 降低数据集的规模，对于每一个聚类采用一种处理方式
 - 对相似的文本进行聚类，只浏览主题符合个人兴趣的文章

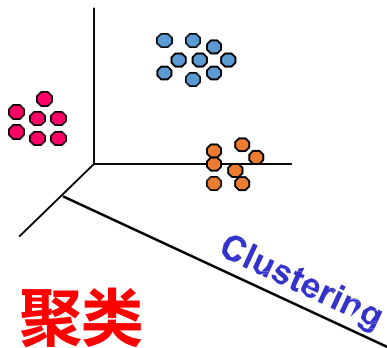
什么不是聚类分析

- 依据简单规则的数据对象划分
 - 例如入大学之后对给位同学进行分班，在世界杯根据抽签结果对参赛队伍进行分组等。
 - 没有考虑到个体之间距离的因素。
- 根据外部属性进行简单划分
 - 例如根据籍贯、民族对人口进行族群划分，根据年龄将人分为少年、中年、老年等。
 - 缺少必要的聚类建模过程。
- 从外部标签学习获得的分类模型(classification)
 - 数据集本身具有明确的类型标签，根据标签训练模型对数据进行划分。
 - 数据类型划分标准是通过外部信息获得的，而非数据集本身。

聚类分析的核心特点

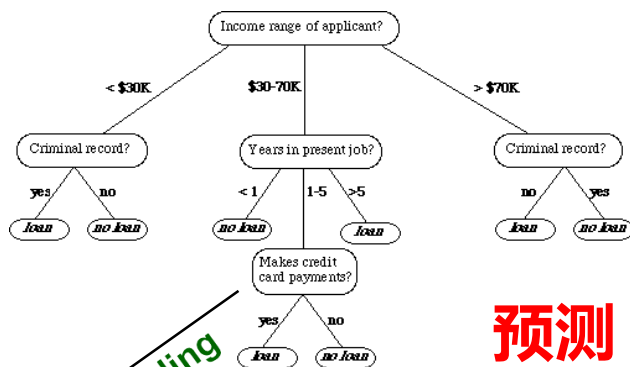
- 根据聚类对象在特征空间的距离
- 对数据进行聚类建模模型
- 将数据无监督的划分为若干组或簇(cluster)

数据挖掘的任务类型



Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



预测
建模

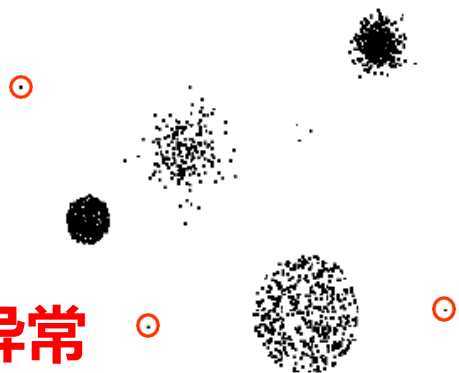
关联
分析

Association
Analysis



Anomaly
Detection

异常
检测



关联规则发现 (Association Rule Discovery)

销售
记录

<i>TID</i>	<i>Items</i>
1	面包(Bread), 可乐(Coke), 牛奶(Milk)
2	啤酒(Beer), 面包(Bread)
3	啤酒(Beer), 可乐(Coke), 尿布(Diaper), 牛奶(Milk)
4	啤酒(Beer), 面包(Bread), 尿布(Diaper), 牛奶(Milk)
5	可乐(Coke), 尿布(Diaper), 牛奶(Milk)



关联
规则

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}



关联规则发现 (Association Rule Discovery)

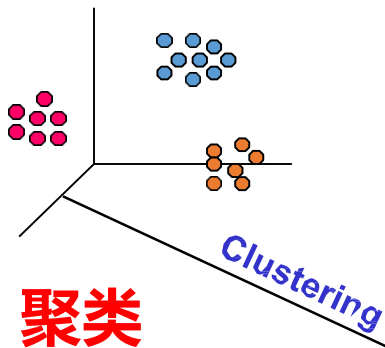
- 问题定义

- 给定一个记录的集合，每一条记录包括若干个项 (item)
- 从集合中找出由一个 item/item set 预测另一个 item/item set 同时出现的规则。

- 表达形式 $X \rightarrow Y$

- 满足X中条件的数据库元素，在一定程度上也满足Y中的条件。
- X被称为前项，Y被称为后项。

数据挖掘的任务类型



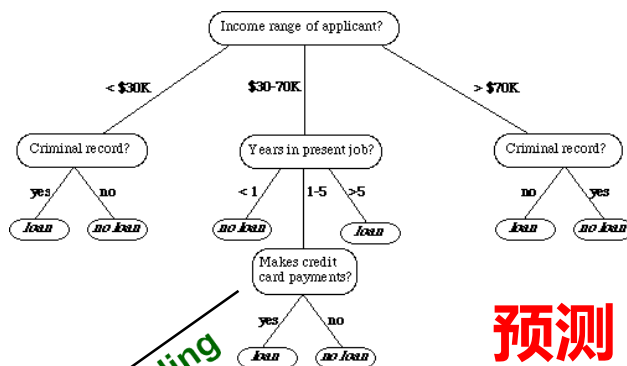
关联分析



Association Analysis

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

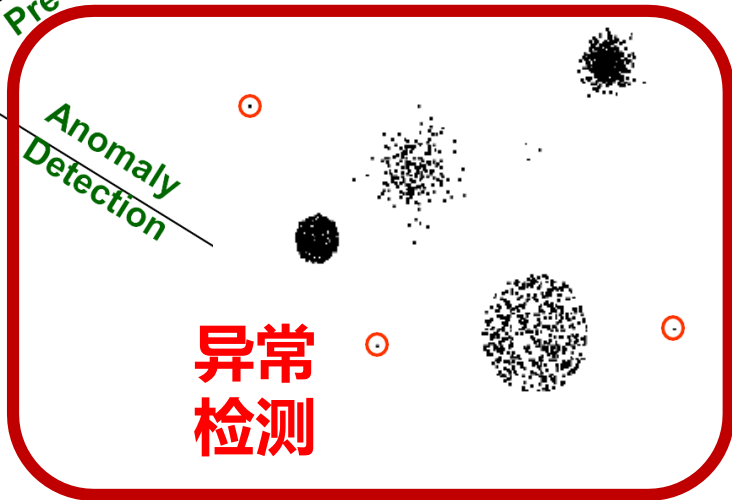


Predictive Modeling

预测建模

Anomaly Detection

异常检测

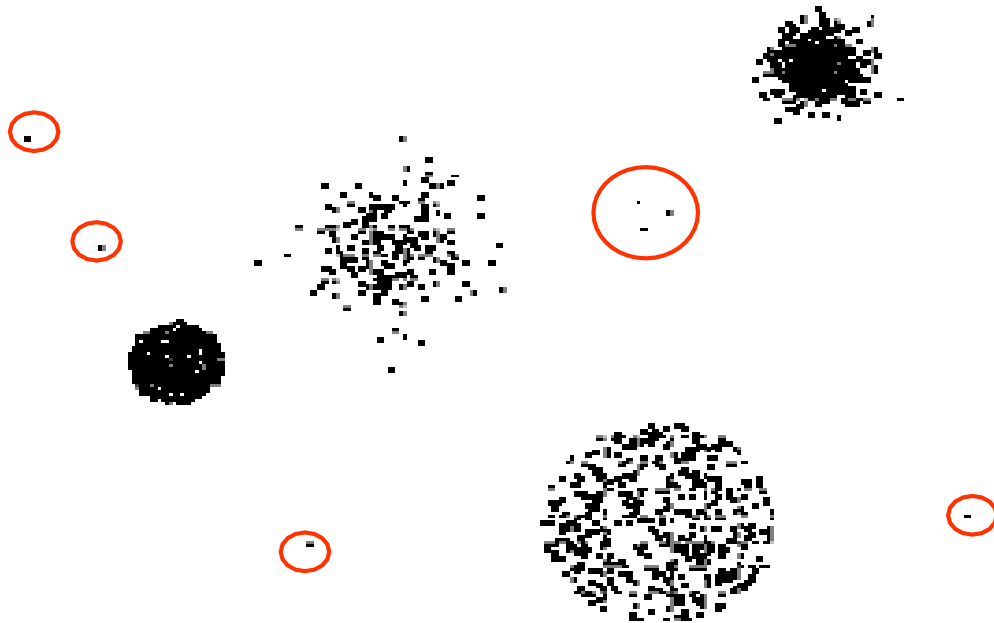


偏离/异常检测 (Deviation/Anomaly Detection)

- 从正常的行为中检测重要的偏离/异常
- 不符合正常实例分布的观测，比如与某种分布下大多数实例不相似
- 应用
 - 消费/电信欺诈检测
 - 网络入侵检测
 - 医疗和公共卫生



偏离/异常检测



异常检测的挑战

- 离群点的数量是未知的
- 分析过程可能是**无监督**的
 - 无监督模型的一个难点是对于分析结果非常难以验证（这一点和聚类问题相同）
- 分类过程如果是**有监督**的
 - “正常”样本的数量是远远多于“异常”样本
 - 异常检测往往可以等价于一种**非对称**的分类问题（类别不平衡）
- 数据不平衡问题(海底捞针问题)：机器学习分类器从大量负类（不感兴趣的）中找到少数正类（感兴趣，或故障）
 - 1. 每年大约有2%的信用卡账户被欺骗。（大多数欺诈检测领域严重不平衡）
 - 2. 工厂生产故障率通常约0.1%。

机器学习和数据挖掘

相关开源工具包、主流期刊和会议

主流的开源工具包（版本号截止2022.09.05）

- **PyTorch (version 1.12.1)**
(Facebook/Meta)
- **Tensorflow (version 2.9.2) (集成 Keres)**
(Google)
- **CNTK (version 2.7)**
(Microsoft)
- **MXNet (version 1.9.1)**
(Amazon)
- **PaddlePaddle (version 2.3.2)-飞桨**
(Baidu)



▣ 顶级会议

(Deadline: <https://aideadlin.es/?sub=ML,CV,CG,NLP,RO,SP,DM,AP,KR>)

- **ICML(International Conference on Machine Learning)** ,
- **NIPS(NeurIPS, Neural Information Processing Systems)**,
- **ICLR(International Conference on Learning Representations)**,
- **COLT(Conference on Learning Theory)**
- 计算机视觉与机器学习混合: **CVPR(每年召开)**, **ICCV**, **ECCV(后2个两年一次, 错开召开)**
- 机器学习与数据挖掘大杂烩: **AAAI**, **IJCAI**
- 主要是数据挖掘领域: **KDD**, **SDM**, **ICDM**
- 其他机器学习领域比较知名的: **UAI**, **AISTATS**
- 其他数据挖掘领域比较知名的: **CIKM**, **PKDD**

机器学习和数据挖掘领域较好的会议和期刊

▣ 顶级期刊

● 机器学习相关领域:

Journal of Machine Learning Research (JMLR),

IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI),

International Journal of Computer Vision (IJCV)

▣ 大多数ML / DL论文也可作为预印本提供 (<https://arxiv.org>) (搜索和查询: <https://arxiv-sanity-lite.com/>)



Open access to 1,490,045 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics

Subject search and browse:

14 Jan 2019: [The annual update from the arXiv team is now available](#)

3 Jan 2019: [Holiday schedule announced for 21 January](#)

5 Sept 2018: [arXiv looks to the future with move to Cornell CIS](#)

See cumulative "What's New" pages. Read [robots beware](#) before attempting any automated download

(careful & critical though, note that not all preprints were later peer-reviewed / were accepted for publication)

课后阅读：

《机器学习》（周志华著）第1章

《数据挖掘导论》（第2版）第1-2章

下节课内容：

数据和特征工程