机 器 学 习 导 论

Foundations of Machine Learning

#L1

Introduction

袁彩霞

yuancx@bupt.edu.cn

计算机学院 智能科学与技术中心

Topics for Today

- What is Machine Learning (ML)?
- Why do we study ML?
- A rather simple sketch of ML history
- Different learning methods
- Key concepts
- Course Plan and Administrivia
- Recommended Textbooks

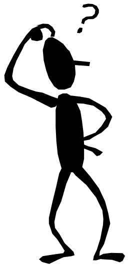
A Few Quotes

- "A breakthrough in machine learning would be worth ten Microsofts" (Bill Gates, Chairman, Microsoft)
- "Machine learning is the next Internet" (Tony Tether, Director, DARPA)
- Machine learning is the hot new thing" (John Hennessy, President, Stanford)
- "Web rankings today are mostly a matter of machine learning" (Prabhakar Raghavan, Dir. Research, Yahoo)
- "Machine learning is going to result in a real revolution" (Greg Papadopoulos, Former CTO, Sun)
- "Machine learning is today's discontinuity" (Jerry Yang, Founder, Yahoo)
- "Machine learning today is one of the hottest aspects of computer science" (Steve Ballmer, Former CEO, Microsoft)

So What Is Machine Learning?

- 例子1：徽章游戏
  - 1994年ICML及COLT国际会议的与会者会收到一个标记有“+”或“-”的徽章。游戏设计者(Haym Hirsh)根据一个与会者的姓名有关的函数，为不同的与会者发放不同的徽章

+ Naoki Abe
- Myriam Abramson
+ David W. Aha
+ Kamal M. Ali
- Eric Allender
+ Dana Angluin
……



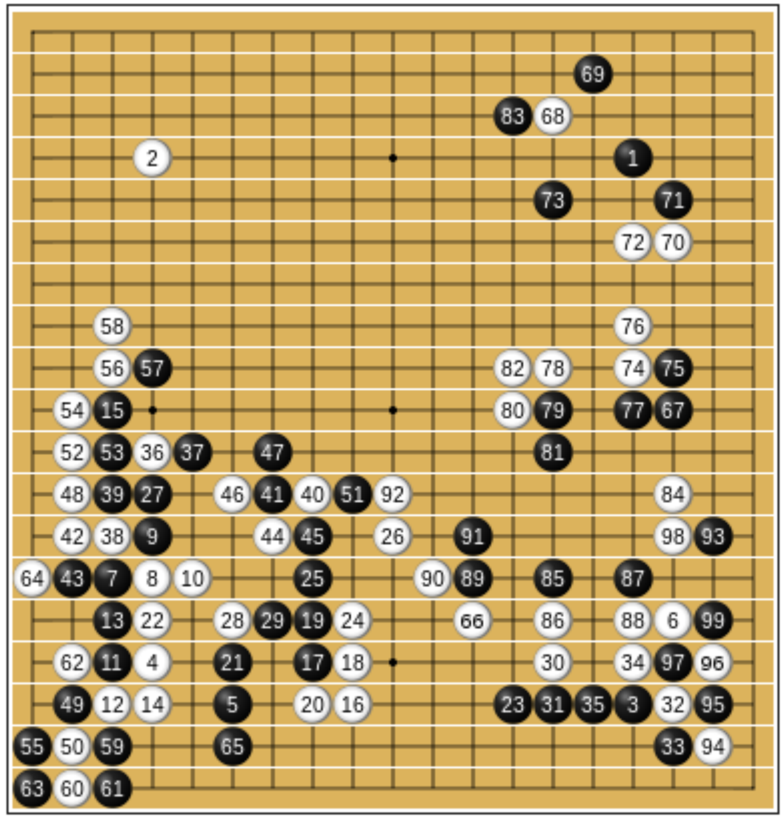
  - 目标：找出这个背后的函数，根据与会者姓名为其发放“+”或“-”徽章

So What Is Machine Learning?

- 例子2：NBA 统计预测
  - 假设已有NBA历史数据：
    - Player regular season stats
    - Player regular season career totals
    - Player playoff stats
    - Player playoff career totals
    - Player all-star game stats
    - Team regular season stats
    - Complete draft history
    - Currently all of the regular season
  - 目标：预测新赛季MVP得主及比赛结果

So What Is Machine Learning?

- 例子3：AlphaGo
  - AI computer program that plays the board game Go



So What Is Machine Learning?

- "Learning is any process by which a system improves performance from *experience*."

  – Herbert A. Simon

- “学习”无处不在：
  - 社会 (e.g., 科学团体)
  - 生物 (e.g., 人类)
  - 机器 (e.g., 计算机围棋)

So What Is Machine Learning?

- 定义：
  - Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
  - Tom Mitchell (1998). Well-posed Learning Problem: A computer program is said to *learn* from experience *E* with respect to some task *T* and some performance measure *P*, if its performance on *T*, as measured by *P*, improves with experience *E*.

So What Is Machine Learning?

任务：T；性能：P； 经验：E

T： 下围棋
P： 比赛中击败对手的百分比
E： 历史比赛数据或和自己进行对弈

T： 识别和分类图像中的手写文字
P： 分类的正确率
E： 已知类别的手写文字数据库

T： 通过视觉传感器在多车道高速公路上驾驶
P： 平均无差错行驶里程
E： 人类驾驶时录制的一系列图像和驾驶指令

相关学科

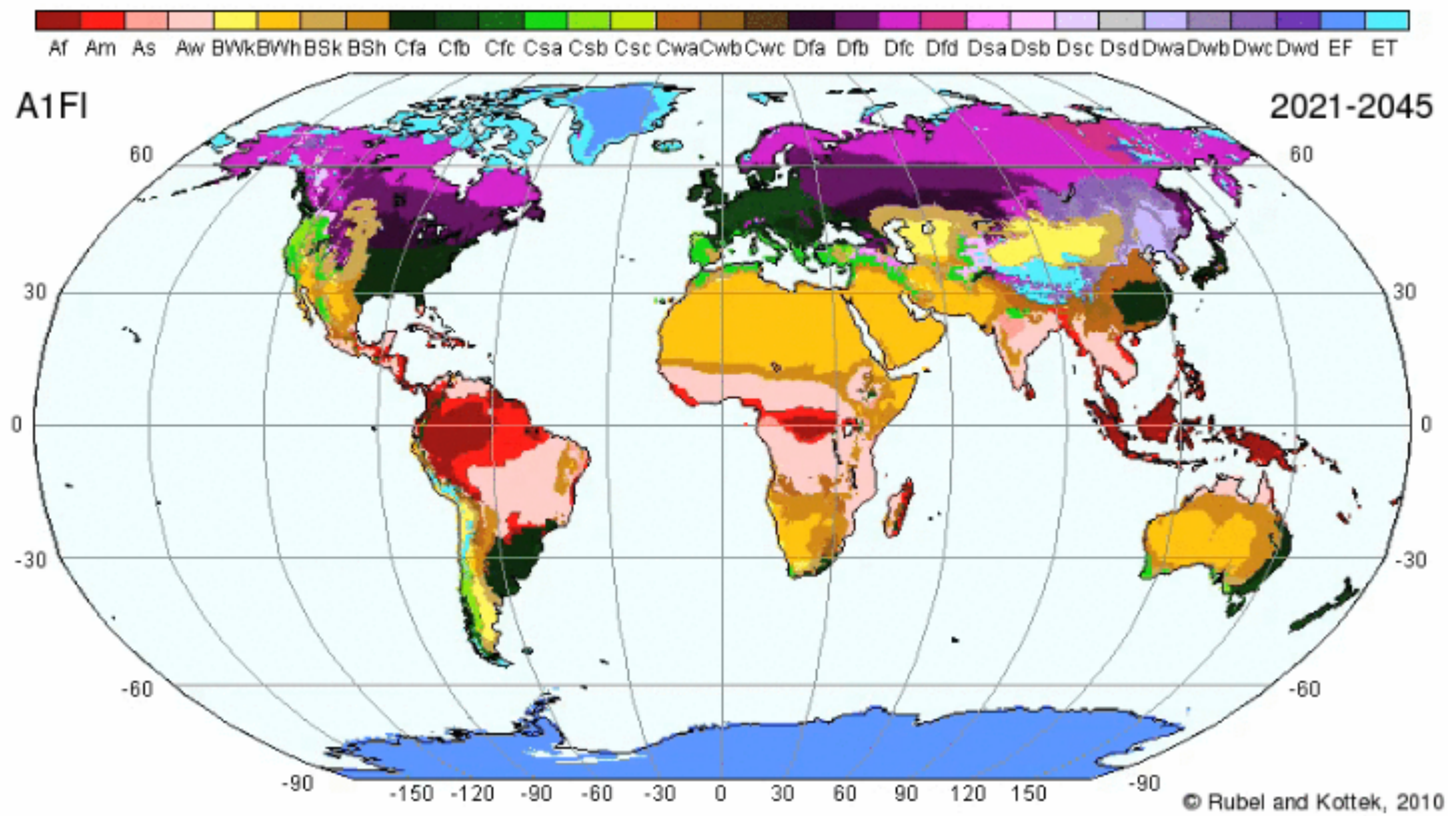
- Makes Use of :
  - 概率论、统计理论、矩阵论、最优化理论、计算理论
- Related to:
  - 哲学、心理学（认知、发展）、神经科学、语言学
- Has applications in:
  - 人工智能 (自然语言处理、计算机视觉、人机交互、Planning…)
  - 工程 (医学、金融、法律…)
  - 计算机科学 (汇编、体系结构、系统、数据库、硬件…)
  - ……

Topics for Today

- What is Machine Learning (ML)?
- Why do we study ML?
- A rather simple sketch of ML history
- Different learning methods
- Key concepts
- Course Plan and Administrivia
- Recommended Textbooks

Why Study Machine Learning?

- 处理海量、复杂的真实世界的数据
- 获得更高层次的认知
- 理解人类及其它生物的学习机制



World map of KG climate classification

Why Study Machine Learning?

- 广泛及迫切的社会需求：
  - 类别/行为/事件分析与预测
    - 医学诊断
    - 金融诈骗检测
    - Email的垃圾邮件过滤
    - 图书、电影、音乐、笑话等推荐
    - 天文图像
  - 任务求解/规划/控制
    - 下西洋棋、象棋
    - 电梯控制
    - 机器人动作控制
    - …

Why Study Machine Learning?

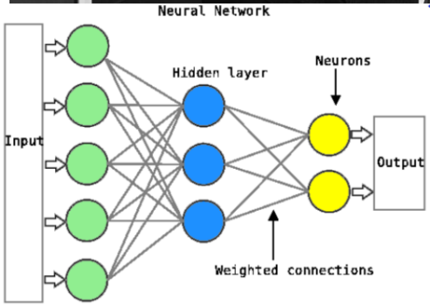
- 计算机系统需要新的能力
  - 很多任务仅靠“编程”难以实现
    - 例如：很难通过写一个程序解决人脸识别问题
  - 很多系统过于复杂，难以人工实现
    - 例如：商品推荐、图像搜索
- 时机到了！
  - 足够完备的算法和理论
  - 足够庞大的数据
  - 足够强大的计算能力

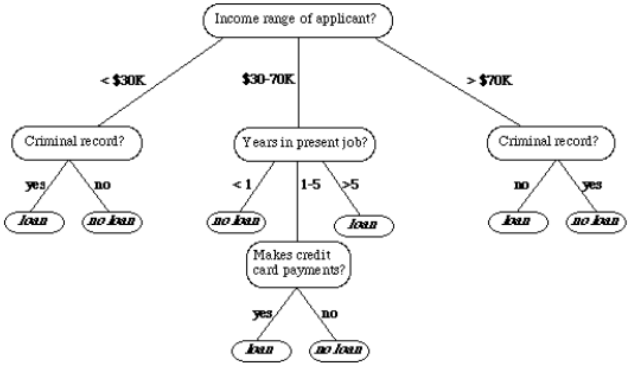
Topics for Today

- What is Machine Learning (ML)?
- Why do we study ML?
- A rather simple sketch of ML history
- Different learning methods
- Key concepts
- Course Plan and Administrivia
- Recommended Textbooks

机器学习简史

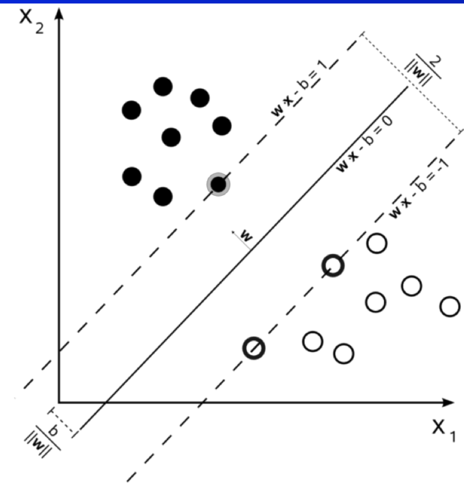
- 1950s
  - Arthur Samuel's checker player
  - Oliver Selfridge's Pandemonium
- 1960s:
  - Donald Hebb: Hebbian学习论
  - F. Rosenblatt: **感知机**及基于NN的连接主义
  - Widrow-Hoff rule 或delta法则
  - Minsky and Papert: XOR问题
- 1970s:
  - 符号概念推理
  - **专家系统**及知识获取瓶颈
  - **Quinlan's ID3**
  - Michalski's AQ and soybean diagnosis
  - Scientific discovery with BACON
  - Mathematical discovery with AM

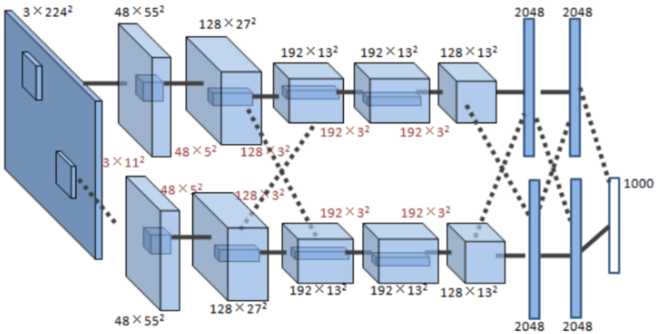




机器学习简史

- 1980s:
  - **决策树**及规则学习
  - Explanation-based Learning (EBL)
  - Learning and planning and problem solving
  - Cognitive architectures
  - **NN的复兴** (MLP, BP)
  - **PAC学习理论及支撑向量机**
  - 关注实验性方法论
- 1990s:
  - Data mining
  - Adaptive software agents and web applications
  - **强化学习(Reinforcement learning)**
  - Inductive Logic Programming (ILP)
  - **融合方法：Bagging, Boosting, Stacking**
  - **Bayes Net学习**
- Since 2000s:
  - **NN的又一次复兴** (CNN, DBN, etc.)
  - **Learning with human-in-the-loop**





Topics for Today

- What is Machine Learning (ML)?
- Why do we study ML?
- A rather simple sketch of ML history
- Different learning methods
- Key concepts
- Course Plan and Administrivia
- Recommended Textbooks

一般的学习过程

- 学习过程：
  - 学习器（计算机程序）学习经验数据D，以实现
  - **either** 对未知数据的预测 (predict)
  - **or** 对已知数据的描述 (discover)
- 例如：
  - 学习器通过学习一系列患者案例及其对应的医疗诊断，可以：
  - **either** 判断某个病人是否患某种疾病 (predict)
  - **or** 发现疾病与症状之间的关系(discover)

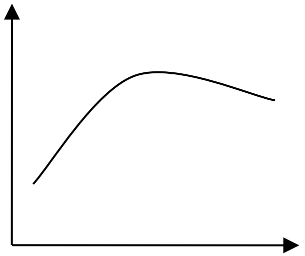
几个不同的学习方法

- 有监督/有指导学习：**Supervised Learning,** learning with a teacher
- 数据：$D=\{d^{(1)}, d^{(2)}, \ldots, d^{(n)}\}$, n个训练样本
  - $d^{(i)}=(x^{(i)}, y^{(i)})$
  - $x^{(i)}$输入向量， $y^{(i)}$为$x^{(i)}$对应的类别/输出(given by a teacher)
- 目标：学习一个映射$f: x \to y$
- 问题类型：
  - 回归：X离散或连续，Y*连续*
  - 分类：X离散或连续，Y*离散*

有监督学习：例子

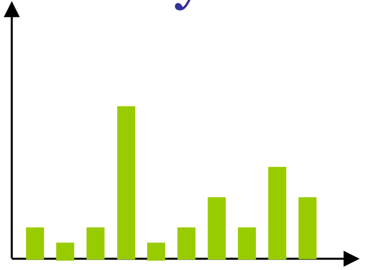
- 回归：Y连续

债务
盈利
未来订单



- 分类：Y离散

手写数字

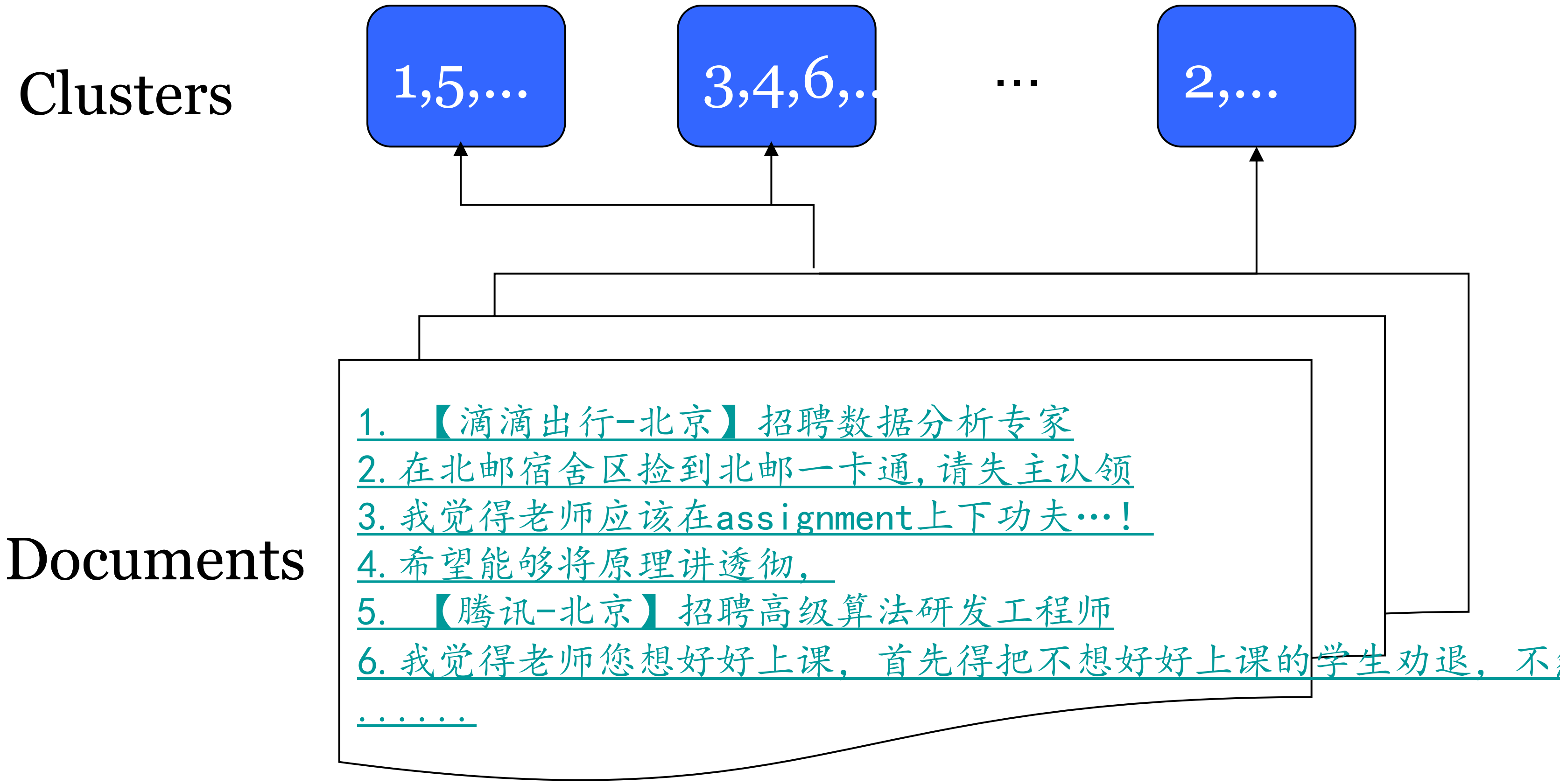


几个不同的学习方法

- 无监督/无指导学习：**Unsupervised Learning,** learning without a teacher
- 数据：D={$d^{(1)}$, $d^{(2)}$, ..., $d^{(n)}$}
  - $d^{(i)}= x^{(i)}$，输入向量
  - 没有目标类别/输出*y*
- 目标：
  - 学习样本间的关联、样本的成分等
- 问题类型：
  - 聚类
    - 将相似样本划分到同一簇，e.g., 患者病例
  - 密度估计：
    - 根据训练样本确定样本总体的概率分布

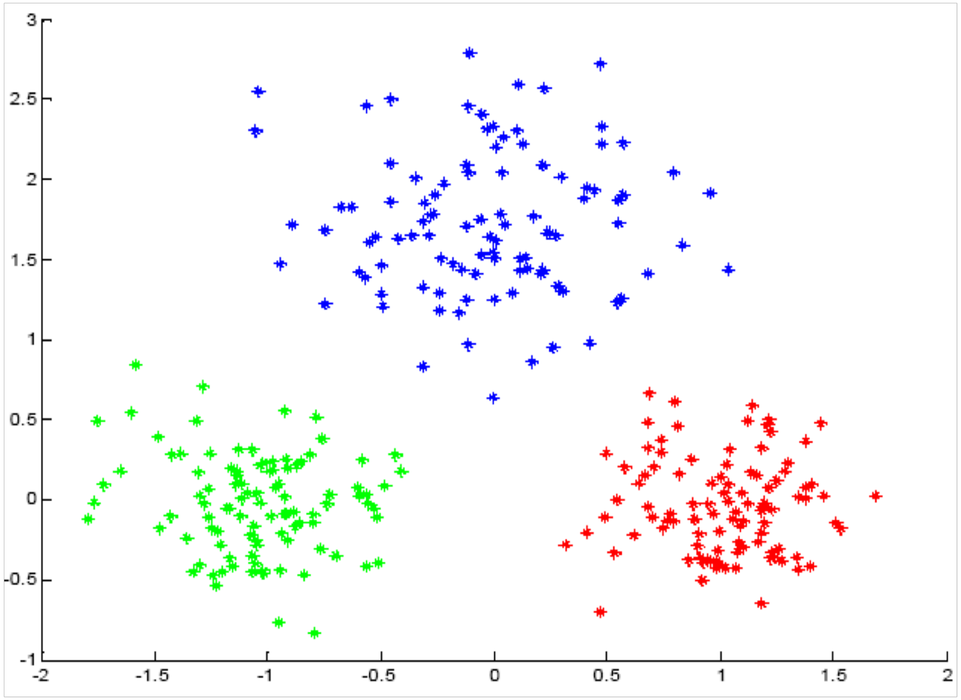
无监督学习：例子

- 聚类：


Clusters
1,5,…
3,4,6,..
…
2,…
Documents
1. 【滴滴出行-北京】招聘数据分析专家
2. 在北邮宿舍区捡到北邮一卡通，请失主认领
3. 我觉得老师应该在assignment上下功夫…！
4. 希望能够将原理讲透彻，
5. 【腾讯-北京】招聘高级算法研发工程师
6. 我觉得老师您想好好上课，首先得把不想好好上课的学生劝退，不
……

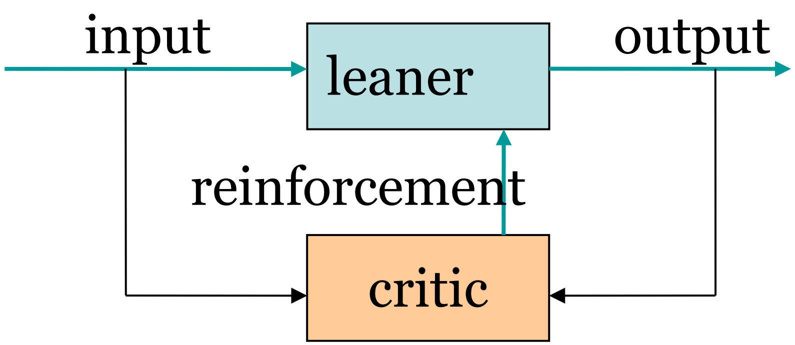
无监督学习：例子

- 密度估计：
  - 构建样本总体的概率模型，可以从中抽样得到 $d^{(i)}=x^{(i)}$



几个不同的学习方法

- 强化学习： **Reinforcement Learning,** learning with a critic
- 需要学习一个映射π: **S→A**
- 给定训练数据x（但没有y）
- 取而代之的是，可以得到一个判别学习器性能优劣的一个反馈（强化，reinforcement ）信号



- 目标：选择一个使得回报最大/损失最小的策略

强化学习：例子

- 动作（行为）：读幼儿园、读小学、读中学、读大学、练摊、学拉丁舞、早恋、间隔年……
- 状态：年龄、健康状况、财政状况……
- 回报：+100, -100
- 目标：人生赢家

Topics for Today

- What is Machine Learning (ML)?
- Why do we study ML?
- A rather simple sketch of ML history
- Different learning methods
- Key concepts
- Course Plan and Administrivia
- Recommended Textbooks

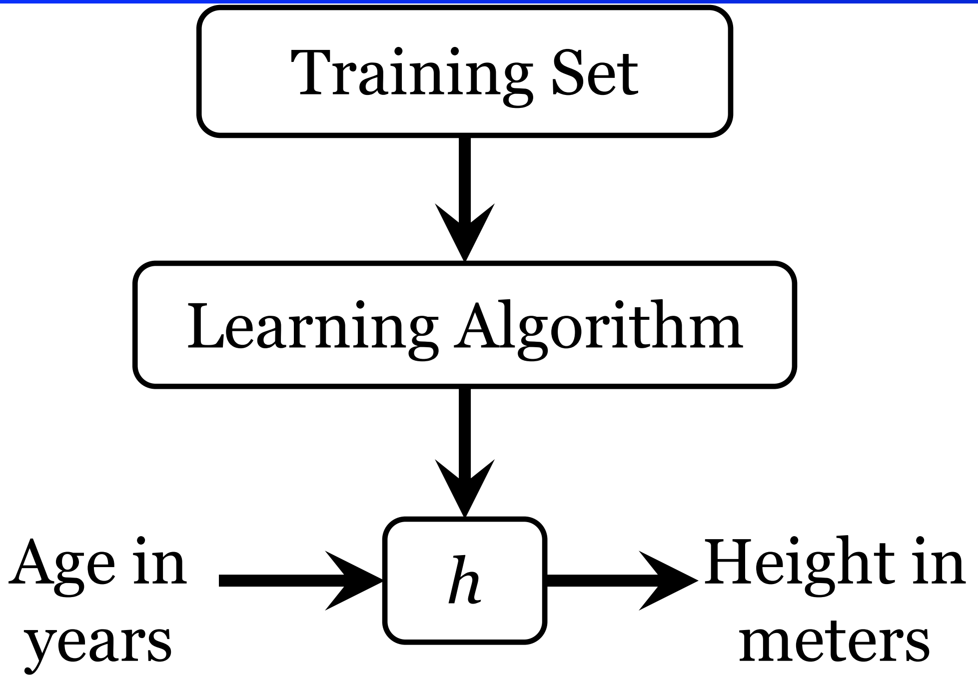
训练数据(Training data)

- E.g., 身高问题

Ages in year(x)	Height in meters(y)
2	0.78
2.6	0.90
3	0.96
5	1.04
…	…

  - x: 输入变量/特征
  - y: 输出变量/目标变量
  - (x, y):一个训练样本(training example)
  - $(x^{(i)}, y^{(i)})$: 第i个训练样本，i=1, …, n
  - n: 训练集(training set)中训练样本的个数

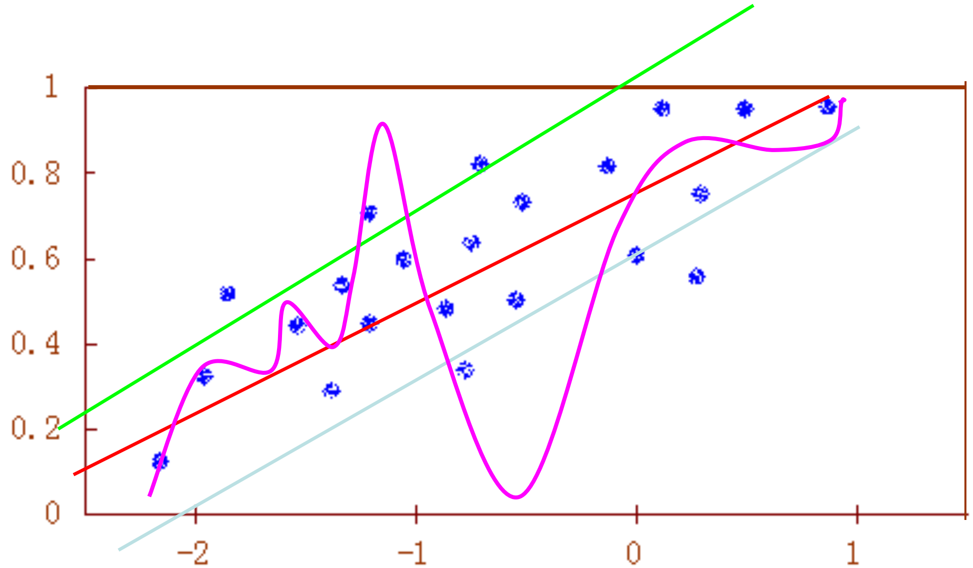
假设(Hypothesis)



- Hypothesis (h): 将x映射到y
- 如何表示h?
- $h_w(x)=w_1x+w_0$　或简写为：$h(x)=w_1x+w_0$
- w_i: 参数(parameters)
- 所有的假设组成的空间被称为假设空间

学习偏置(Learning Bias)

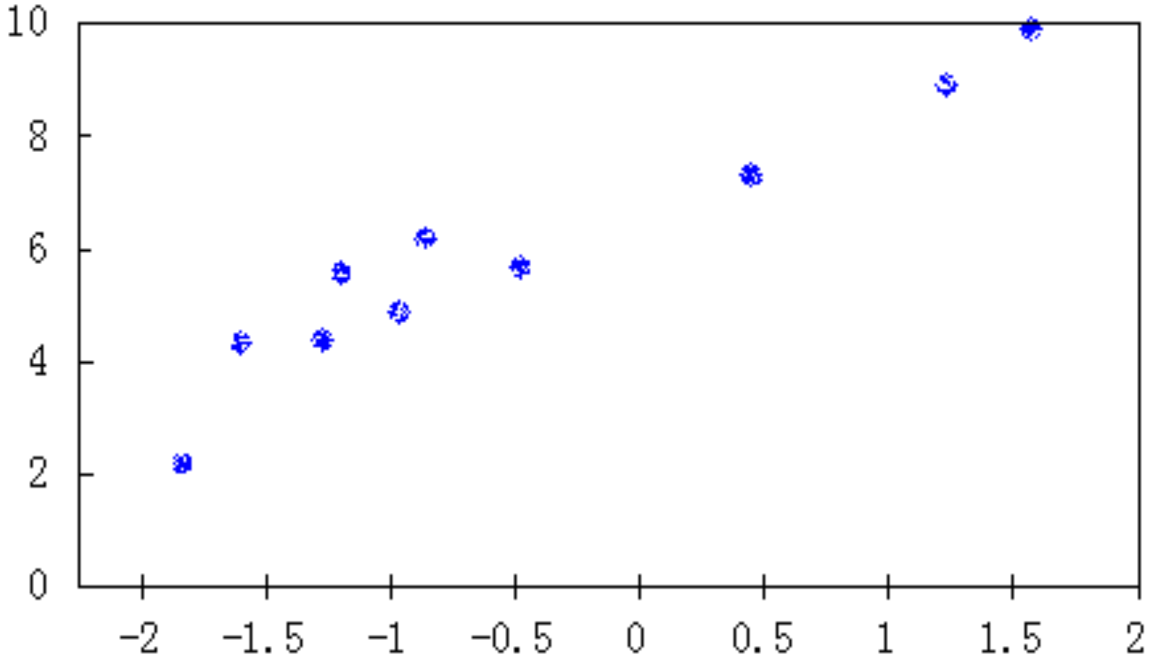
- 假设给定训练样本 (x, y)，希望学习得到一个映射 $h: x \to y$，为任意给定的x找到它对应的y



- 选择哪一个假设？仍然有很多样本未出现！
- 归纳偏置允许学习器在选择候选函数时有所倾向
- 任何一个有效的学习器都存在着学习偏置

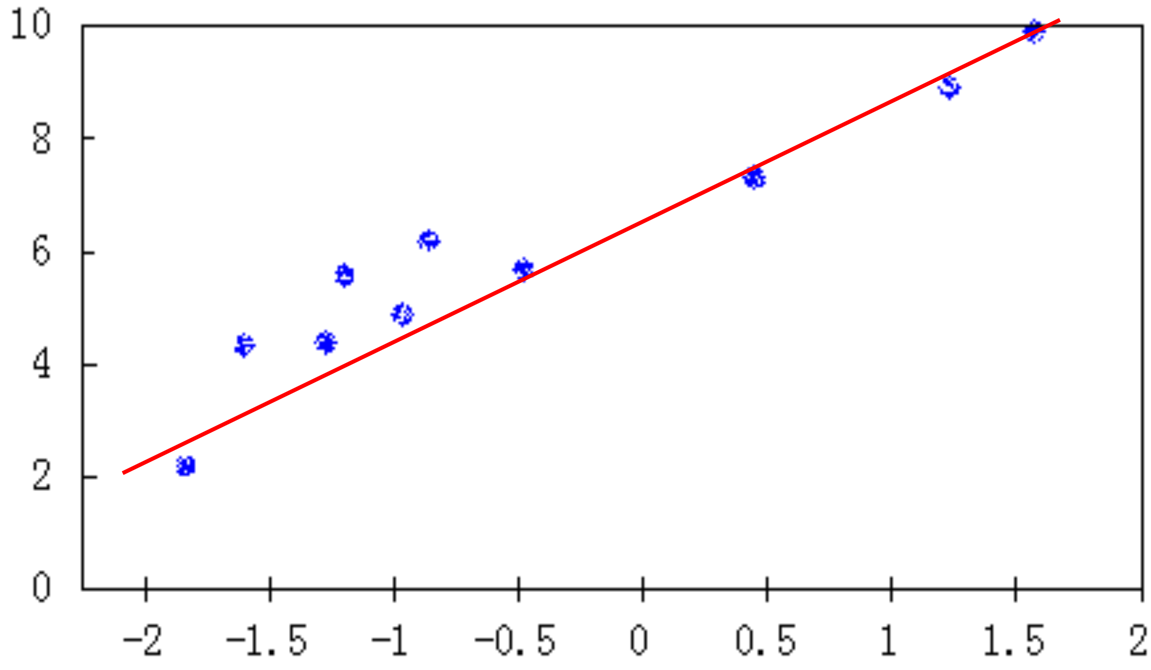
过拟合(Overfitting)

- 假设有10个训练样本，选择一个多项式函数 (polynomial function) 作为候选假设



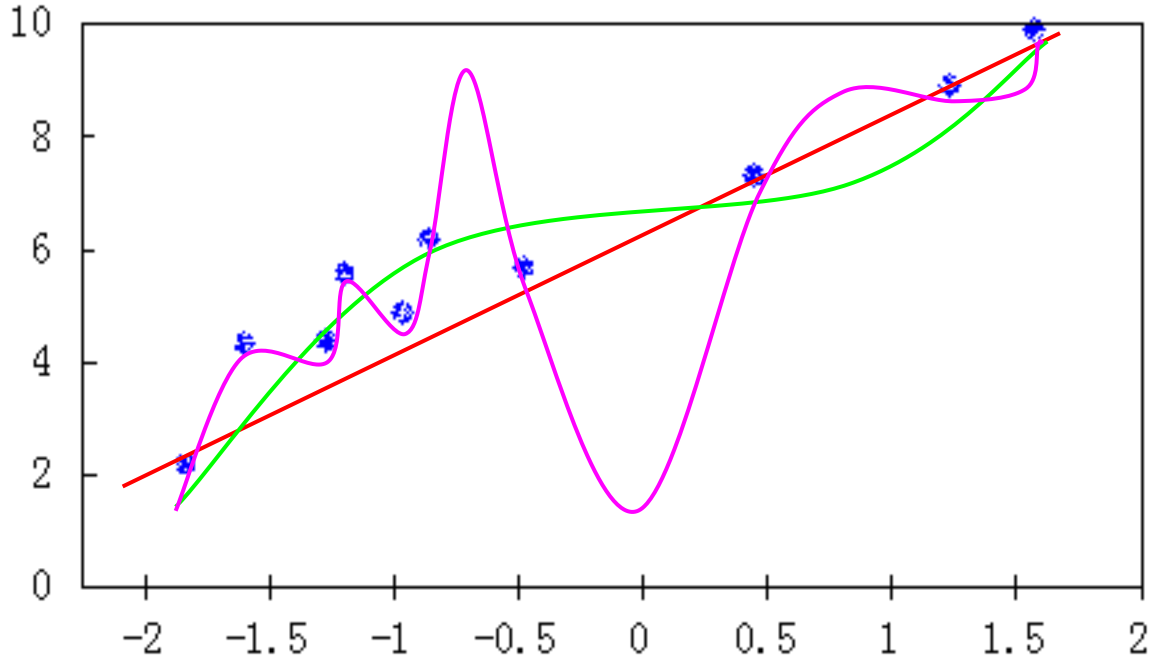
过拟合(cont.)

- 可以选择一个使均方误差最小的一次函数（线性函数）



过拟合(cont.)

- 可以选择一个使均方误差最小的三次函数
- 也可以选择更高阶的多项式函数 (以得到更小的误差)



过拟合(cont.)

- 对于10个样本点，可以选择一个9次多项式函数完美的拟合数据：误差为零
- 但是一味地最小化训练误差是否是一个好的选择？
  - NO!!!
- 更重要的是：模型在未知数据上的表现如何？
  - 一个情形：训练误差很小，但在未知数据上误差很大
- 原因：
  - 模型参数过多(自由度过大)
  - 训练数据太少(与模型的复杂度相比)
- 一个问题：如何使学习器避免过拟合？

两类误差

- 通常，模型选择及求解的三个步骤：
  - 选择一个或一组模型（含参数）
    - E.g. $h(x)=w_0 + w_1x$
  - 选择一个目标函数（可最优化）
    - E.g. 误差函数 (Error function/Cost function):
      - e.g., mean squared error function

$$\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - h(x^{(i)}))^2$$

  - 求解使目标函数最优的一组模型参数
    - E.g. 使误差最小的模型参数

两类误差

- 问题：
  - 通过过往的经验（观测数据）来学习模型参数
  - 但真正感兴趣的是：学习得到一个在所有数据（包括未知数据）上表现良好的一个模型
- 两类误差：
  - 训练误差(training error)：

$$\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-h(x^{(i)}))^2$$

  - 真实误差或泛化误差：(true error, generalization error):

$$E_{(x,y)}[(y-h(x))^2]$$

  - 往往需要采用训练误差逼近真实误差
  - 问题：一个好的训练误差是否一定意味着好的泛化误差？

两类误差

- 最优化(平均)训练误差可能会导致过拟合：

$$\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-h(x^{(i)}))^2$$

- 如何估计泛化误差？
- 两个思路：
  - **理论上：** 大数定律
    - 真实误差和样本平均误差之间的差异存在一个统计边界
  - **实际上：** 实验时从训练数据中抽取出m个数据用于模型的测试
    - **(平均)测试误差(test error)**

$$\frac{1}{m}\sum_{j=1}^{m}(y^{(j)}-h(x^{(j)}))^2$$

Topics for Today

- What is Machine Learning (ML)?
- Why do we study ML?
- A rather simple sketch of ML history
- Different learning methods
- Key concepts
- Course Plan and Administrivia
- Recommended Textbooks

课程内容

- Tentative Topics:
  - Linear models and generalized linear models: 1 unit
  - Neural networks: from shallow to deep: 2 units
  - Support vector machines: 2 units
  - Bayesian learning and naïve Bayesian classifier: 1 unit
  - Modeling sequential data: from HMM to CRFs: 2 units
  - K-means, Gaussion mixture model and EM algorithm: 2 units
  - Feature learning: 2 unites
  - Reinforcement learning: 2 units
  - Future Readings : 1 unit

课程内容

I (1 unit) **Introduction**

1. What is Machine Learning
2. A Simple Learning Example
3. Supervised & Unsupervised & Reinforcement Learning

II (1 unit) **Linear Models**

1. Linear Regression
2. Logistic Regression
3. Cost Function
4. Gradient Descent: Intuition and Algorithm
5. Overfitting and Regularization

III (2 units) **Neural Networks: From Shallow To Deep**

1. What are Neural Networks
2. Types of Neural Network Architectures
3. Perceptrons: Intuitions and Why the Learning Works
4. Multilayer Perceptrons (Feedforward Deep Networks)
5. The Backpropagation Algorithm
6. Deep Neural Networks
7. Three Classes of Deep Learning Networks
8. Stochastic Gradient Descent by Backpropagation
9. Why Are DNN Hard to Train

课程内容

IV (2 units) **Support Vector Machine**

1. Maximize minimum margin
2. Support vector
3. Hard linear model
4. Soft linear model with kernal methods

V (1 unit) **Bayesian learning**

1. Bayes rule
2. Bayes optimal classifier
3. Naïve Bayesian classifier

VI (2 units) **Modeling Sequential Data**

1. Examples of Complex Output Spaces
2. Hidden Markov Model
3. Discriminative Learning of HMM
4. From HMMs to Linear Chain CRFs
5. More Possibilities of CRFs

课程内容

VII (2 units) **Unsupervised Learning**

1. K-Means Algorithm
2. Gaussian Mixture Model
3. EM Algorithm

VIII (2 units) **Feature Learning**

1. PCA: a Motivating Example
2. From PCA to probablistic PCA
3. Kernal PCA

IX (2 units) **Reinforcement Learning**

1. Examples: Learning to Walk
2. Markov Decision Process
3. Passive Reinforcement Learning
4. Active Reinforcement Learning
5. Q-learning

X (1 unit) **Future Readings**

Deep Reinforcement Learning

考核方式

- Assignments during the course (80%)
  - 10+个不同分值的候选作业题目，须从中至少选够100分（也可自行命题，须提前讨论）
  - 期末时Oral presentation
- Others(20%)
  - 课程参与度：提问、回答等

  - Please don’t copy old solution!!

Topics for Today

- What is Machine Learning (ML)?
- Why do we study ML?
- A rather simple sketch of ML history
- Different learning methods
- Key concepts
- Course Plan and Administrivia
- Recommended Textbooks

推荐参考书

- 主要参考教材：
  - 周志华. *机器学习*. 清华大学出版社, 2016. -- elementary and interesting
  - 李航. *统计学习方法*.清华大学出版社, 2012. -- essential and but systematically poor
  - T. Mitchell. *Machine Learning*. McGraw Hill, 1997. (Chinese ed., 2003). -- old but classical
  - Duda, Hart, Stork. *Pattern Classification*. 2nd edition. J. Wiley and Sons, 2000. -- elementary and interesting
  - Hastie, Tibshirani and Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009. -- difficult but essential
  - Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. -- heavy but charming
- 其它：
  - Alpaydin, E. *Introduction to Machine learning*. MIT Press, 2004.
  - Flach, P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.
  - Ian Goodfellow, Yoshua Bengio, Aaron Courville. *Deep Learning*.人民邮电出版社, 2017.

推荐公开教程

- A portion of slides are borrowed from
  - Michael I. Jordan, UC Berkeley at http://www.cs.berkeley.edu/~jordan/courses/
  - Andrew W. Moore, CMU at http://www.cs.cmu.edu/~awm/tutorials
  - Yoshua Bengio, U. Montreal at https://ift6266h16.wordpress.com/
  - Ng. Andrew, Stanford University at http://www.stanford.edu/class/cs229/
  - Feifei-Li, Stanford University at http://cs231n.stanford.edu/syllabus.html
  - ...
  - And many of the figures are provided by Chris Bishop's textbook: "Pattern Recognition and Machine Learning"
- Open courses
  - Machine Learning
  - Neural Networks for Machine Learning
  - deaplearning.ai
  - ...

Useful resources & links

- Conferences:
  - International Conference on Machine Learning (ICML)
  - Neural Information Processing Systems Conference (NIPS)
  - Annual Conference on Learning Theory(COLT)
  - Association for the Advancement of Artificial Intelligence (AAAI)
  - International Joint Conferences on Artificial Intelligence (IJCAI)
  - ICLR (International Conference on Learning Representations)
  - International Conference on Data Mining (ICDM)
  - ACM SIGIR
  - ACM SIGKDD
  - ACM CIKM
  - Association for Computational Linguistics (ACL)
  - IEEE Conference on Computer Vision and Pattern Recognition
  - ...
- Journals:
  - Machine Learning Journal
  - Journal of Machine Learning Research
  - Artificial Intelligence Journals
  - ACM Transactions on Knowledge Discovery from Data
  - Data Mining and Knowledge Discovery
  - IEEE-Transactions on Pattern Analysis and Machine Intelligence
  - Neural Computation
  - IEEE-Transactions on Neural Networks and Learning Systems

- Home work: Basic Mathematics Review
- Next lecture: Linear Models
  - Ref. 周志华, Ch3

Thank you!