《机器学习》课件

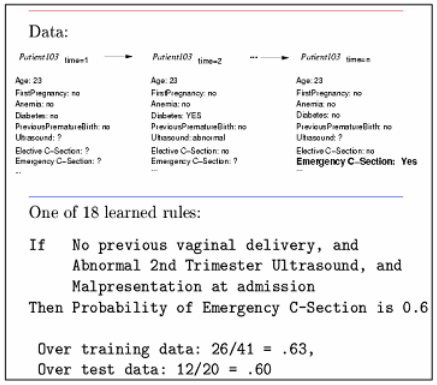
机器学习绪论

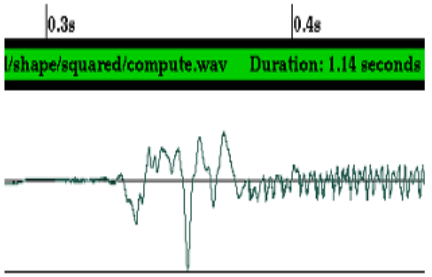
一些现象的总结

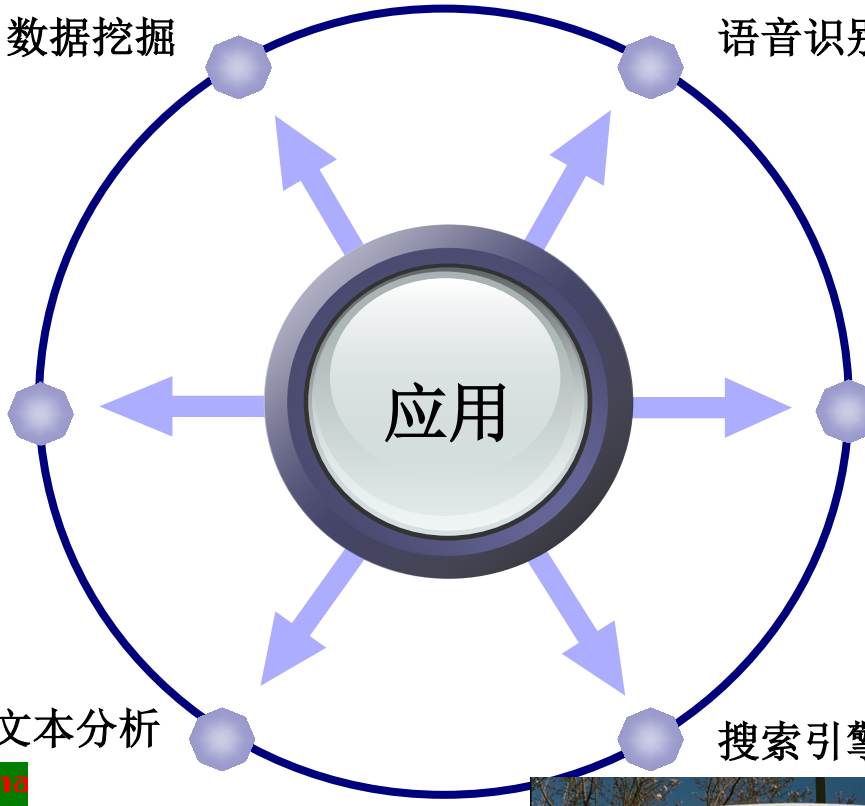
- 传统机器学习算法
  - 模型选择，NN，SVM，Adaboost
  - 学习策略与优化 SGD，解析
- 深度学习
  - 模型选择参数化 结构手工调整以及优化
  - 神经网络结构搜索 模型选择自动化

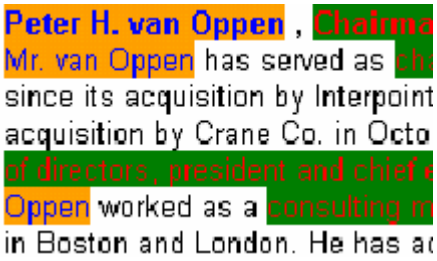
自动搜出网络

机器学习—应用


Data:
Patient103 time=1
Patient103 time=2
Patient103 time=n
Age: 23
FirstPregnancy: no
Anemia: no
Diabetes: no
PreviousPrematureBirth: no
Ultrasound: ?
Elective C-Section: ?
Emergency C-Section: ?
Age: 23
FirstPregnancy: no
Anemia: no
Diabetes: YES
PreviousPrematureBirth: no
Ultrasound: abnormal
Elective C-Section: no
Emergency C-Section: ?
Age: 23
FirstPregnancy: no
Anemia: no
Diabetes: no
PreviousPrematureBirth: no
Ultrasound: ?
Elective C-Section: no
Emergency C-Section: Yes
One of 18 learned rules:
If No previous vaginal delivery, and
Abnormal 2nd Trimester Ultrasound, and
Malpresentation at admission
Then Probability of Emergency C-Section is 0.6
Over training data: 26/41 = .63,
Over test data: 12/20 = .60


0.3s
0.4s
/shape/squared/compute.wav Duration: 1.14 seconds


数据挖掘
语音识别
应用
目标识别
控制学习
文本分析
搜索引擎


Peter H. van Oppen , Chairma
Mr. van Oppen has served as ch
since its acquisition by Interpoint
acquisition by Crane Co. in Octo
of directors, president and chief
Oppen worked as a consulting m
in Boston and London. He has a

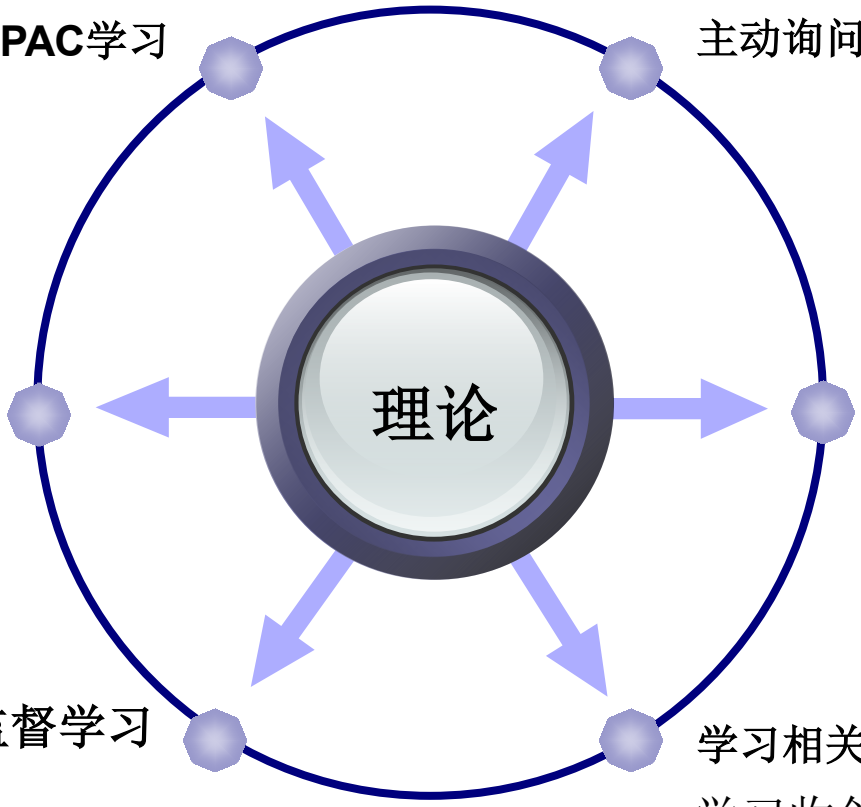

Google
2400 Bayshore Parkway

机器学习—理论

有监督概念学习

实例—成败概率（δ）

假设—误差（ε）



机器学习的发展及趋势

- 80年代

  1983年，《机器学习：通往人工智能的途径》 R.S.Michalski 等著

1986年，Machine Learning杂志创刊

主要方法：增强学习 AlphaGO

符号主义，代表：ILP　　归纳逻辑程序设计

连接主义， 代表：NN　　神经网络

- 90年代

  统计学习，代表：SVM　　支持向量机

统计学习的发展及趋势

引导这一革命的60年代四项发现：

- 解决不适定问题的正则化原则

  Tikhonov，Ivanov，Phillips
- 非参数统计学

  Parzen，Rosenblatt
- 泛函空间的大数定律，及其与学习过程的关系

  Vapnik，Chervonenkis
- 算法复杂性与归纳推理的关系

  Kolmogorov，Solomonoff，Chaitin

  学习理论：小样本统计学习理论

机器学习的发展及趋势

- 目前及趋势

  集成学习（ensemble learning）

  普适机器学习:基础方法、广泛应用
- 机器学习面临的挑战
  - **泛化能力：学习推广能力，SVM，理论->实践；EL ，实践->理论**
  - **学习速度：** “训练速度” 和 “测试速度”，训练速度快则测试速度慢：k近邻；测试速度快则训练速度慢：神经网络
  - **可理解性（理论层面）：学到了什么？“黑盒子”：NN、SVM、EL，“白盒子”问题**

    deep learning

机器学习的发展

- 未标记数据处理：充分利用各种数据，如何利用

问题：经典机器学习方法：标记数据->事件结果

未标记数据：遥感数据、web、海军舰队

“坏”数据->噪声污染、属性缺失、不一致，处理：扔掉

- 代价敏感：减低错误率，漏警与虚警

问题：把“正确”当成“错误”

把“错误”当成“正确” 一样吗？

达到较低错误基础上，如何趋利避害？

- 高维数据处理：成千上万个属性
- 结构数据学习：挖掘“结构”隐含的信息
- 领域知识利用：特定领域的“最优学习器”

机器学习国外研究现状

- DARPA启动PAL计划

  Perceptive Assistant Learns

  感知辅助学习

  启动：2003年

  Defence Advanced Research Projects Agency

  时间：5年

  美国国防部高等研究计划局

  首期投入：2千9百万美元/1-1.5年

  核心：机器学习

  涵盖：知识表达、知识推理、自然语言处理

  目标：获得新的有价值的技术，可用于军事、商业、科学研究；开发新软件，可帮助决策者处理并发多任务及意外事件等复杂问题

机器学习国外研究现状

- R Reflective Agents with Distributed Adaptive Reasoning
  分布式自适应推理智能单元

  承担单位：CMU（卡内基梅陇大

  Boeing, CMU, Dejima Inc., Fetch Tech
  MIT, Oregon HSU, Stanford, SUNY-S
  Berkeley, UMass, UMich, UPenn, R
  UT Austin, UW, Yale

  'soldier's assistant'

- CALO子计划

  承担单位：SRI（斯坦福研究院）等20家单位

  首期：2千2百万美元

  目标：用户示教和通过工作自动学习的软件，可以处理广泛的相关决策任务，可以执行常规任务，辅助解决突发事件

机器学习国外研究现状

分类：将观察信息分为相关学习组

Category: Learn relevant groupings for observed information

关系：从实例中学习相互关系

Relational: Learn relationships among entities

时序：学习用户行动的动态结构

Sequential: Learn the dynamic structure of ongoing activity of the user

程序：通过规划学习处理新任务

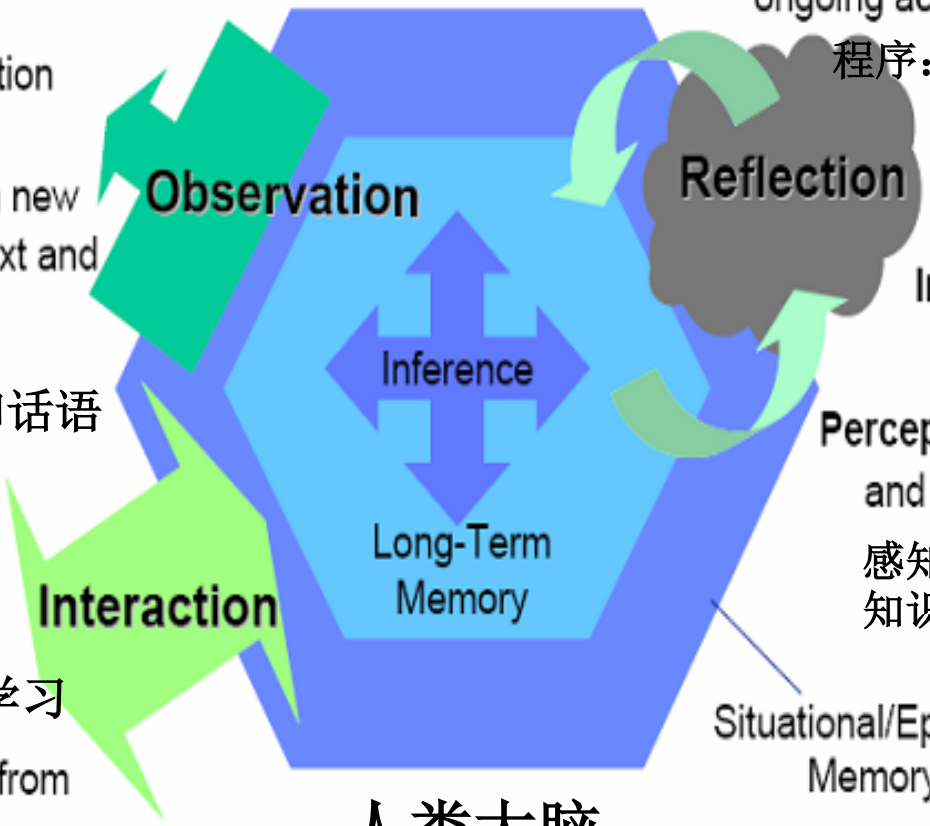
Procedural: Learn to handle new tasks through planning

推断：对学习的新事实推理

Inferential: Reason to learn new facts

Language: Learn new information from text and utterances

语言：从文本和话语学习新的信息



Perceptual: Learn to associate images and sounds with other knowledge

感知：用图像、声音等其他知识辅助学习

建议：向用户学习

Advice: Learn from the user

Situational/Episodic Memory

情景/片段记忆

人类大脑

机器学习主要研究方法

- 机器学习系统结构 机器学习与人类的共生，课本大概率是对的其实

学习算法定义：给定算法A、任务T、性能评估量P、经验E，当A对T进行操作时，如果A随E的增加使得P完善，则称A具有从E进行学习的能力。

举例：给定算法A

T=买房子	目标分类	机器人自动驾驶
P=正确率；	分类正确次数/总次数	无差错里程/总里程
E=学习样本	目标训练样本	人类驾驶录像及驾驶指令

房子的价钱是由包括面积、房间的个数、房屋的朝向等等因素去决定的。广义的线性函数

机器学习主要研究方法

- 选择训练经验E：
  - **为P提供直接或间接反馈**
  - **学习器控制训练顺序的能力**

    被动学习，施教者提供样本顺序

    半主动学习，自动选择样本，询问施教者

    主动学习，不依赖施教者，试验新样本，自动修改学习结果

机器学习主要研究方法

- 训练样本集S分布与T分布的匹配程度：泛化能力

- 选择目标函数
  - **知识表达：状态空间描述**
  - **目标函数V（x）：**

$$\hat{V}(\mathrm{x}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6$$

  - **选择函数逼近算法**

    训练样例：< x, V_{train} (x)>

    训练过程：估计训练值

    训练信息：结果-胜？负？

机器学习主要研究方法

调整权值：最小二乘（best fit），最小均方LMS（Least Mean Squares）

$$E \equiv \sum_{(b_i, V_{train(i)}(b) \in S)} (V_{train(i)}(x) - \hat{V}_i(x))^2$$

$< x_i, V_{train(i)}(x_i) >, i$ – 训练样本序号

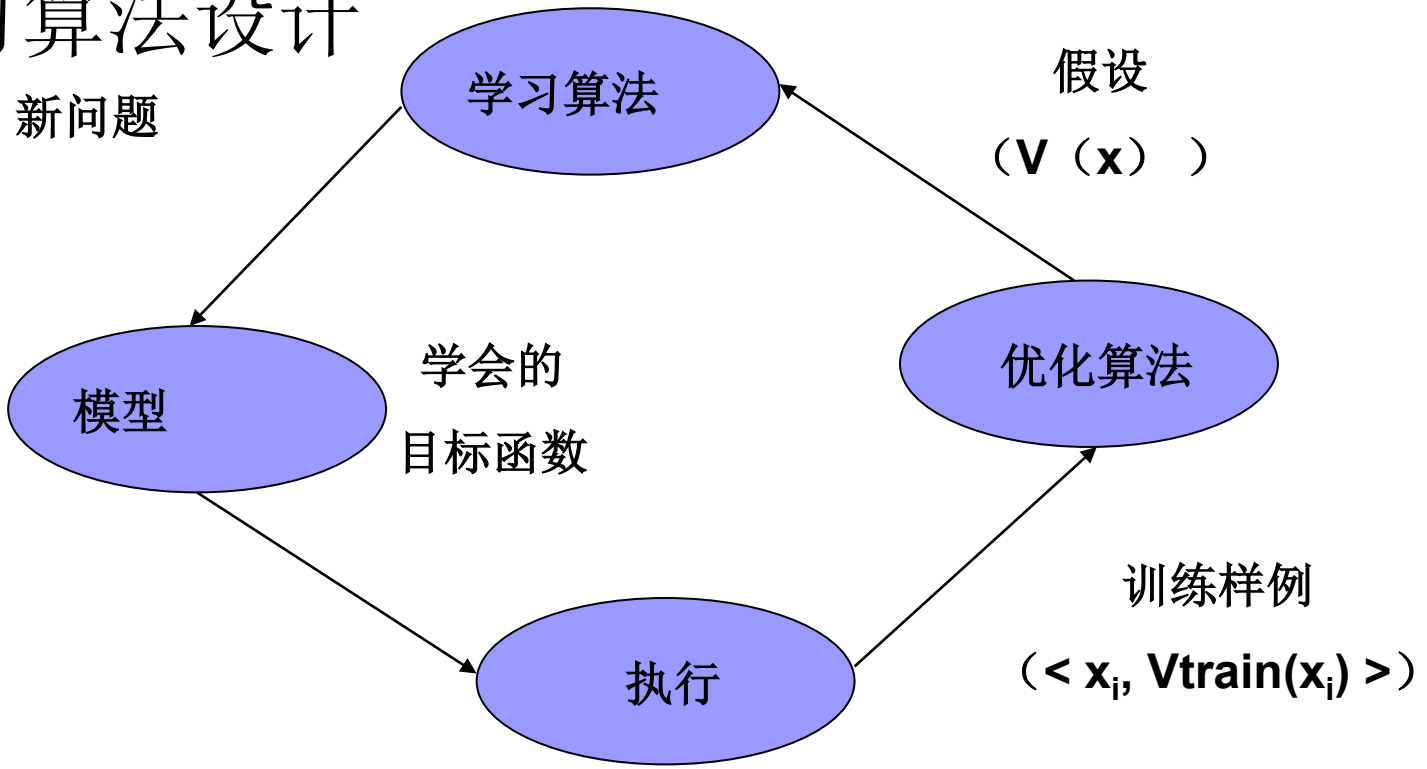
$$\hat{V}_i(x) = w_{0i} + \sum_{j=1}^{6} w_{ji} x_{ji}$$

$$w_{ji+1} = w_{ji} + \eta(V_{train(i)}(x) - \hat{V}_i(x))$$

$$j = 0, 1, 2, ..., 6$$

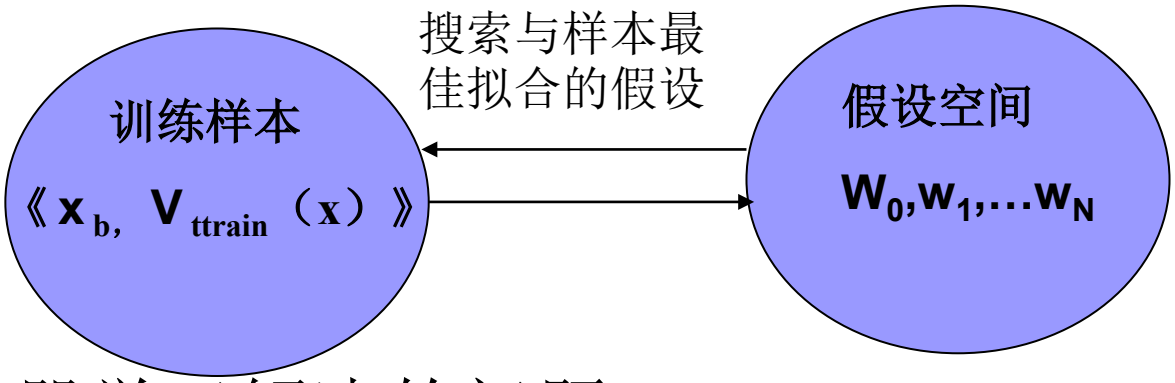
机器学习主要研究方法

- 学习算法设计



机器学习主要研究方法

- 机器学习---假设空间搜索



- 机器学习解决的问题：
  - 特定数据→学习→一般目标函数；数据充分->条件？->算法收敛到目标函数；现有算法的适用性和局限性？
  - 训练数据的充分性;
  - 先验知识引导泛化的规律

- 平时Project 30%
- 考试 70%
- 推荐教材 机器学习与视觉感知，张宝昌等，清华大学出版社 第二版