机器学习与数据挖掘

Machine Learning & Data Mining

权小军 教授

中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

1. Course Information

机器学习与数据挖掘

维基百科：

- **机器学习**是人工智能的一个分支。人工智能的研究历史有着一条从以“推理”为重点，到以“知识”为重点，再到以“学习”为重点的自然、清晰的脉络。显然，**机器学习是实现人工智能的一个途径**，即以机器学习为手段解决人工智能中的问题。

机器学习与数据挖掘

维基百科：

- **机器学习**是人工智能的一个分支。人工智能的研究历史有着一条从以“推理”为重点，到以“知识”为重点，再到以“学习”为重点的自然、清晰的脉络。显然，**机器学习是实现人工智能的一个途径**，即以机器学习为手段解决人工智能中的问题。

- 机器学习在近30多年已发展为一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、计算复杂性理论等多门学科。**机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法**。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与推断统计学联系尤为密切，也被称为**统计学习理论**。算法设计方面，机器学习理论关注可以实现的，行之有效的学习算法。

机器学习与数据挖掘

维基百科：

- **机器学习**是人工智能的一个分支。人工智能的研究历史有着一条从以“推理”为重点，到以“知识”为重点，再到以“学习”为重点的自然、清晰的脉络。显然，**机器学习是实现人工智能的一个途径**，即以机器学习为手段解决人工智能中的问题。
- 机器学习在近30多年已发展为一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、计算复杂性理论等多门学科。**机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法**。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与推断统计学联系尤为密切，也被称为**统计学习理论**。算法设计方面，机器学习理论关注可以实现的，行之有效的学习算法。
- 机器学习已广泛应用于**数据挖掘**、计算机视觉、自然语言处理、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、DNA序列测序、语音和手写识别等领域。

机器学习与数据挖掘

维基百科：

- **数据挖掘**（英语：data mining）是一个跨学科的计算机科学分支。**它是用人工智能、机器学习、统计学和数据库的交叉方法在相对较大型的数据集中发现模式的计算过程**。 数据挖掘过程的总体目标是从一个数据集中提取信息，并将其转换成可理解的结构，以进一步使用。

机器学习与数据挖掘

维基百科：

- **数据挖掘**（英语：data mining）是一个跨学科的计算机科学分支。**它是用人工智能、机器学习、统计学和数据库的交叉方法在相对较大型的数据集中发现模式的计算过程。 数据挖掘过程的总体目标是从一个数据集中提取信息**，并将其转换成可理解的结构，以进一步使用。
- 除了原始分析步骤，它还涉及到**数据库和数据管理**方面、数据预处理、模型与推断方面考量、兴趣度度量、复杂度的考虑，以及发现结构、可视化及在线更新等后处理。

机器学习与数据挖掘

维基百科：

- **数据挖掘**（英语：data mining）是一个跨学科的计算机科学分支。**它是用人工智能、机器学习、统计学和数据库的交叉方法在相对较大型的数据集中发现模式的计算过程。 数据挖掘过程的总体目标是从一个数据集中提取信息**，并将其转换成可理解的结构，以进一步使用。
- 除了原始分析步骤，它还涉及到**数据库和数据管理**方面、数据预处理、模型与推断方面考量、兴趣度度量、复杂度的考虑，以及发现结构、可视化及在线更新等后处理。
- 数据挖掘是“**数据库知识发现**”（Knowledge-Discovery in Databases, KDD）的分析步骤，**本质上属于机器学习的范畴**。

机器学习和数据挖掘的关系

- 数据挖掘试图从海量数据中找出有用的知识。

机器学习和数据挖掘的关系

- 数据挖掘试图从海量数据中找出有用的知识。
- 机器学习是数据挖掘的重要工具。

机器学习和数据挖掘的关系

- 数据挖掘试图从海量数据中找出有用的知识。
- 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。

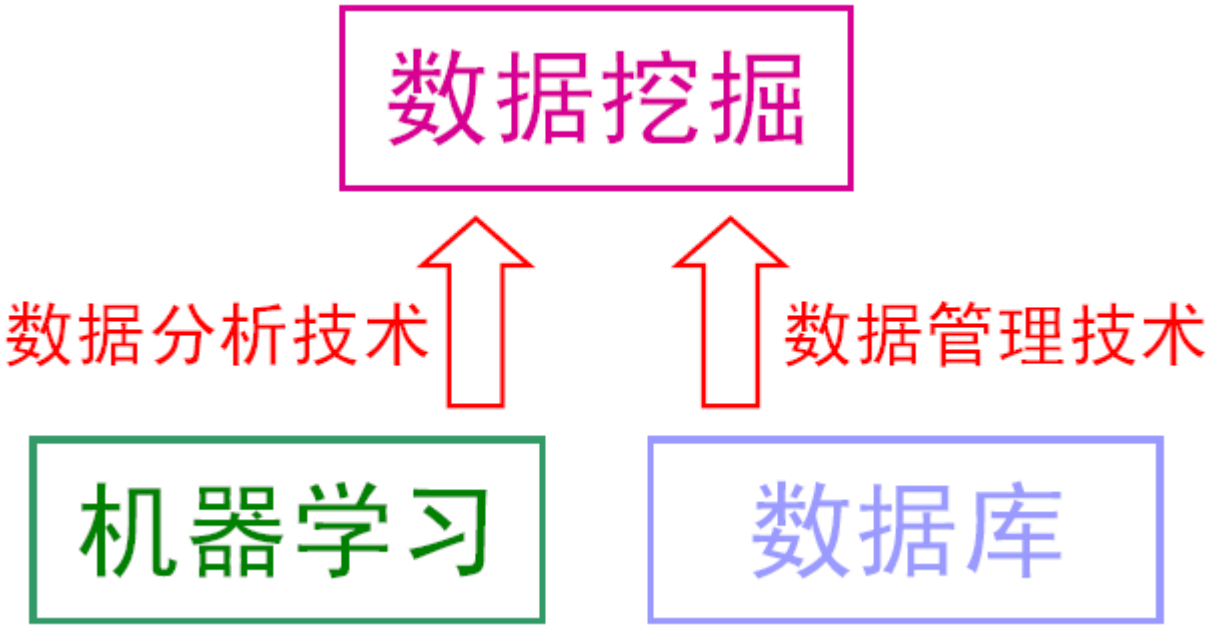
机器学习和数据挖掘的关系

- 数据挖掘试图从海量数据中找出有用的知识。
- 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。

机器学习和数据挖掘的关系

- 数据挖掘试图从海量数据中找出有用的知识。
- 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。
- 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**。

机器学习和数据挖掘的关系



机器学习 vs 人工智能?

Difference between machine learning and AI:

a) If it is written in Python, it's probably machine learning.

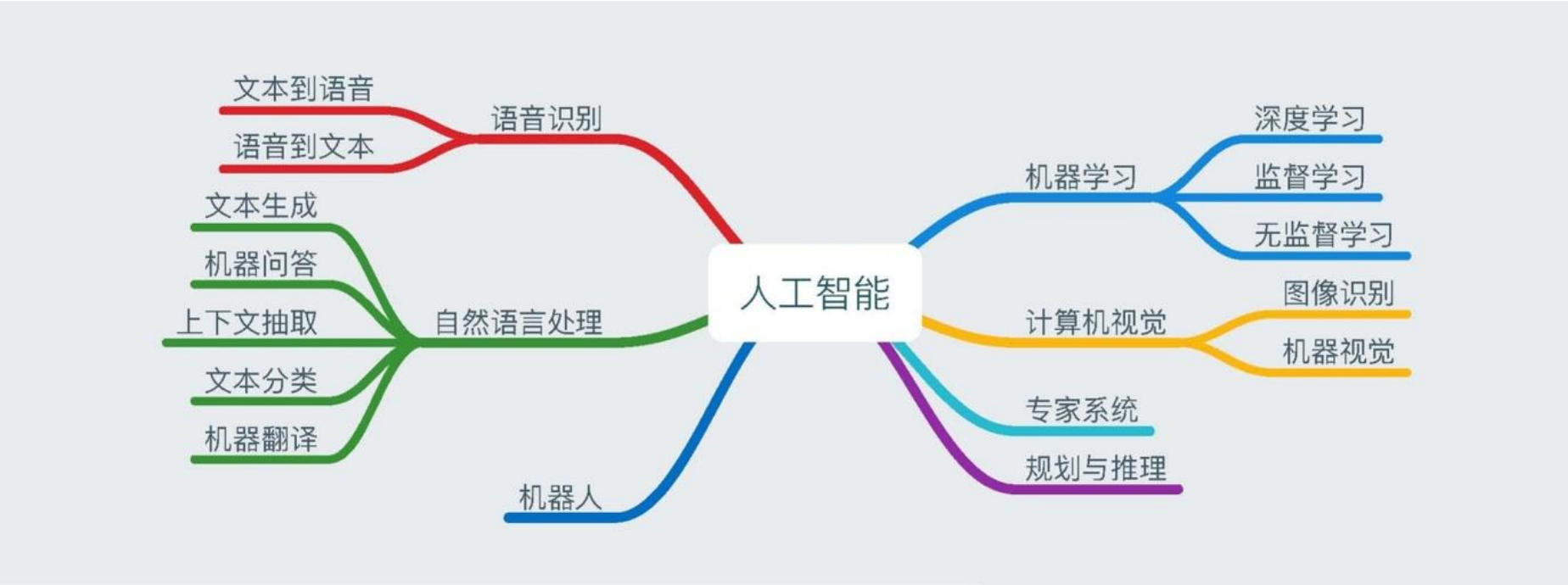
b) If it is written in PowerPoint, it‘s probably AI.

—互联网

机器学习 vs 人工智能?

—互联网

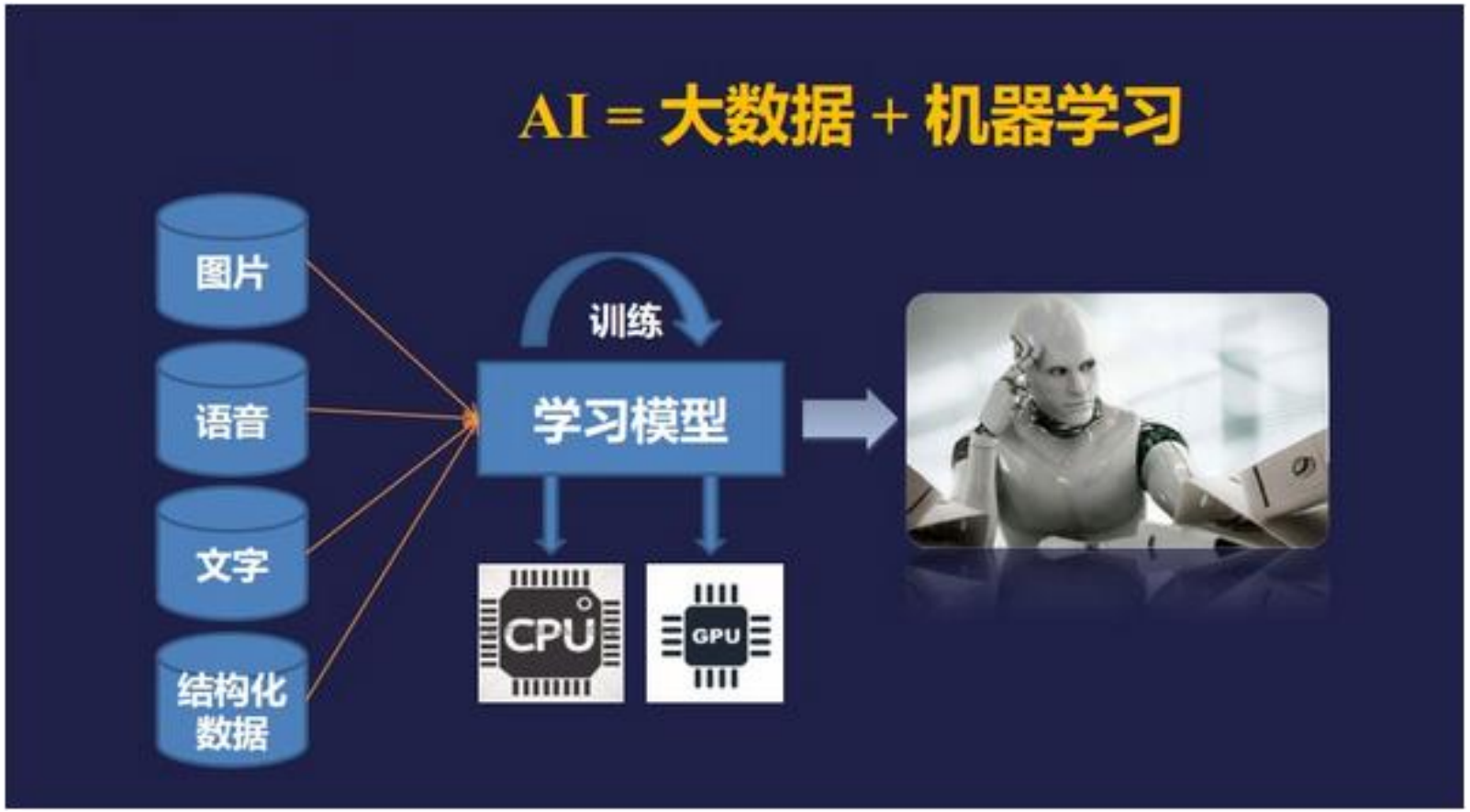
机器学习 vs 人工智能?



—互联网

机器学习 vs 人工智能?

机器学习 vs 人工智能?



—互联网

课程信息

- 任课老师：权小军
  - 中山大学数据科学与计算机学院教授
  - 研究方向：文本数据挖掘，自然语言处理，机器学习
  - Email: quanxj3@mail.sysu.edu.cn
- 助教：沈维州，李云昊
  - 研究方向：自然语言处理，机器学习
  - 办公地点：超算中心502

课程信息

- Nearest Neighbours
- Decision Trees
- Ensemble Models (Random Forest, Adaboost)
- Linear Regression/Classification
- Support Vector Machines (SVMs)
- Deep Learning (basics, optimization)
- Deep Learning (advanced deep models)
- Principal Components Analysis
- Probabilistic Models
- K-Means
- Expectation-Maximization
- Matrix Factorizations
- Bayesian Linear Regression
- Gaussian Processes
- Reinforcement Learning

课程信息

考核方式

- 闭卷考试：50%；
- 课程设计：30%（1个）- 代码，报告；
- 课堂作业：10%；
- 课堂考勤：10%
  - 1st absence：exemption（不扣分）；
  - 2nd absence: half exemption（扣2分）；
  - ≥3：每次扣5分（上限30分）；

网课期间的出勤情况仅作为参考

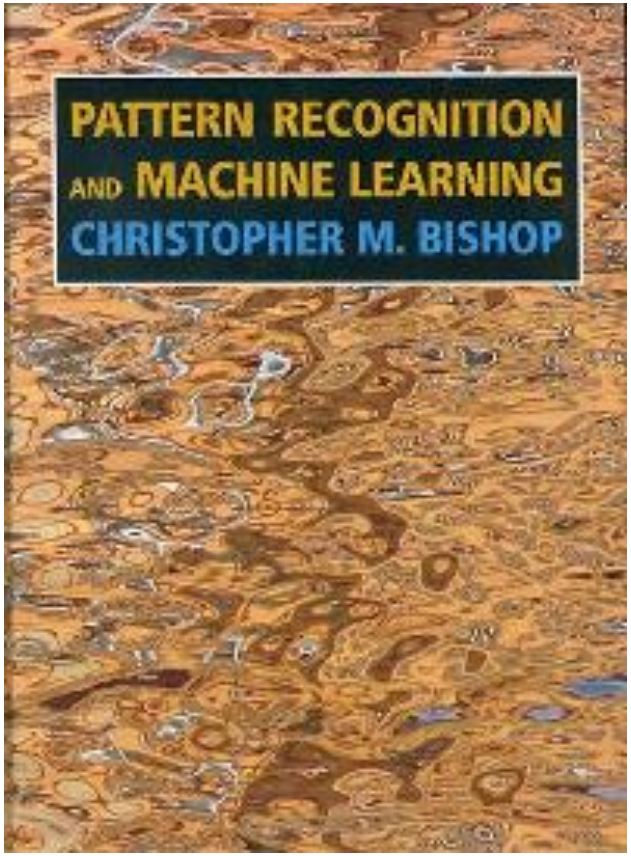
课程信息

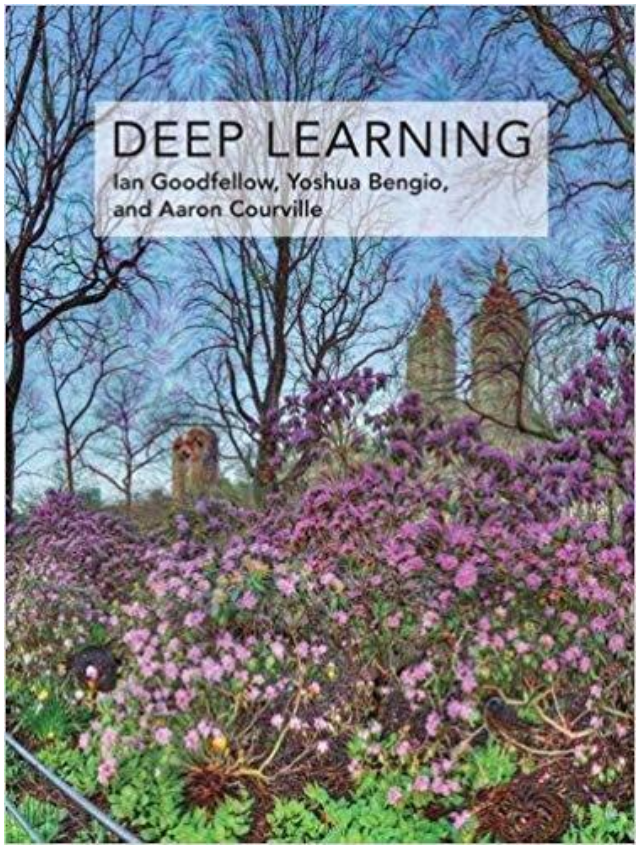
- 课件地址
  - 课件地址：https://pan.baidu.com/s/16xSQQrR70EClZnnl8Efaew
  - 提取码：j43l

课程信息

入门读物

课程信息


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville

进阶读物

2. Qualifications

Qualifications

工作经历:

- 2017.07-至今: 中山大学数据科学与计算机学院，教授
- 2014.01-2017.05: 新加坡科技研究局资讯通信研究院，研究科学家
- 2012.09-2013.12: 香港城市大学中文、翻译与语言学系，博士后研究员
- 2011.07-2011.12: 普渡大学计算机系，访问学者
- 2010.06-2010.08: 罗格斯大学商学院，访问学者
- 2008.07-2009.09: 香港城市大学计算机系，研究助理

教育背景:

- 2009.10-2012.08: 香港城市大学计算机系，博士
- 2005.09-2008.06: 中国科学技术大学计算机系，硕士

Qualifications

- 2005.09-2008.06：中国科学技术大学计算机系，硕士研究生
  - 语义计算与数据挖掘实验室
  - 研究方向：文本分类
  - 导师：陈恩红教授

Qualifications

- **2008.07-2012.08：香港城市大学计算机系，研究助理&博士研究生**
  - 问答系统小组
  - 研究方向：短文本处理&情感分析
  - 导师：刘文印教授

Qualifications

- 2010.06-2010.08：罗格斯大学商学院，访问学者
  - 研究方向：情感分析
  - 导师：熊辉教授

Qualifications

- **2011.07-2011.12：普渡大学计算机系，访问学者**
  - 智能信息检索实验室
  - 研究方向：新闻情感分析
  - 导师：司罗教授

Qualifications

- **2012.09-2013.12：香港城市大学中文、翻译与语言学系，博士后研究员**
  - 自然语言处理实验室
  - 研究方向：机器翻译
  - 导师：揭春雨教授

Qualifications

- **2014.01-2017.05：新加坡科技研究局资讯通信研究院，研究科学家**
  - 文本挖掘实验室
  - 研究方向：文本挖掘
  - 领导或参与多项工业界合作项目

3. Machine Learning Basics

Content

- 引言
- 基本术语
- 发展历程
- 应用现状
- 阅读材料

机器学习定义

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

-Tom Mitchell

机器学习定义

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Tom Mitchell

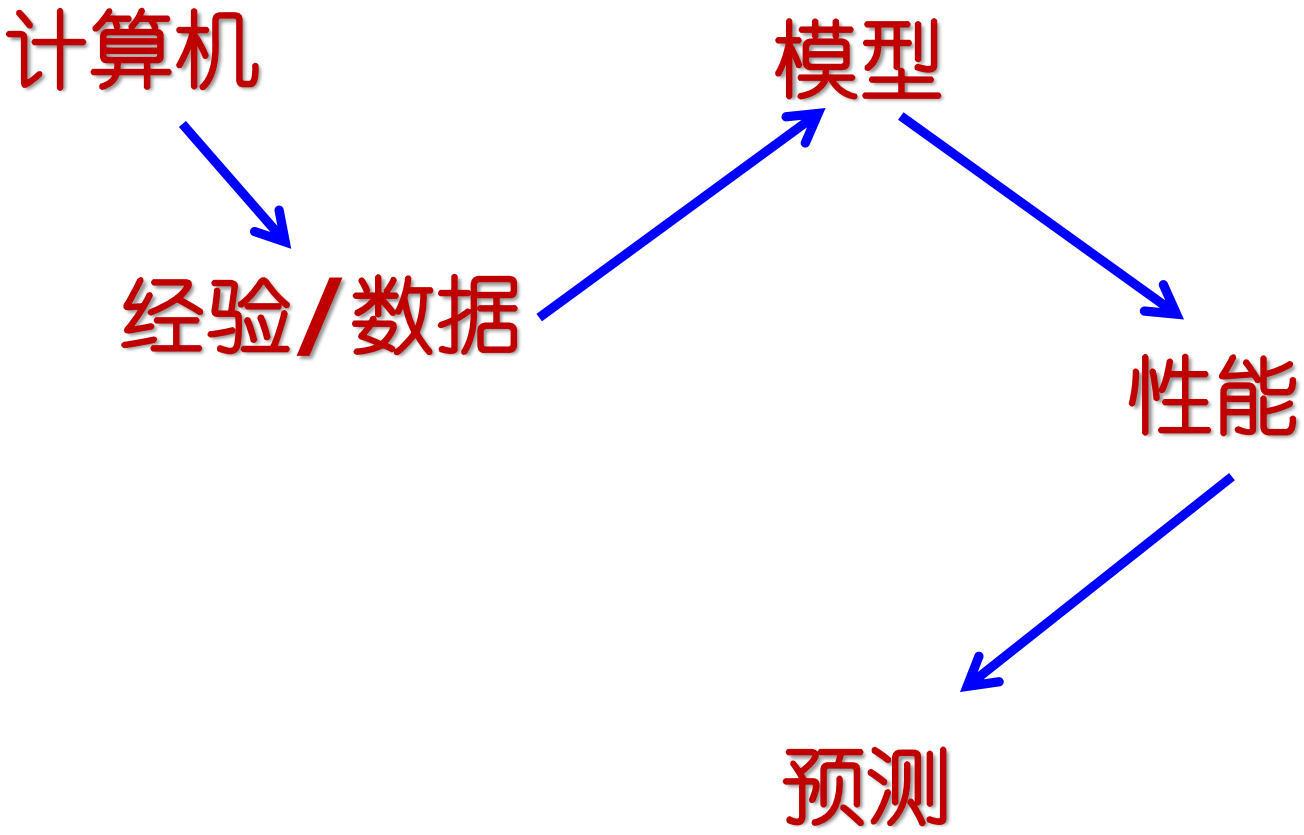
机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能，从数据中产生"模型"，用于新的情况的判断。

机器学习其它定义

- 机器学习是对能通过经验自动改进的计算机算法的研究

- 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

机器学习定义：关键词



机器学习和算命

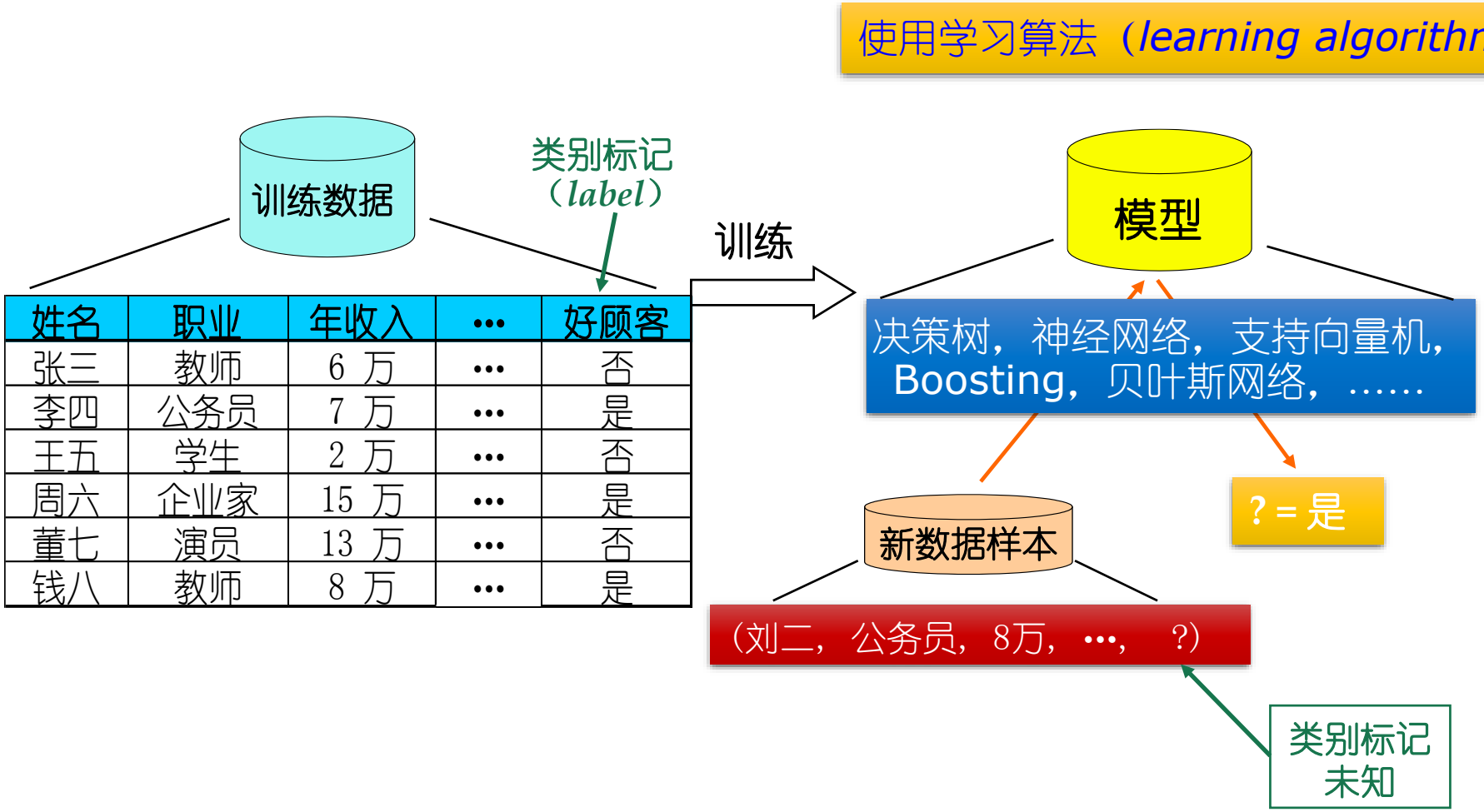
求签
算命

机器学习和算命

算命也是一个“机器学习”的过程！

- 特征：面相，语言，脸与手的纹路，出生八字等；
- 辅助技巧；

典型的机器学习过程


使用学习算法（learning algorithm）
训练数据
类别标记
（label）
训练
模型
姓名 职业 年收入 … 好顾客
张三 教师 6 万 … 否
李四 公务员 7 万 … 是
王五 学生 2 万 … 否
周六 企业家 15 万 … 是
董七 演员 13 万 … 否
钱八 教师 8 万 … 是
决策树，神经网络，支持向量机，
Boosting，贝叶斯网络，……
新数据样本
? = 是
（刘二，公务员，8万，…， ?）
类别标记
未知

典型的机器学习过程

- 过程：通常包括一个训练过程，一个测试过程；
- 数据：训练和测试数据分离；

典型的机器学习过程

- 过程：通常包括一个训练过程，一个测试过程；
- 数据：训练和测试数据分离；
- **特殊情况：训练和测试过程也可能融合；**

Content

- 引言
- 基本术语
- 发展历程
- 应用现状
- 阅读材料

基本术语：数据

	特征			标记
	编号	色泽	根蒂	敲声	好瓜
训练集 ←	1	青绿	蜷缩	浊响	是
	2	乌黑	蜷缩	沉闷	是
	3	青绿	硬挺	清脆	否
	4	乌黑	稍蜷	沉闷	否
测试集 ←	1	青绿	蜷缩	沉闷	?

基本术语：任务

- 有无标记信息
  - 监督学习：分类、回归
  - 无监督学习：聚类
  - 半监督学习：两者结合

基本术语：泛化能力

机器学习的目标是使得学到的模型能很好的适用于“新样本”，而不仅仅是训练集合，我们称模型适用于新样本的能力为泛化(generalization)能力。

通常假设样本空间中的样本服从一个未知分布 $\mathcal{D}$,样本从这个分布中独立获得，即“独立同分布”(i.i.d)。一般而言训练样本越多越有可能通过学习获得强泛化能力的模型

基本术语：监督学习

- Instance，feature vector，feature space
- 输入实例x的特征向量：

$$x=(x^{(1)},x^{(2)},\cdots,x^{(i)},\cdots,x^{(n)})^{\mathrm{T}}$$

- 表示多个输入变量中的第i个

$$x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^{\mathrm{T}}$$

- 训练集：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

- 输入变量和输出变量：
  - 分类问题、回归问题、标注问题

基本术语：监督学习

- 假设空间
  - 监督学习目的是学习一个由输入到输出的映射，称为模型
  - 模式的集合就是假设空间（hypothesis space）

基本术语

半监督学习

- 少量标注数据，大量未标注数据
- 利用未标注数据的信息，辅助标注数据，进行监督学习
- 较低成本

主动学习

- 机器主动给出实例，教师进行标注
- 利用标注数据学习预测模型

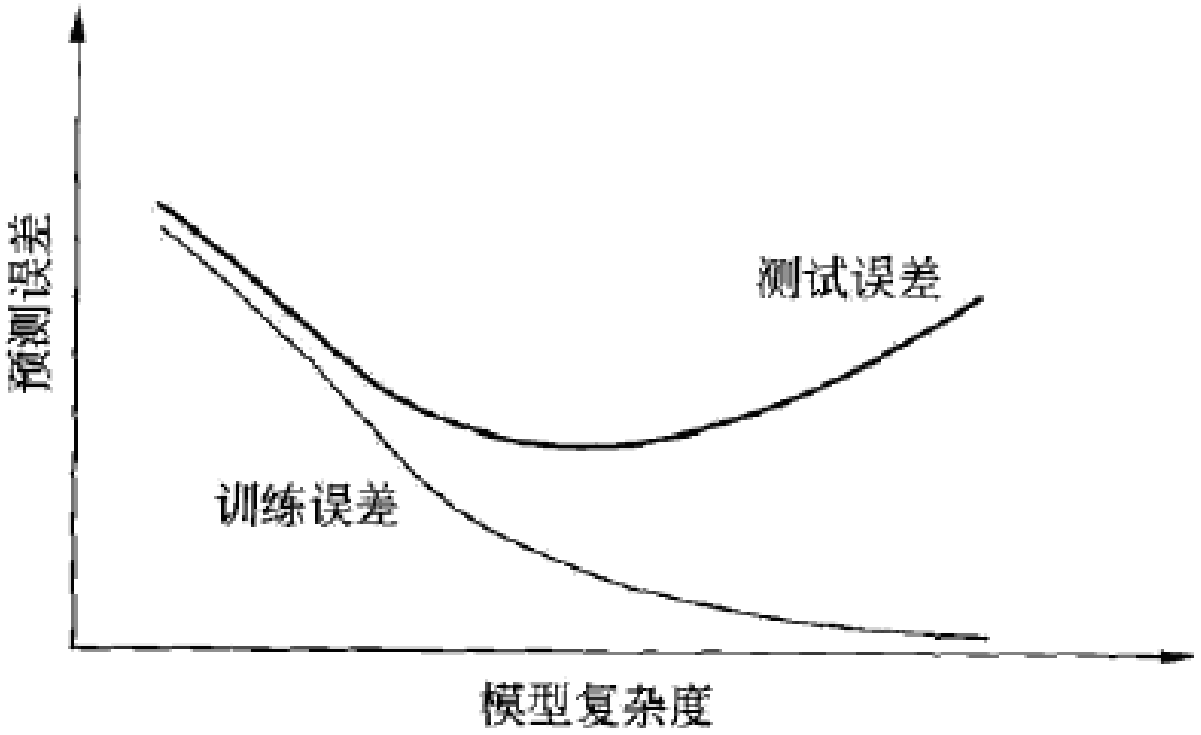
基本术语

模型评估与模型选择

- 训练误差，训练数据集的平均损失 $R_{\text{emp}}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$
- 测试误差，测试数据集的平均损失 $e_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} L(y_i, \hat{f}(x_i))$
- 损失函数是0-1 损失时: $e_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} I(y_i \neq \hat{f}(x_i))$
- 测试数据集的准确率: $r_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} I(y_i = \hat{f}(x_i))$

基本术语

模型评估与模型选择



Content

- 引言
- 基本术语
- 发展历程
- 应用现状
- 阅读材料

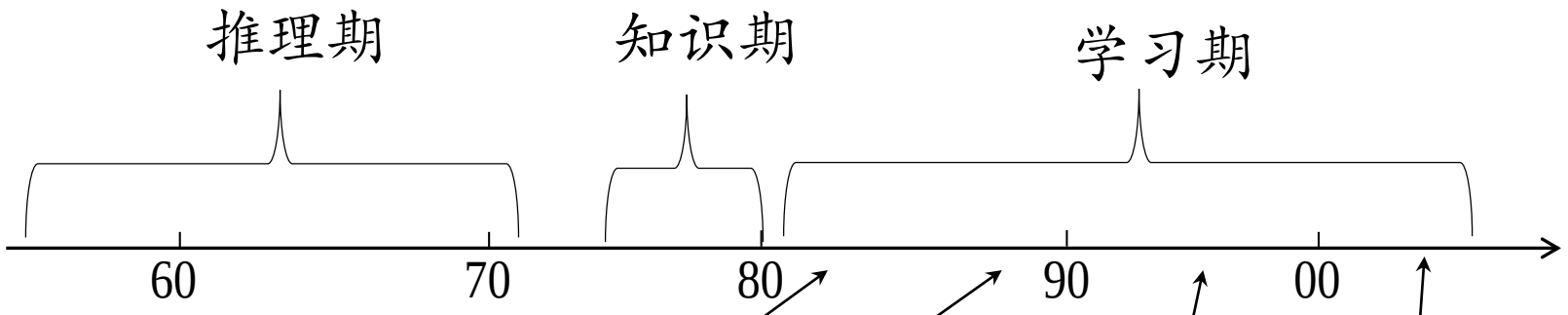
发展历程

- 推理期:
  - A. Newell和H. Simon的“逻辑理论家”(Logic Theorist)程序以及此后的“通用问题求解”(General Problem Solving)程序等在当时取得了令人振奋的结果。
  - 2006年卡耐基梅隆大学宣告成立第一个“机器学习系”，机器学习奠基人之一T.Mitchell教授任系主任。
- 知识期:
  - 大量专家系统问世，在很多应用领域取得大量成果;
  - 但是由人来总结知识再交给计算机相当困难。

发展历程

- 学习期：
  - 符号主义学习
    - ✓决策树：以信息论为基础，最小化信息熵，模拟了人类对概念进行判定的树形流程
    - ✓基于逻辑的学习：使用一节逻辑进行知识表示，通过修改扩充逻辑表达式对数据进行归纳
  - 连接主义学习
    - ✓神经网络
  - 统计学习
    - ✓支持向量机及核方法

发展历程



符号主义学习：决策树和基于逻辑的学习

连接主义学习：基于神经网络

统计学习：支持向量机和核方法

连接主义学习：深度学习

Content

- 引言
- 基本术语
- 发展历程
- 应用现状
- 阅读材料

应用现状

- 数据挖掘
- 计算机视觉
- 自然语言处理
- 生物特征识别
- 搜索引擎
- 医学诊断
- 检测信用卡欺诈
- 证券市场分析
- DNA序列测序
- 语音和手写识别
- 战略游戏
- 机器人

应用现状

Text to speech and speech recognition

应用现状

Computer vision

应用现状

Bioinformatics

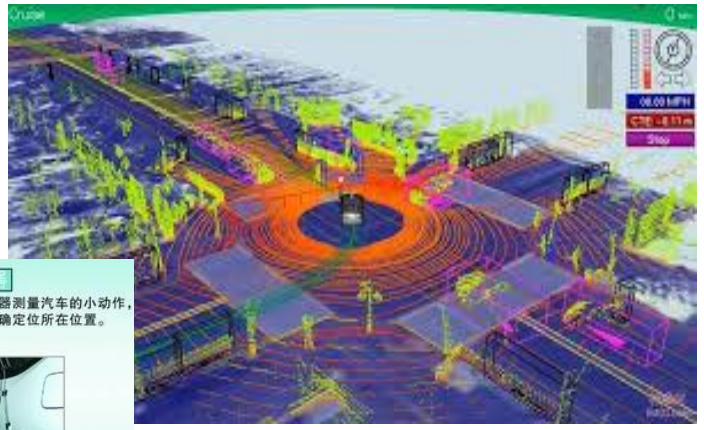
- Gene

应用现状

Financial Information

应用现状

Robotic Control



应用现状

Deep Learning

Google的猫脸识别：人工智能的新突破

Leon 发表于 2012/06/27-13:29 Google /人工智能 /猫脸识别 /神经网络 /tech

分享到QQ 分享 快成为第一个分享 的人吧! 分享到

Artificial Intelligence，也就是人工智能，就像长生不老和星际漫游一样，是人类最美好的梦想之一。虽然计算机技术已经取得了长足的进步，但是到目前为止，还没有一台电脑能产生“自我”的意识。是的，在人类和大量现成数据的帮助下，电脑可以表现的十分强大，但是离开了这两者，它甚至都不能分辨一个喵星人和一个汪星人。

可喜的是，我们还有 Google 这类“不靠谱”的公司。据 纽约时报报道，Google X 实验室近日开发出了一套具备自主学习能力的神经网络系统。这套系统有什么神奇之处呢？不借助任何外界信息帮助，它就能从一千万张图片中找出那些有小猫的图片。

Content

- 引言
- 基本术语
- 发展历程
- 应用现状
- 阅读材料

阅读材料

机器学习相关学术期刊和会议

- **机器学习**
  - 学术会议：NIPS、ICML
  - 学术期刊：Machine Learning和Journal of Machine Learning Research
- **数据挖掘**
  - 学术会议：SIGKDD、ICDM、SDM、PKDD和PAKDD
  - 学术期刊：IEEE Transactions on Knowledge and Data Engineering
- **人工智能**
  - 学术会议：IJCAI和AAAI
- **自然语言处理**
  - 学术会议：ACL、EMNLP、COLING

《机器学习与数据挖掘》大作业

多标签用户人格分类

作业背景

MBTI理论认为一个人的个性可以从四个角度进行分析，用字母代表如下：

- 驱动力的来源：外向E--内向I
- 接受信息的方式：感觉S--直觉N
- 决策的方式：思维T--情感F
- 对待不确定性的态度：判断J--知觉P

按照不同的组合，可以产生16种人格类型

作业背景

本次大作业要求利用机器学习方法，通过用户的发言记录对用户的人格类型进行分类

数据链接 https://www.kaggle.com/datasnaek/mbti-type

作业内容

1. 使用集成学习方法完成人格分类。

- 对数据进行预处理。
- 使用集成学习模型（AdaBoost或Random Forest）进行人格分类。
- 提交报告及代码。

作业内容

2. 使用SVM进行人格分类

- 数据预处理。
- 使用SVM进行人格分类。
- 提交报告及代码。

作业内容

3. 使用深度学习模型进行人格分类

- 数据预处理。
- 使用深度学习模型（不限）进行人格分类。
- 提交代码和报告。

评价指标

- 单独对每种类别进行评价
  - F1 & Accuracy
- 整体评价
  - F1 & Accuracy

分数分布

- 集成学习方法：8分
- SVM方法：8分
- 深度学习方法：14分

关键DDL

- 6月10日：提交集成方法报告和代码；
- 6月30日：提交SVM方法报告和代码；
- 7月31日：提交深度学习方法报告和代码；

注意事项

- 深度学习方法需要有方法上的创新，如果只简单使用开源代码或框架，则最多只能拿8分!

- 加分项：使用英文撰写报告，设计合理实验（对比不同模型表现/不同超参数对性能影响），自己撰写模型代码，尽可能少的调用工具。

Thank you!

权小军 中山大学数据科学与计算机学院