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.08

无监督学习（I）

*图片均来自网络或已发表刊物

目录

无监督学习介绍

- 机器学习算法分类
  - 贪婪 vs. 懒惰
  - 参数化 vs. 非参数化
  - 有监督 vs. 无监督 vs. 半监督
  - ……

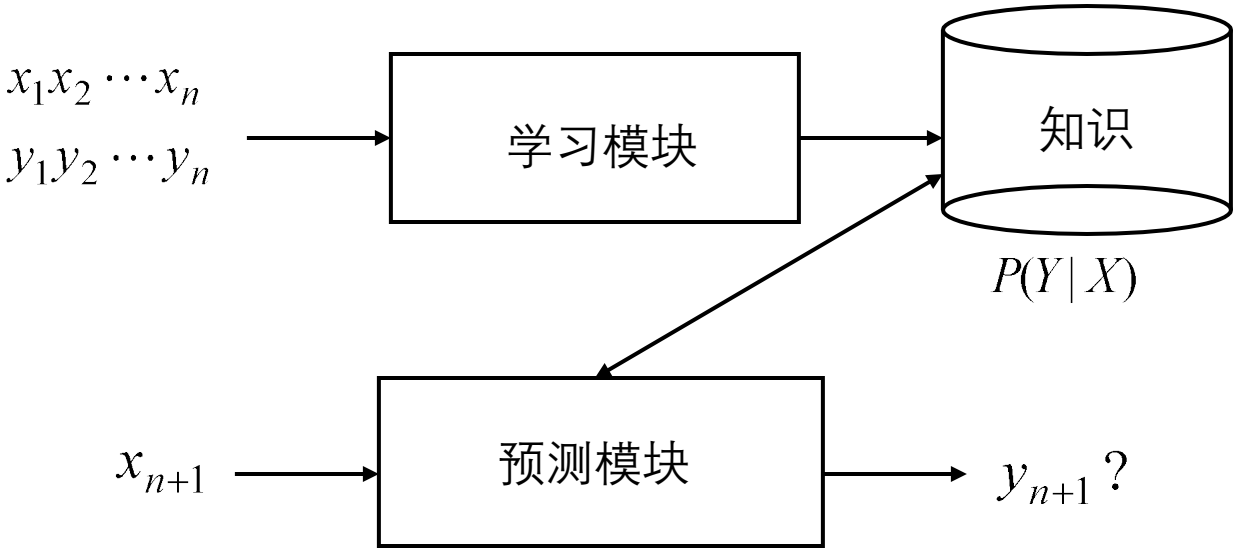
什么是无监督学习？（unsupervised learning）

- 无监督学习，无指导学习
- 解释 1
  - 有监督：涉及人力的介入
  - 无监督：不牵扯人力
- 解释 2
  - 给定一系列数据: $x_1, x_2, \ldots, x_N$
  - 有监督：期望的输出同样给出 $y_1, y_2, \ldots, y_N$
  - 无监督：没有期望输出

什么是无监督学习？

- 解释 3
  - 有监督：学习的知识关注条件分布 $P(Y|X)$
    - X = 样例（用其特征来表示），Y = 类别
  - 无监督：学习的知识关注联合分布 $P(X)$,
    - X：$X_1, X_2, \ldots, X_n$
  - 半监督学习：通过一些（少量）有标注的数据和很多无标注的数据学习条件分布 $P(Y|X)$

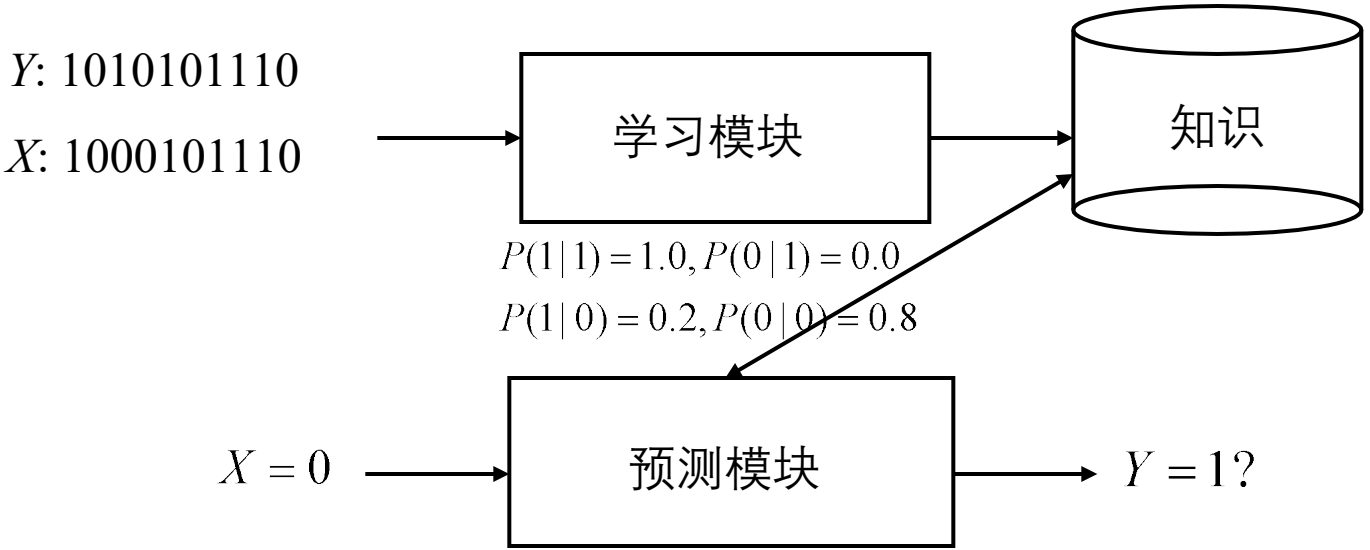
监督学习



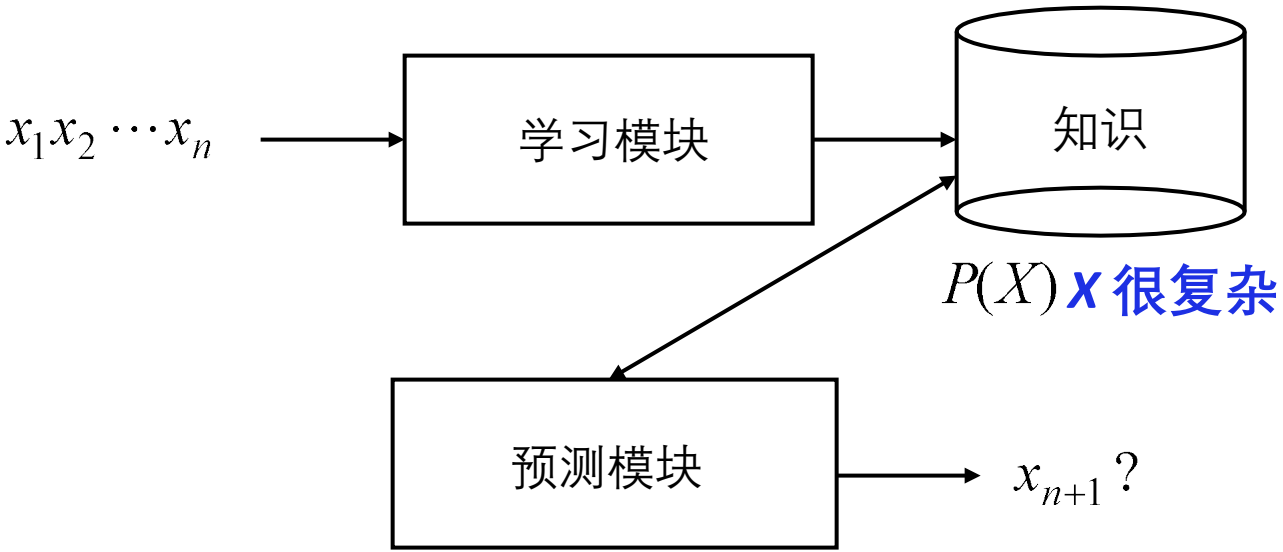
Y 简单

X 一般很复杂

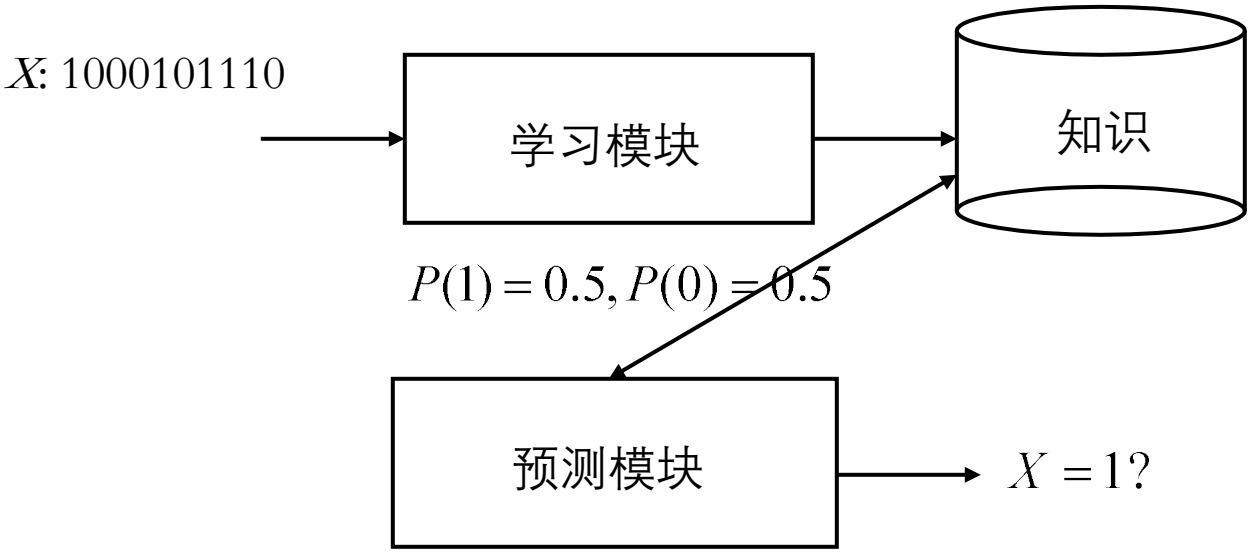
监督学习示例


Y: 1010101110
X: 1000101110
学习模块
知识
$P(1|1)=1.0, P(0|1)=0.0$
$P(1|0)=0.2, P(0|0)=0.8$
$X=0$
预测模块
$Y=1?$

无监督学习



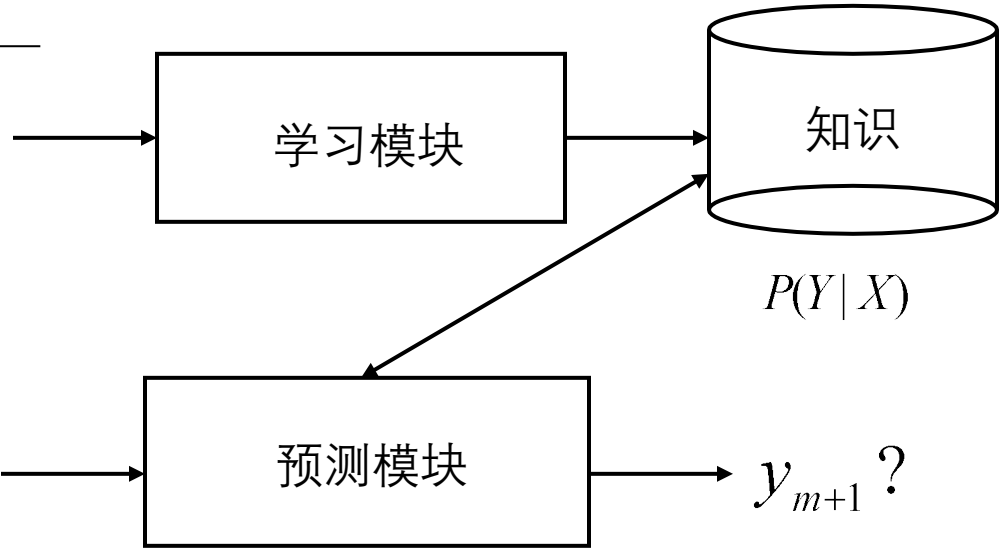
无监督学习示例



半监督学习

$$\frac{x_1 x_2 \cdots x_n}{y_1 y_2 \cdots y_n}$$
$$x_{n+1} x_{n+2} \cdots x_m$$

***n* 很小**
***m* 很大**



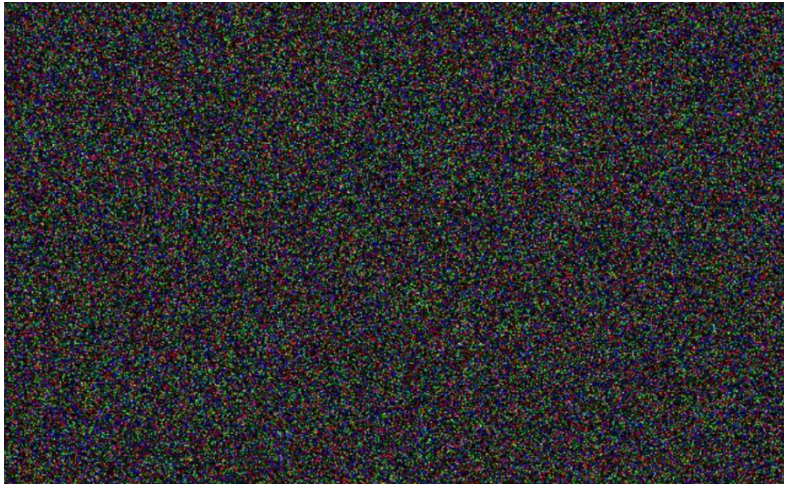
无标注数据的结构

- 构建模型找到输入的合理表示
  - 可以用来做决策、预测未知数据、将输入高效迁移到其他学习器等
- 发现数据的结构
  - 一篇学术论文包含题目、摘要.....
  - 半结构化网页中蕴含结构化信息
  - 图片中的像素不是随机生成的
  - 不同的用户兴趣组

数据的结构蕴含极为重要的信息

← 原始图片

随机交换像素点和 RGB 值之后 →



我们可以用无标注数据干什么？

- 数据聚类
  - 在没有预先定义的类别时将数据分为不同的组
- 降维
  - 减少所需要考虑的变量数量
- 离群点检测
  - 识别机器学习系统在训练中未发现的新数据/信号

    Identification of new or unknown data or signal that a Machine Learning system is not aware of during training
- 刻画数据密度

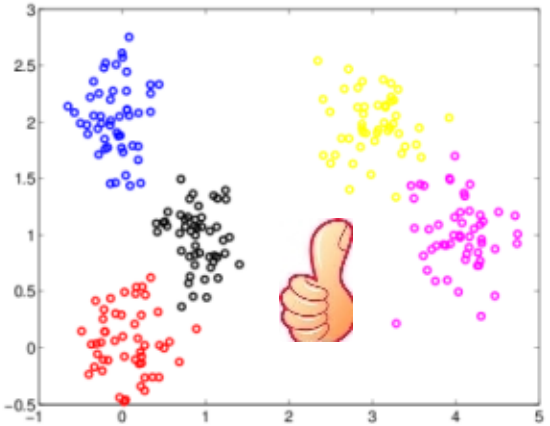
目录

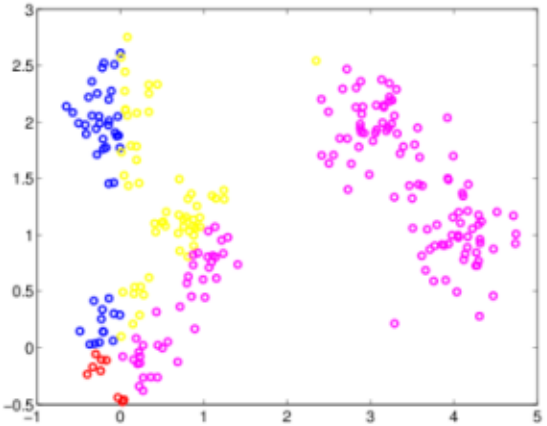
- 无监督学习介绍
- 聚类介绍
- 层次聚类
- K-means 聚类
- K-medoids 聚类

什么是聚类？

- 将相似的对象归入同一个“类”
  - “Birds of a feather flock together.” “物以类聚，人以群分”
  - 发现数据的结构
- 使得同一个类中的对象互相之间关联更强
  - 同一个类中的对象相似
  - 不同类中的对象有明显差异
- 核心问题：相似度定义
  - 簇/类内(intra-cluster)相似度
  - 簇/类间(inter-cluster)相似度

什么好的聚类？





- 类内距离小
- 类间距离大

聚类的类型 (1)

- 软聚类（soft clustering） vs. 硬聚类（hard clustering）
  - 软：同一个对象可以属于不同类
  - 硬：同一个对象只能属于一个类

示例（1）

e.g. 非动词间共现的数据

	eat	drink	make
wine	0	3	1
beer	0	5	1
bread	4	0	2
rice	4	0	0

硬聚类

示例（2）

e.g. 文档数据

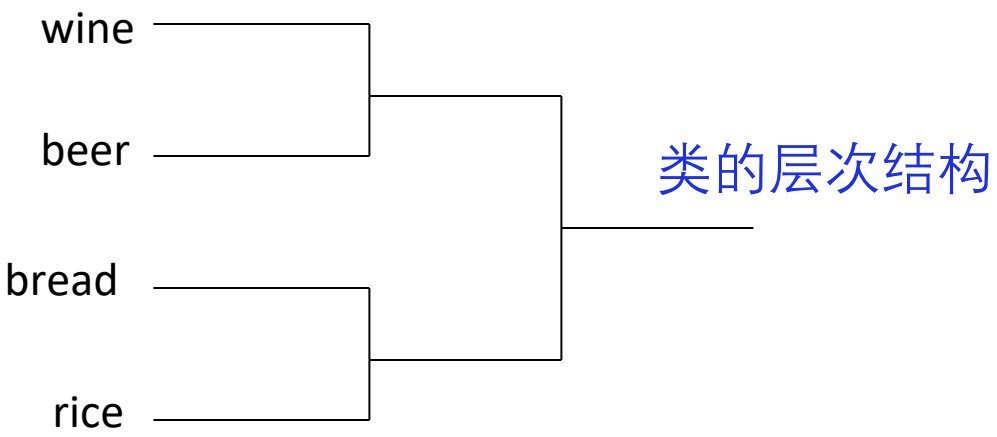
软聚类

	parsing	estimation	prediction	translation
document 1	0	3	2	0
document 2	0	5	1	0
document 3	1	1	1	0
document 4	4	0	0	3

聚类的类型 (2)

- 软聚类 vs. 硬聚类
  - 软：同一个对象可以属于不同类
  - 硬：同一个对象只能属于一个类
- 层次聚类 vs. 非层次聚类
  - 层次：
    - 一个所有类间的层次结构（tree）
  - 非层次：
    - 平的，只有一层

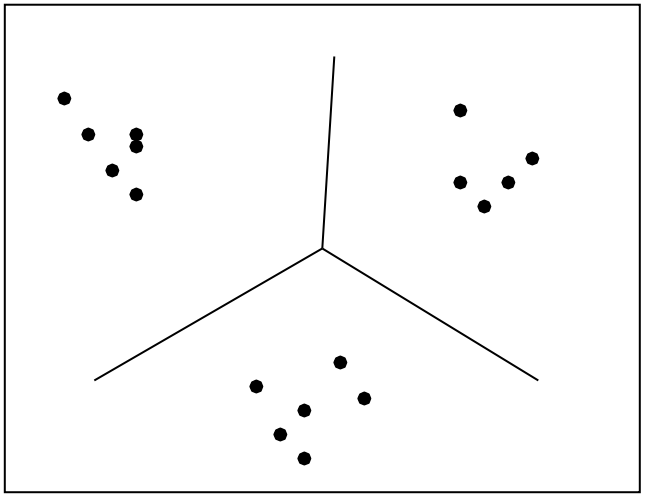
示例（3）



示例 (4)

向量空间的数据点

非层次聚类



应用

- 生物学
  - 将同源序列分组到基因家族中
  - 基因数据的相似度往往在聚类中被用于预测种群结构
- 图像处理 e.g. 自动相册
- 经济 – 尤其是市场商务智能
  - 找到不同的顾客群体，e.g. 保险
- WWW
  - 文档/事件 聚类，e.g. 每周新闻摘要
  - WEB日志分析，e.g. 找到相似的用户
- ……

数据聚类需要什么？

- 无标注数据
- 对象间的 距离 或 相似度度量
- （可选）类间的距离或相似度度量
- 聚类算法
  - 层次聚类
  - K-means、K-mediods
  - ……

数据

- 向量 $x \in D_1 \times D_2 \cdots \times D_N$
- 类型
  - 实数值 Real: D=R
  - 二值 Binary: $D = \{v_1, v_2\}$ e.g., {Female, Male}
  - 非数值 Nominal: $D = \{v_1, v_2, ..., v_M\}$ e.g., {Mon, Tue, Wed, Thu, Fri, Sat, Sun}
  - 有序值 Ordinal: $D = R$ or $D = \{v_1, v_2, ..., v_M\}$
    - 用于顺序非常重要的场景 e.g., 排名

相似度度量 (1)

- 相似度 = (距离)$^{-1}$
- 实数值数据
  - 内积
  - 余弦相似度
  - 基于核
  - …
- 回顾基于实例学习中的距离度量
  - Minkowski 距离
    - Manhattan 距离、Euclidean 距离、Chebyshev 距离
  - ……

相似度度量 (2)

- 非数值
  - E.g. "Boston", "LA", "Pittsburgh"
  - 或 “男”, “女",
  - 或 “弥散”, “球形”, “螺旋”, “风车"

相似度 (2)

- 非数值
  - E.g. "Boston", "LA", "Pittsburgh",
  - 或 “男”, “女",
  - 或 “弥散”, “球形”, “螺旋”, “风车"
  - 二值
    - If $x_i = x_j$, then $sim(x_i, x_j) = 1$, else $sim(x_i, x_j) = 0$
  - 用对应的语义属性
    - E.g. $Sim(Boston, LA) = \alpha dist(Boston, LA)^{-1}$,
    - $Sim(Boston, LA) = \alpha(|size(Boston) - size(LA)|) / Max(size(cities))$
  - 用相似度矩阵

相似度度量 (2): 相似度矩阵

相似度度量 (2): 相似度矩阵

	Tiny	Little	Small	Medium	Large	Huge
Tiny	1.0	0.8	0.7	0.5	0.2	0.0
Little		1.0	0.9	0.7	0.3	0.1
Small			1.0	0.7	0.3	0.2
Medium				1.0	0.5	0.3
Large					1.0	0.8
Huge						1.0

- 对角线一定是 1.0
- 不需要线性（插值）假设
- 必须满足传递性

相似度度量 (3)

- 有序值
  - E.g. “小” , “中”, “大”, “特大"
  - 归一化成 [0,1] 间的实数值：
    - max(v)=1, min(v)=0, 其他进行插值
    - E.g. “小”=0, “中"=0.33, etc.
  - 然后就可以使用实数值变量的相似度度量
  - 可以用相似度矩阵

目录

- 无监督学习介绍
- 聚类介绍
- 层次聚类 （待续）
  - 凝聚式层次聚类
  - 分裂式层次聚类
- K-means 聚类
- K-medoids 聚类