机器学习

Machine Learning

主讲人：张敏 清华大学长聘副教授

ML 机器学习

MACHINE LEARNING-MIN ZHANG

Unit.12

机器学习算法总结

*图片均来自网络或已发表刊物

总览

- 基础概念
- 机器学习方法
- 一些深入话题
- 实验相关
- 学习理论分析
- 总结

I. 基本概念

C1. 什么是机器学习?

- 学习

  = 在某种任务上基于经验不断进步

- T （Task）
- P （Performance）
- E （Experience）

C2. 归纳学习假设

- 归纳学习假设:

  Any hypothesis found to approximate the target function well over **a sufficiently large set** of training examples will also approximate the target function well over unobserved examples.

  (任一假设若在足够大的训练样例集中很好地逼近目标函数,它也能在未见实例中很好地逼近目标函数)

II. 机器学习方法

有监督和无监督学习

	有监督	无监督
训练样例	(X, Y) 对, 通常包含人为的努力	仅 X, 通常不涉及人力

有监督和无监督学习

	有监督	无监督
训练样例	(*X*, *Y*) 对, 通常包含人为的努力	仅 *X*, 通常不涉及人力
学习目标	学习 *X* 和 *Y* 的关系	学习 *X* 的结构

有监督和无监督学习

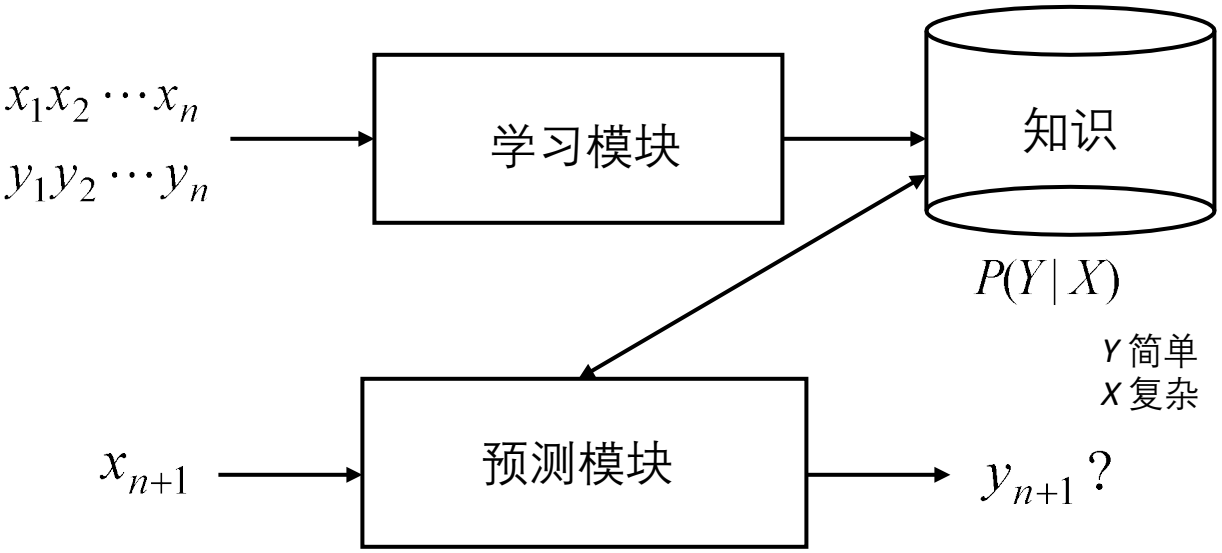
	有监督	无监督
训练样例	(*X*, *Y*) 对, 通常包含人为的努力	仅 *X*, 通常不涉及人力
学习目标	学习 *X* 和 *Y* 的关系	学习 *X* 的结构
效果衡量	损失函数	无

有监督和无监督学习

	有监督	无监督
训练样例	(*X*, *Y*) 对, 通常包含人为的努力	仅 *X*, 通常不涉及人力
学习目标	学习 *X* 和 *Y* 的关系	学习 *X* 的结构
效果衡量	损失函数	无
应用	预测: *X*=输入, *Y*=输出	分析: *X*=输入

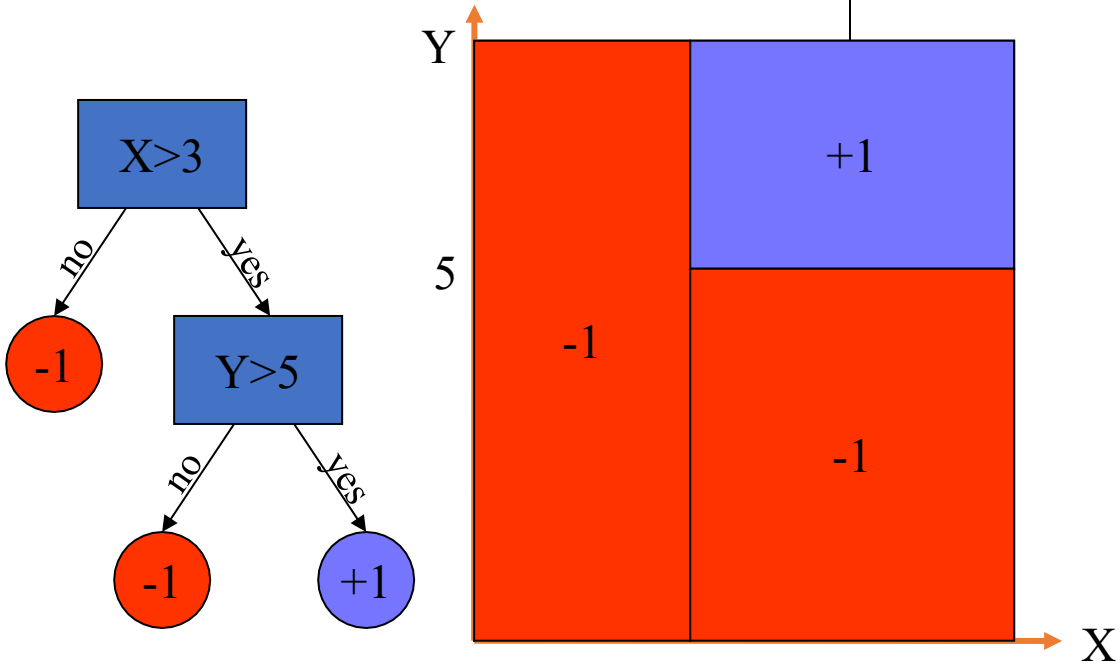
II. 机器学习方法 (part I)
——监督学习

监督学习



A1. 决策树

- 通过各个属性的分割决策，对数据进行有效模型化



- 决策树:
  用 概念/规则 表示假设

- 直观上来看，所学到的假设很容易得到对应的解释
- 但如果我们无法从观测到的数据中得到显式的规则怎么办？

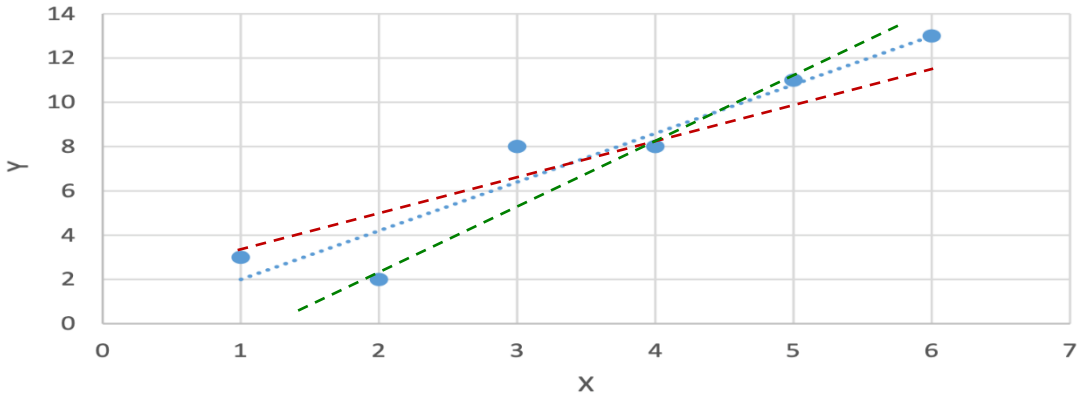
用基于统计的方法

A2. 回归

- 线性假设：目标变量可以被各个属性线性表出

在线课程满意度 = 1.6 + 0.11 × 平台交互性 + 0.15 × 教学资源 + 0.27 × 教学设计

常数项 (截距)　　系数 (斜率)



线性回归

- 基于误差平方和的最小二乘法拟合
- 求解简单容易计算，可以直接得到预测用的公式
- 对于离群点较为敏感

- 如果属性和目标之间不满足线性假设怎么办？

A3. 贝叶斯学习

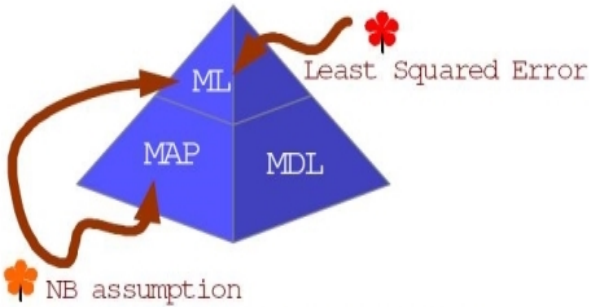
- 条件 → 结果
  - e.g. 肺炎 → 肺癌?
  - 很难直接判断
- 反过来想

  (结果 → 诱因)
  - e.g. 有多少肺癌患者同时患有肺炎？

贝叶斯学习

- 贝叶斯定理
  - 用先验概率来推断后验概率
- Max A Posterior, MAP，h_{MAP}，极大后验假设
  - 通常来说我们希望得到给定训练数据下最有可能的假设
- Maximum Likelihood, ML, h_{ML}，极大似然假设
  - 如果知道 *p*(h)，最聪明的人总是能最大限度地从经验中学习
  - ML vs. LSE (Least Square Error)
- Naïve Bayes, NB, 朴素贝叶斯
  - 独立性假设
  - NB vs. MAP
- Minimum description length, MDL,最小描述长度
  - Tradeoff: 假设复杂度 vs. h 的误差
  - MDL vs. MAP

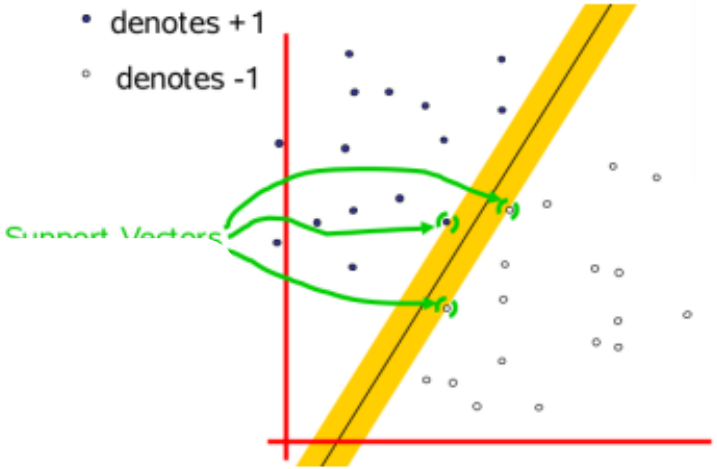
$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

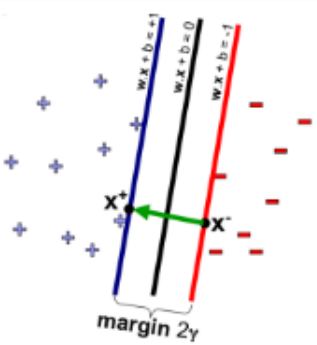


A4. 核方法以及非线性 SVM

最大化间隔

- 定义：线性分类器的间隔指从分界面向两边扩展直到第一次遇到数据点所形成的最大宽度
- 最大化间隔





$$\max_{w,b} \frac{1}{< w, w >}$$

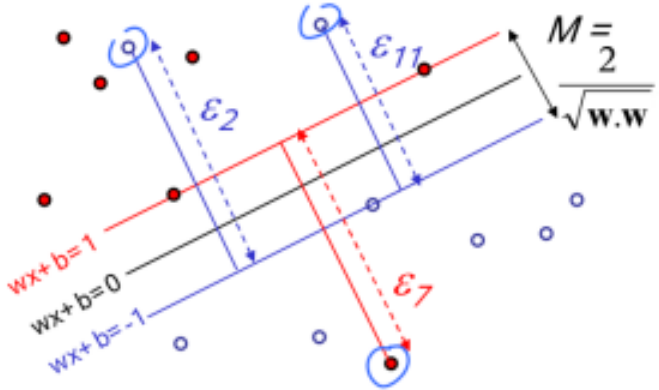
Separable case

$$\min_{w,b} \frac{1}{2} < w, w >$$

$$s.t. y_i \left(< w, x_i > + b \right) \geq 1$$

不可分的情况

- 最小化训练误差



Non-separable Case

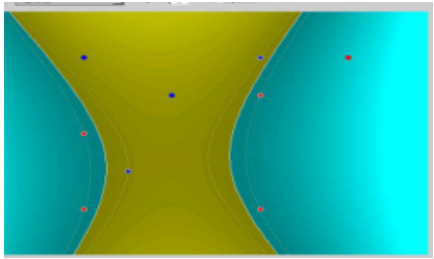
$$\min_{w,b} \frac{1}{2} < w, w > + C \sum_i \varepsilon_i$$

$$s.t. \left(< w, x_i > + b \right) y_i \geq 1 - \varepsilon_i$$

$$\varepsilon_i \geq 0$$

非线性 SVM

- 输入空间 → 特征空间 $\Phi(x): R^n \mapsto F$
- 低维下的非线性 → 高维的线性超平面
- 常见的核函数
  - Polynomials of degree d $K(x,y) = (< x, y >)^d$
  - Polynomials of degree up to d $K(x,y) = (< x, y > +1)^d$
  - Gauss Kernel $K(x,y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right)$
  - Sigmod Kernel $K(x,y) = \tanh(\eta < x, y > +v)$
- 软件
  - LIBSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm
  - SVMlight http://svmlight.joachims.org



- 之前的学习方法
  - 估计问题特性 (e.g. 分布)
  - 做一个模型假设
  - 找到最优的参数

↓

但有时我们在学习之前什么也不知道

↓

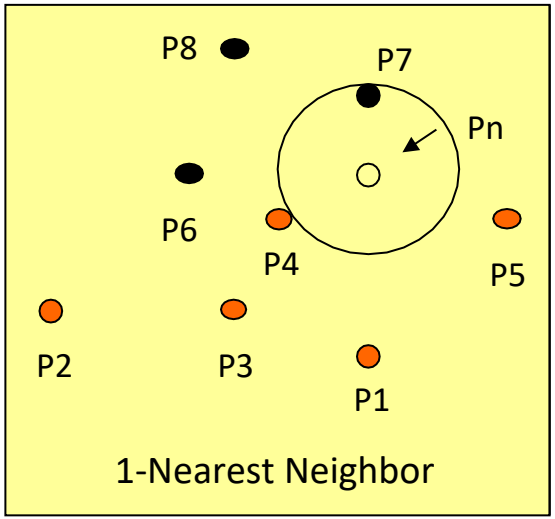
是否有一种学习方法不遵循

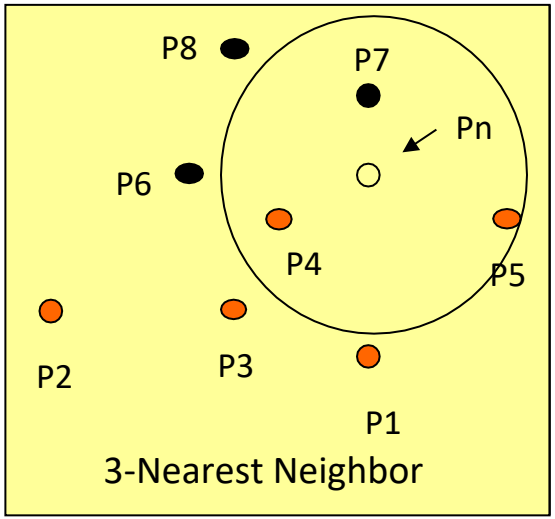
“模型假设 + 参数估计”？

A5. k-Nearest Neighbor (KNN)

- 思考即回忆、进行类比
- One takes the behavior of one's company

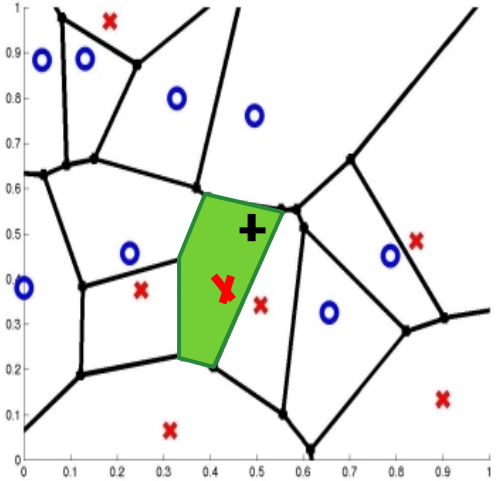
  “近朱者赤，近墨者黑”





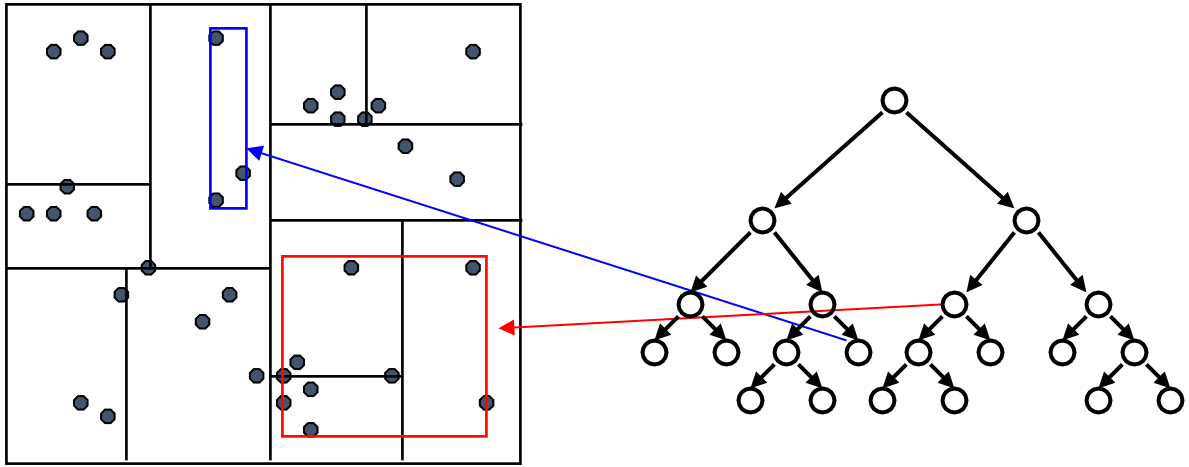
KNN

- 主要的假设
  - 存在一种有效的距离度量
- 非参数化
- 概念简单，但可以建模任何函数



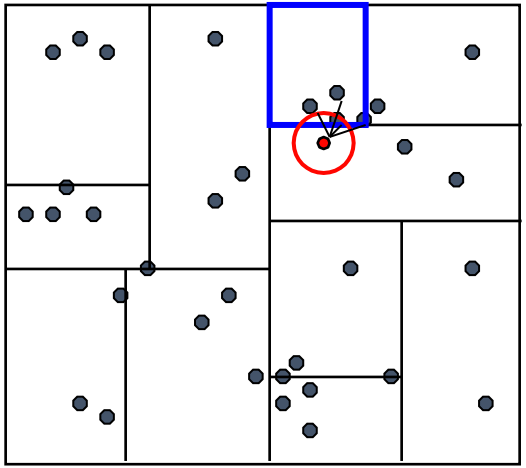
- 内存开销大
- CPU 开销大
- 特征选择问题
  - 不相关的特征 对距离度量有消极的影响
- 对如何表示数据很敏感

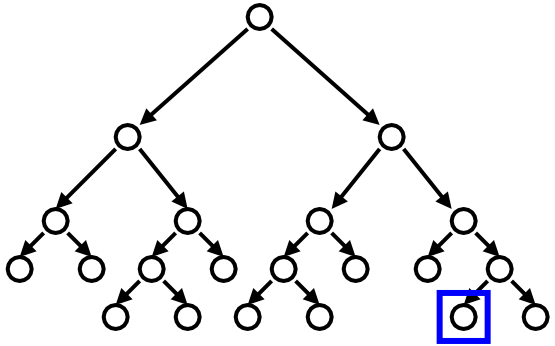
效率问题 – KD-Tree (构建)



每个节点维护一个额外的信息：**这个节点包含数据点的（紧）边界**

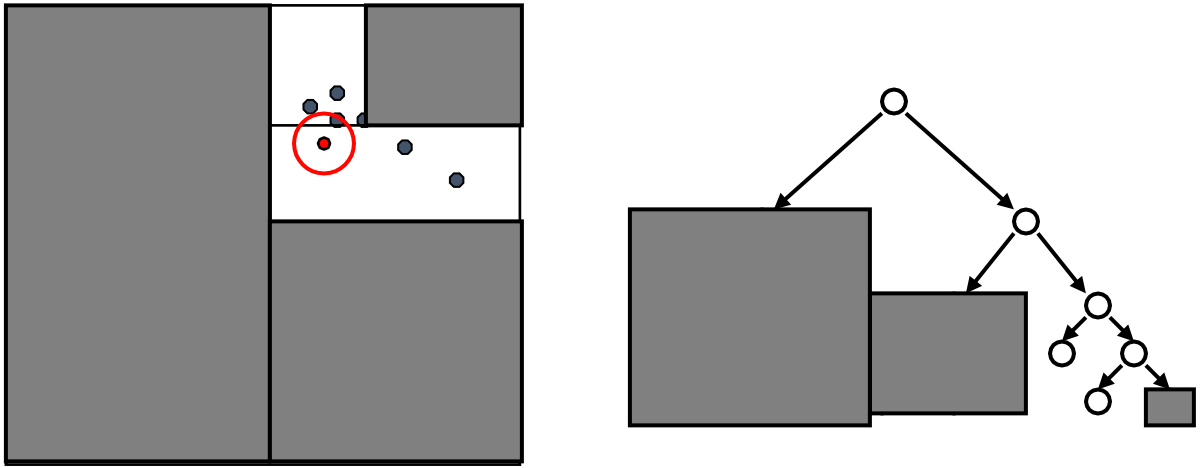
效率问题 – KD-Tree (查询)





每次发现一个新的最近的点，就更新距离上界

效率问题 – KD-Tree (查询)



利用这个最近距离以及每个树节点下数据的边界信息，我们可以对一部分**不可能**包含最近邻居的分支进行剪枝

基于记忆的学习器：4个要素

1. 一种距离度量

   Euclidian / Scaled Euclidian / ……

2. 使用多少个邻居？

   1, k 或全部

3. 一个加权函数（加权）

   $w_i = exp(-D(x_i, query)^2 / K_w^2)$

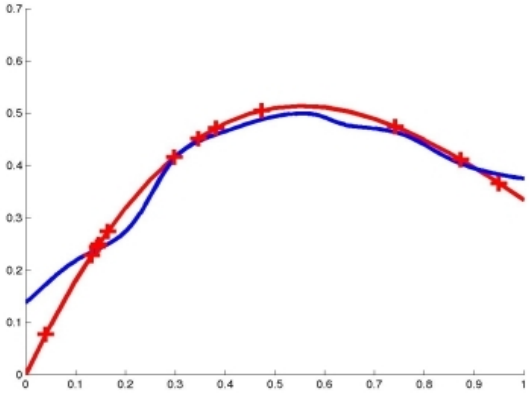
4. 如何使用已知的邻居节点？

   最近的邻居，或

   K 个邻居投票，或

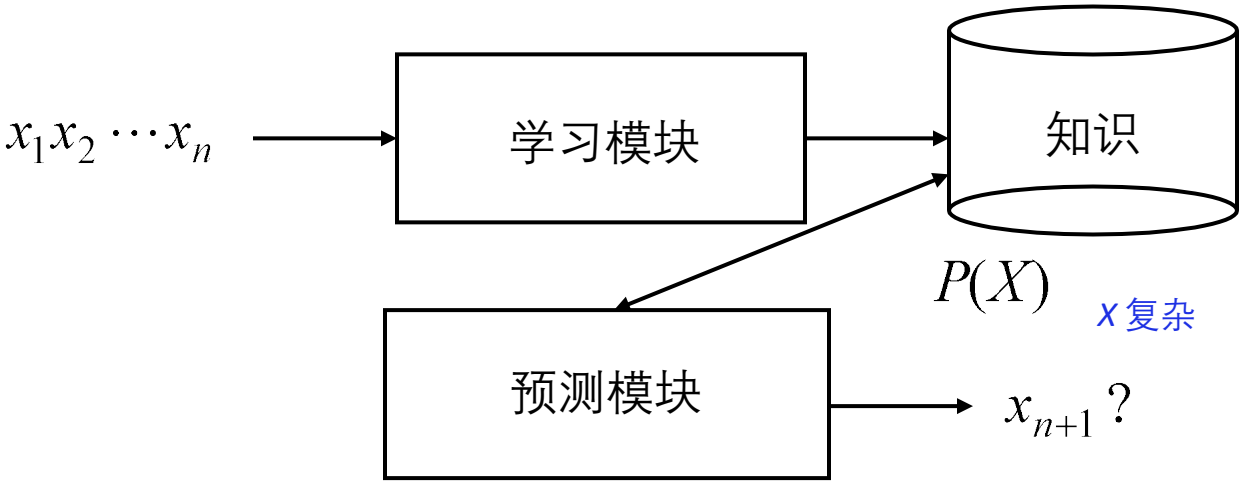
   输出的加权平均

   predict = $\Sigma w_i y_i / \Sigma w_i$



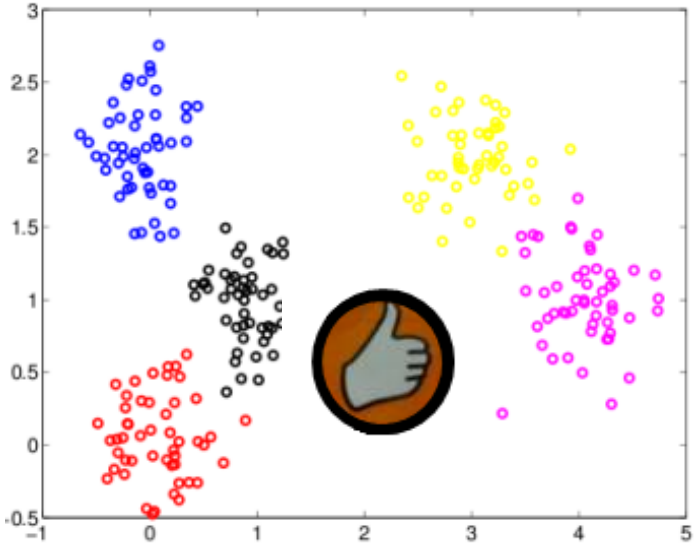
II. 机器学习方法(part II)
——无监督学习

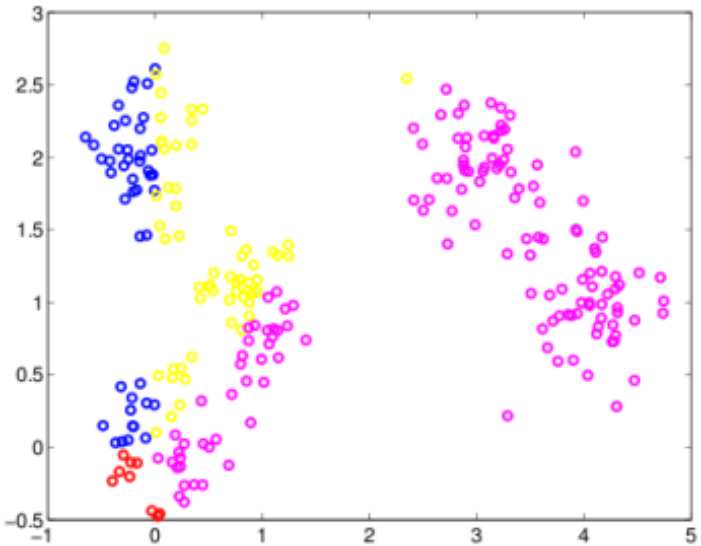
无监督学习



- 构建一个模型并找到输入的有用表示，使得这个表示可以用来做决策、预测未来的输入、高效输入其他学习器等
- 找到数据中独立于非结构化噪音之外的模式（发现结构）

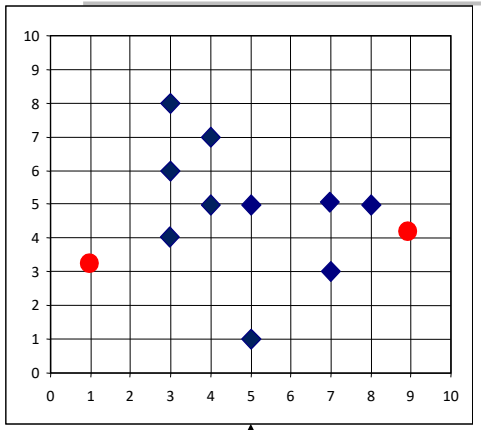
什么是好的聚类？





- 类内的距离小
- 类间的距离大

A6. K-Means



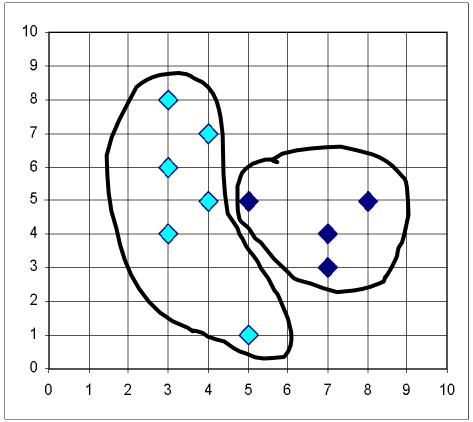
K=2

随机选择 K 个对象作为类中心

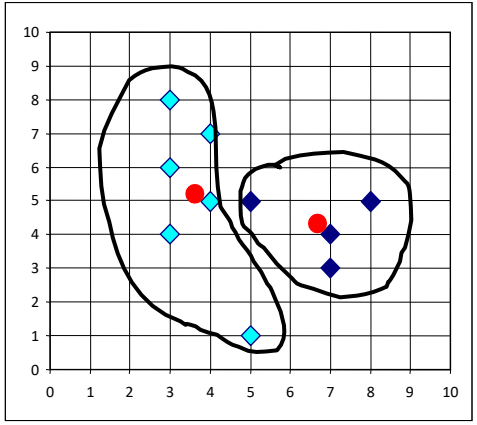
循环

直到没有变化

把每个对象分配到最近的类中心

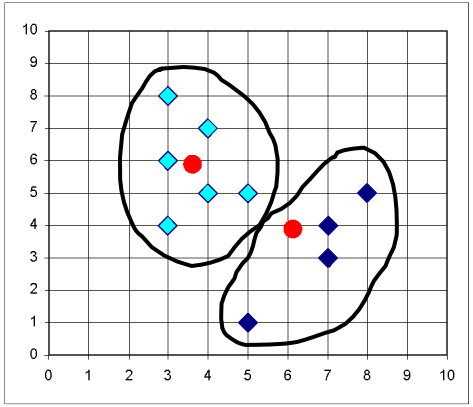


更新类中心

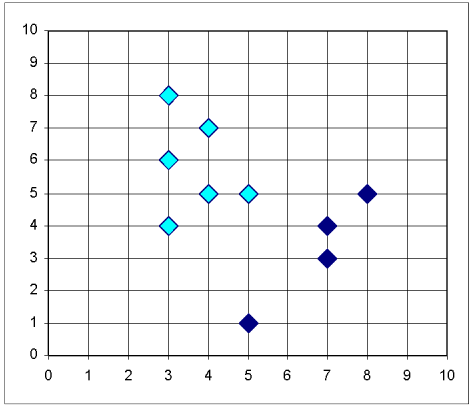


重新分配

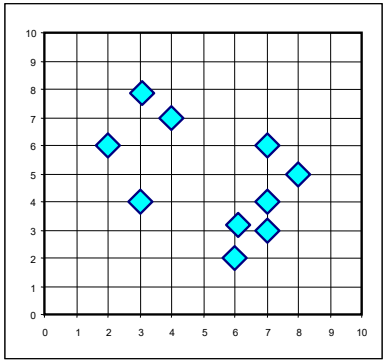
重新分配



更新类中心

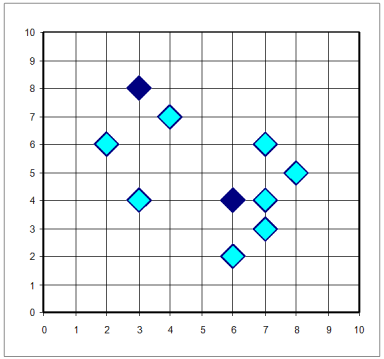


A7. K-Medoids



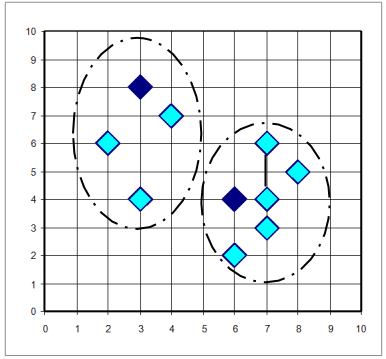
K=2

随机选择 k 个对象作为初始中心

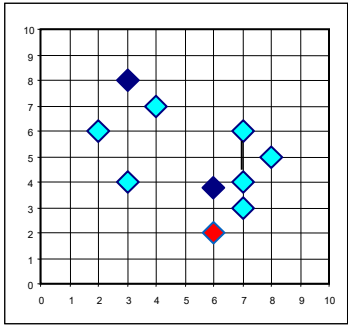


将每个对象分配给最近的中心

总代价 = 20



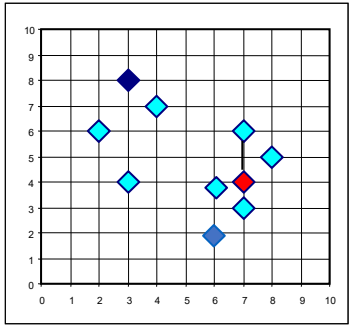
随机选择一个非中心点 O_{random}



计算总代价

代价= 26

不交换



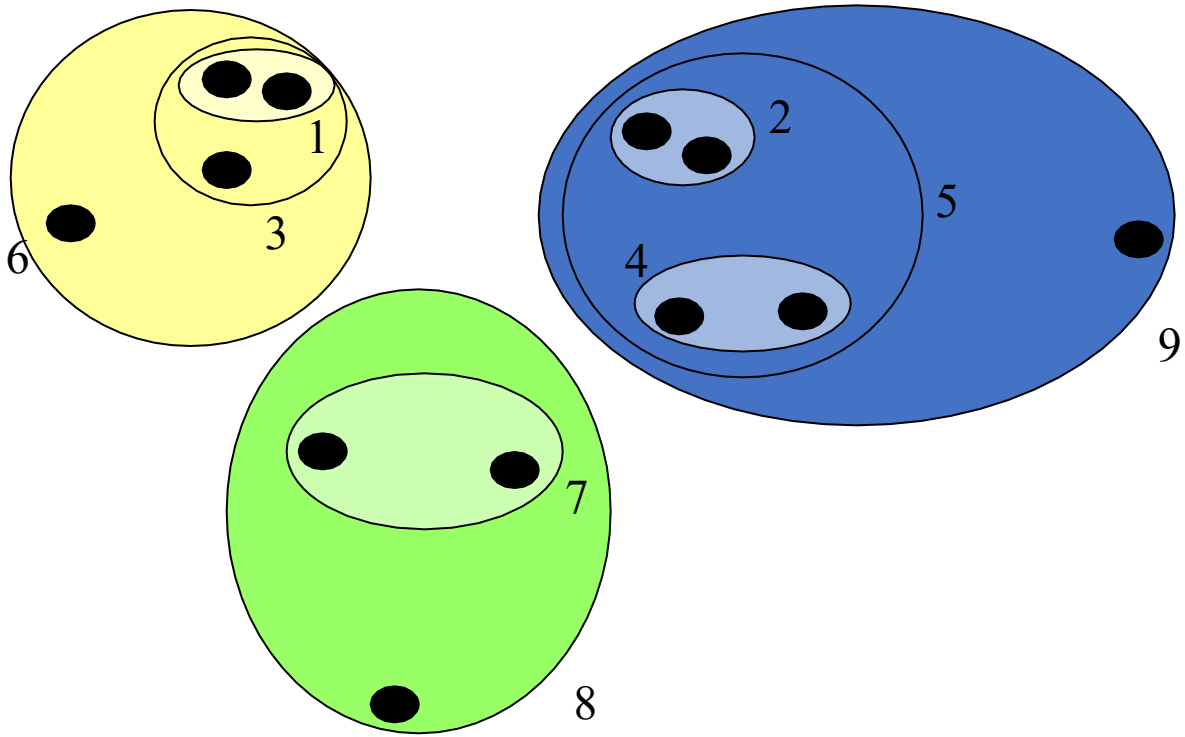
交换 O 和 O_{random}

如果质量提升

循环

直到没有变化

A8. 层次聚类 (凝聚式)

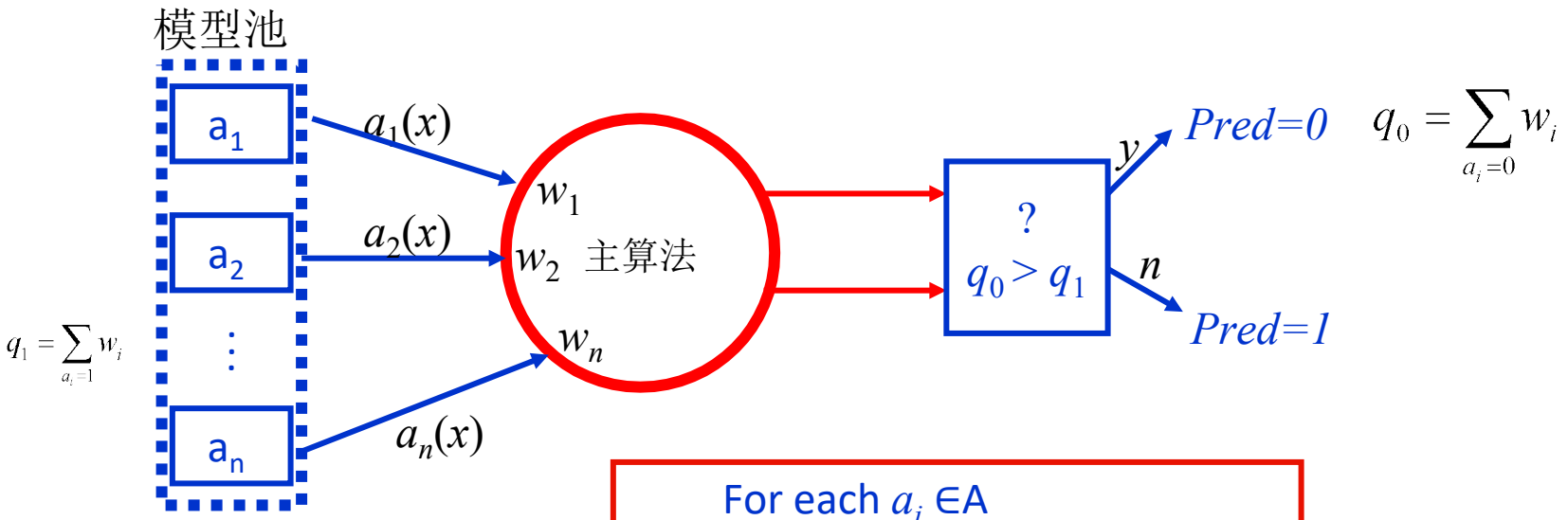


III. 一些深入话题
——集成学习

“Two heads are better than one.”

“三个臭皮匠，顶一个诸葛亮”

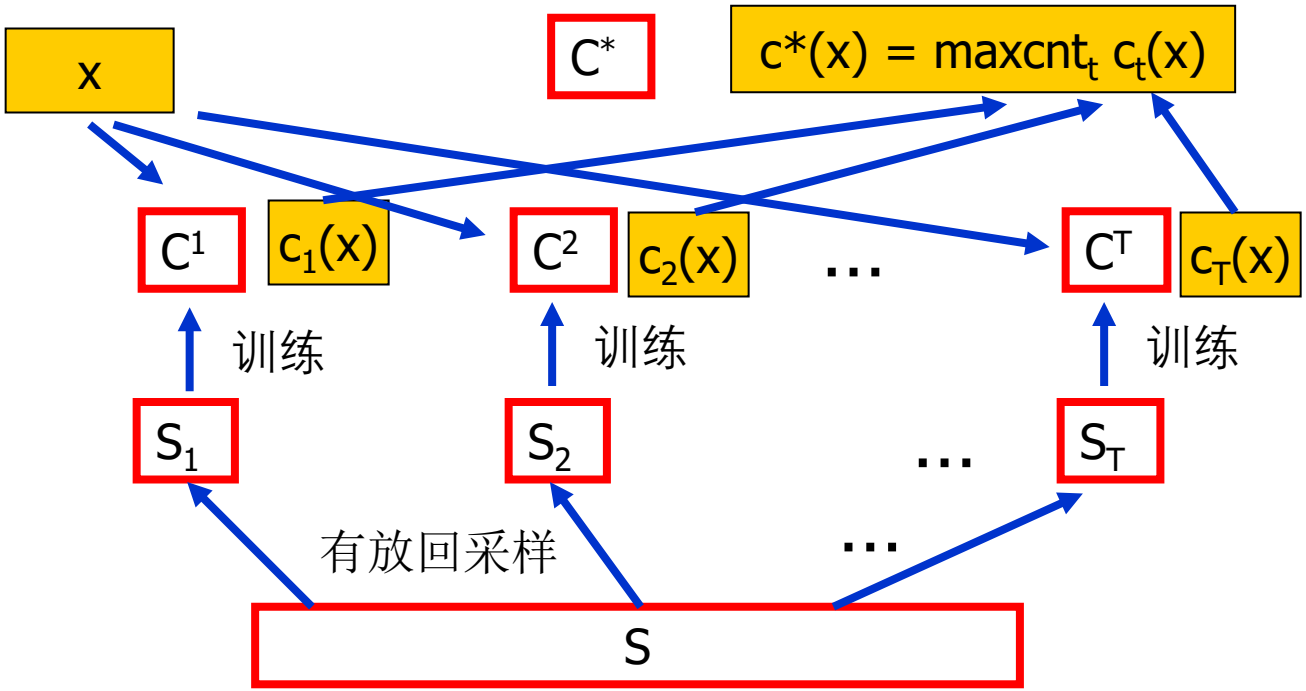
A9. 加权多数算法



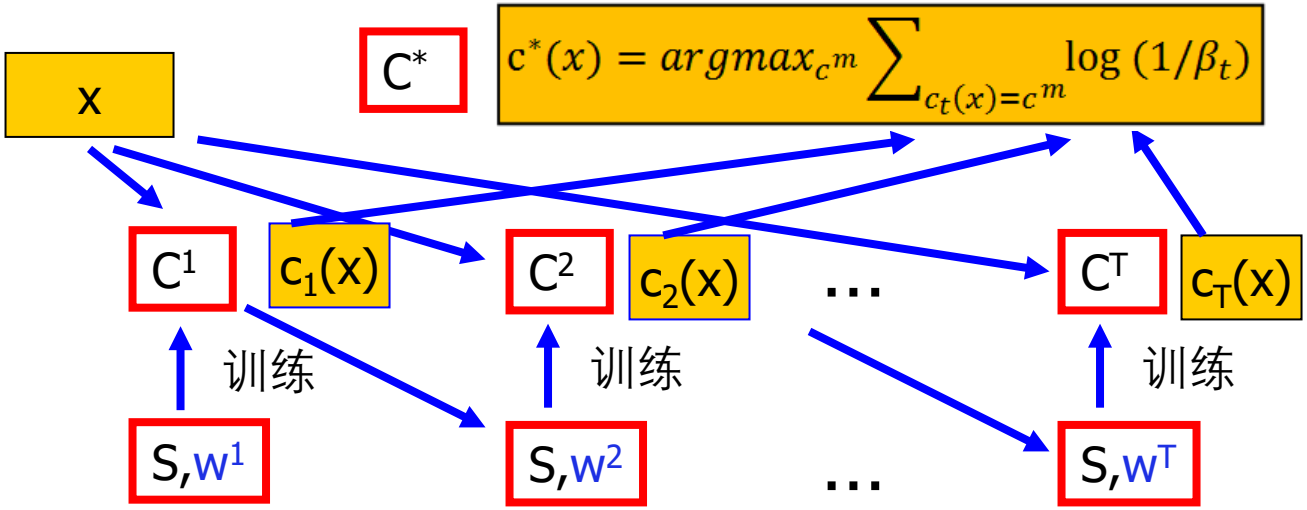
假设：二值输出 {0,1}

For each $a_i \in A$
If $a_i(x) \neq c(x)$, then $w_i \leftarrow \beta w_i$ $\beta \in [0,1)$

A10. Bagging (boostrap aggregating)



A11.Boosting (从错误中学习)



怎样是一个好的弱学习器？

弱学习器的准则（特征）应当包括：

- 不稳定: 训练数据的小变化，能够造成模型的大改变
- 简单: 在非平凡的加权训练误差下能够高效地进行拟合，根据观测样本对预测的计算应该很快
- 模型小: 避免过拟合

重加权 vs. 重采样

III. 一些深入话题
——深度学习

A12. 深度学习：什么时候有用？

- 有充足的计算资源
- 有充足的数据
  - 通常有较复杂的网络结构
- 当不知道怎么挑选好的特征时
- Deep = Deep nets (网络有很多层)
- 当前在 DL 方法的学习过程中很少用到深入的理论知识

我们大概介绍了什么

- Multi-layer perceptron (MLP)
- Convolutional neural nets (CNN)
- Sequential neural nets
  - Recurrent neural networks (RNN)
  - Long short-term memory (LSTM)
  - Gated recurrent unit (GRU)
- 应用

我们大概介绍了什么

- 讨论
  - 令人欣喜的结果，网络变得越来越深
  - 在大规模数据上十分有用，但模型也越来越大
    - 并行计算
  - 需要更多的理论基础
  - 缺少可解释性
  - 对于恶意的攻击不鲁棒
  - 无监督学习还有很大研究空间
- 充分利用在线资源

IV. 实验相关
——过拟合问题

E1. 过拟合问题

- 假设空间 H
  - 考虑的假设集合

过拟合问题

- 一致的假设
  - 成功拟合了所有数据

- $h \in H$ overfits training data if there's an alternative h' $\in H$ such that:

$$err_{train}(h) < err_{train}(h')$$

AND

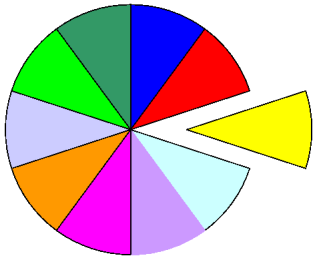
$$err_{test}(h) > err_{test}(h')$$

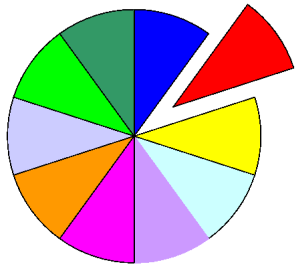
两点 tips:
- 泛化性能
- 在同一个数据集比较两个算法！
  (如果使用不同的数据集: 不同数据 → 不同表现)

IV. 实验相关
——有限数据

在有限数据上学习 (1)：E2. 交叉验证

- 当数据十分有限时
  - 如何更好地用这些数据去同时学习一个假设以及验证它的准确性？
- k- fold cross validation 交叉验证
  - 用平均误差去顾及整体误差





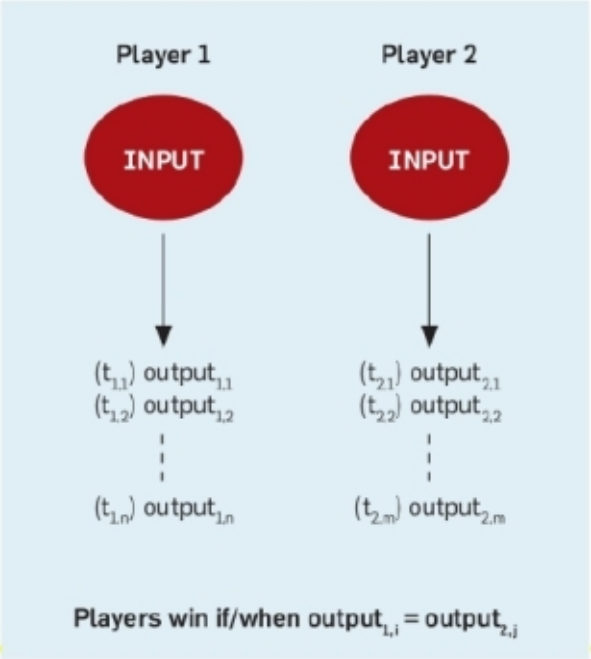
在有限数据上学习 (2)： E3. Boostrap 采样

- Bootstrap 采样
  - 给定一个包含 m 个训练样例的集合 D
  - 有放回地从 D 中均匀随机采样 m 个样例组成 D_i

IV. 实验相关
——用GWAP收集数据

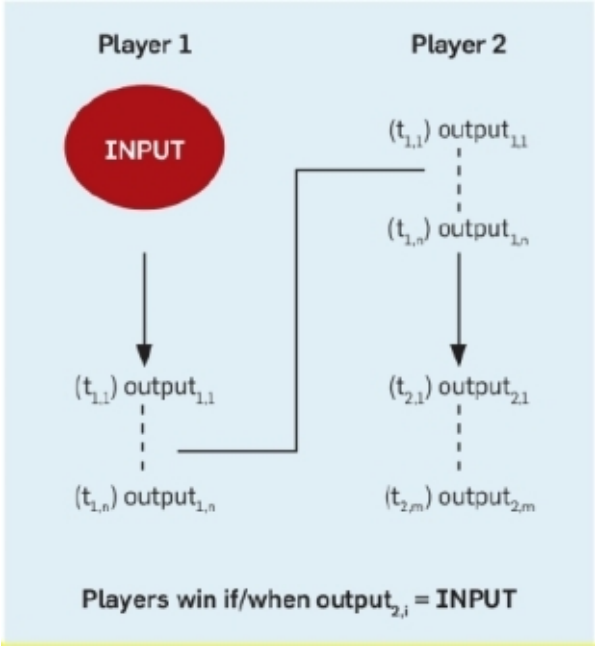
E4. 三种游戏结构 (1)

- 输出一致游戏（Output-agreement games）
  - ESP 游戏



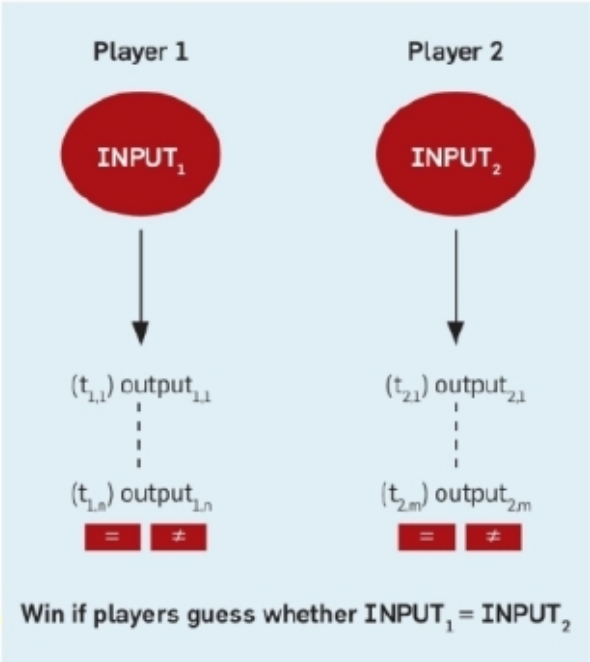
E4. 三种游戏结构 (2)

- 反演问题游戏（Inversion-problem games）
  - Peekaboom
  - Phetch



E4. 三种游戏结构 (3)

- 输入一致游戏（Input-agreement games）
  - Tag a tune



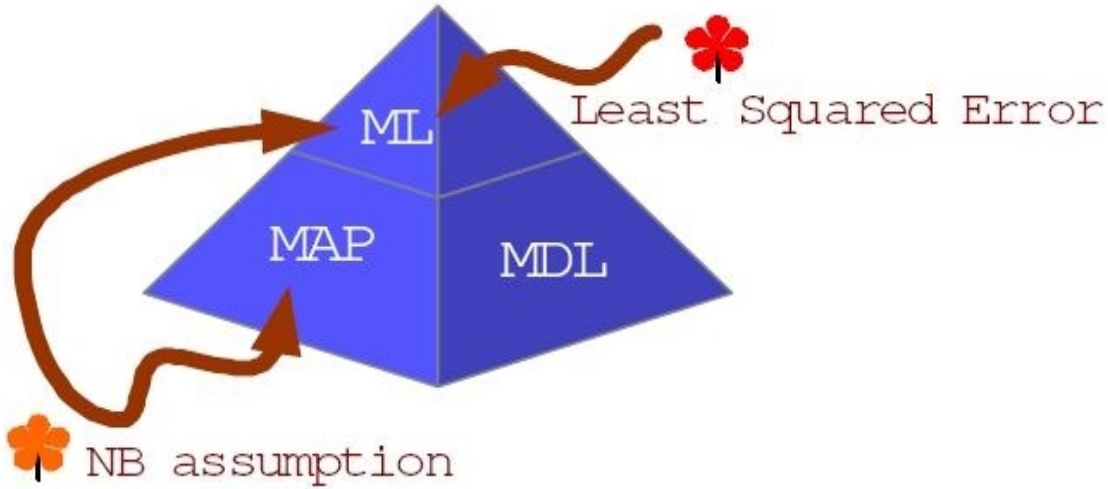
IV. 实验相关
——准则

ML 实验准则

- 不要在训练集上进行测试
- 重复实验
- 对比分析
- 统计显著性检验

V. 学习理论分析

T1. Bayesian statistics



T2. 最小描述长度 (MDL)

$$h_{\text{MDL}} = \underset{h \in \mathcal{H}}{argmin} \{L_{C_1}(h) + L_{C_2}(D|h)\}$$

- Tradeoff: 假设复杂性 vs. 假设犯的错误
  - 偏向于一个更简单的、犯比较少错误的假设
  - 而不是更复杂、能完美地分类训练数据的假设

解决过拟合问题

$$\begin{aligned} h_{\text{MAP}} &= \underset{h \in \mathcal{H}}{argmax}\, \text{P}(D|h)\text{P}(h) \\ &= \underset{h \in \mathcal{H}}{argmax} \{\log_2 \text{P}(D|h) + \log_2 \text{P}(h)\} \\ &= \underset{h \in \mathcal{H}}{argmin} \{-\log_2 \text{P}(D|h) - \log_2 \text{P}(h)\} \\ &= h_{\text{MDL}} \end{aligned}$$

$L_{C_2}(D|h)$ $L_{C_1}(h)$

VI. 总结

概念: 2

算法: 12 (深度学习部分记为 1 个)

实验方法: 4

理论: 2