

1.2 机器学习任务类型

CSDN学院
2017年11月



► 机器学习

- http://en.wikipedia.org/wiki/Machine_learning :



机器学习是人工智能的一个分支，主要关于构造和研究可以从数据中学习的系统。

▶ 数据

- 数据通常以二维数据表形式给出
 - 每一行：一个样本
 - 每一列：一个属性 / 特征
- 例：纽约公寓租赁数据
 - 训练数据：49352个样本，14个输入变量、1个输出变量

| | bathrooms | bedrooms | building_id | created | description | display_address | features | interest_level | latitude | listing_id | longitude | manager_id | photos | price | street_address |
|--------|-----------|----------|----------------------------------|---------------------|---|---------------------|---|----------------|----------|------------|-----------|----------------------------------|---|-------|-------------------------|
| 10 | 1.5 | 3 | 53a5b119ba8f7b61d4e010512e0dfc85 | 2016-06-24 07:54:24 | A Brand New 3 Bedroom 1.5 bath ApartmentEnjoy ... | Metropolitan Avenue | [] | medium | 40.7145 | 7211212 | -73.9425 | 5ba989232d0489da1b5f2c45f6688adc | [https://photos.renthop.com/2/7211212_1ed4542e... | 3000 | 792 Metropolitan Avenue |
| 10000 | 1.0 | 2 | c5c8a357cba207596b04d1afd1e4f130 | 2016-06-12 12:19:27 | | Columbus Avenue | [Doorman, Elevator, Fitness Center, Cats Allow... | low | 40.7947 | 7150865 | -73.9667 | 7533621a882f71e25173b27e3139d83d | [https://photos.renthop.com/2/7150865_be3306c5... | 5465 | 808 Columbus Avenue |
| 100004 | 1.0 | 1 | c3ba40552e2120b0acfc3cb5730bb2aa | 2016-04-17 | Top Top West Village location, | W 13 Street | [Laundry Ir Building, Dishwashe | high | 40.7388 | 6887163 | -74.0018 | d9039c43983f6e564b1482b273bd7b01 | [https://photos.renthop.com/2/6887163_de85c427... | 2850 | 241 W 13 |

y：目标 / 响应 / 自变量

► 机器学习任务类型

- 监督学习(Supervised Learning)
 - 分类 (Classification)
 - 回归 (Regression)
 - 排序 (Ranking)
- 非监督学习 (Unsupervised Learning)
 - 聚类 (Clustering)
 - 降维 (Dimensionality Reduction)
 - 概率密度估计 (density estimation)
- 增强学习 (Reinforcement Learning)
- 半监督学习 (Semi-supervised Learning)
- 迁移学习 (Transfer Learning)
- ...

- 监督学习：学习到一个 $\mathbf{x} \rightarrow y$ 的映射 f ，从而对新输入的 \mathbf{x} 进行预测 $f(\mathbf{x})$
 - 训练数据包含要预测的标签 y （标签在训练数据中是可见变量）

训练数据集

$$\mathcal{D} = \{ \mathbf{x}_i, y_i \}_{i=1}^N$$

训练样本数目

第 i 个训练样本的输入，
亦被称为特征、属性或
协变量

第 i 个训练样本的输出，
亦被称为响应，如类别标签、
序号或数值

► 回归

- 在监督学习任务中，若输出 $y \in \mathbb{R}$ 为连续值，则我们称之为一个回归（Regression）任务。
- 例：预测二手车的价格
 - 输入/协变量(covariate) \mathbf{x} ：车辆属性
 - 输出 y ：车辆价格

► 回归

- 回归：根据训练数据 $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ 学习一个从输入 \mathbf{x} 到输出 y 的映射 f ，然后对新的测试数据 \mathbf{x} ，用学习到的映射对其进行预测： $\hat{y} = f(\mathbf{x})$
 ← 预测：带帽表示预测
 - 回归器： $y = f(\mathbf{x} | \theta)$
 - $f()$ ：模型，如线性回归模型 $y = f(\mathbf{x} | \mathbf{w}) = \mathbf{w}^T \mathbf{x}$
 - θ ：参数，如在上述线性回归模型中， \mathbf{w} 为模型参数
 - 学习的目标：在训练集上预测值与真值之间的差异
 - 损失函数：度量模型预测值与真值之间的差异，例如

$$L(f(\mathbf{x}), y) = \frac{1}{2} (f(\mathbf{x}) - y)^2$$
 - 则目标函数为 $J(\theta) = \frac{1}{N} \sum_{i=1}^N L(f(\mathbf{x}_i | \theta), y_i)$ 经验风险最小化
- 但仅仅经验风险最小化还不够，还需考虑模型复杂度

► 分类

- 在监督学习任务中，若输出 y 为离散值，我们称之为分类。
- 标签空间： $\mathcal{Y} = \{1, 2, \dots, C\}$
- 例：信用评分
 - 输入 x ：客户的存款（savings）和收入（income）
 - 输出 y ：客户的风险等级（risk）
 - 高风险、低风险

► 分类

- 分类：学习从输入 \mathbf{x} 到输出 y 的映射 f ：

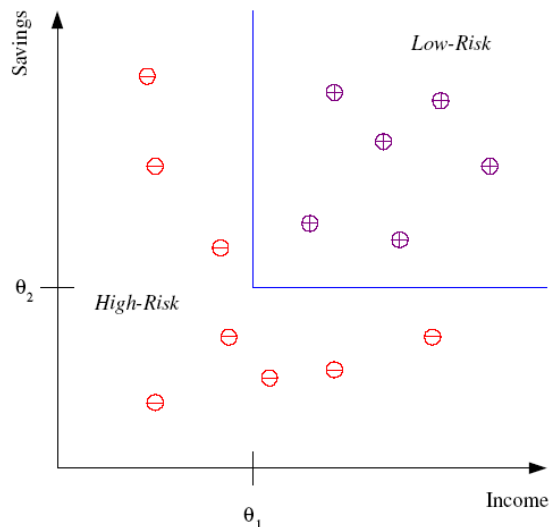
$$\hat{y} = f(\mathbf{x}) = \arg \max_c p(y = c | \mathbf{x}, \mathcal{D})$$

- 可形式化为函数逼近，学习的目标是给定训练数据集，估计函数 f ，然后进行预测 $\hat{y} = f(\mathbf{x})$
- 主要目标是对新的输入进行预测：推广性/泛化能力
([generalization](#))

► 例：分类

- 信用评分

- 给定样本 { (savings, income, risk) }
- 找到预测 “规则” : $\text{risk} = f(\text{savings}, \text{income})$



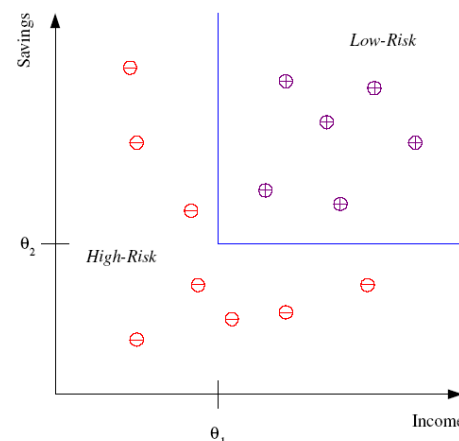
决策树:

Rule: IF $\text{income} > \theta_1$ AND $\text{savings} > \theta_2$
THEN low-risk ELSE high-risk

► 例：分类

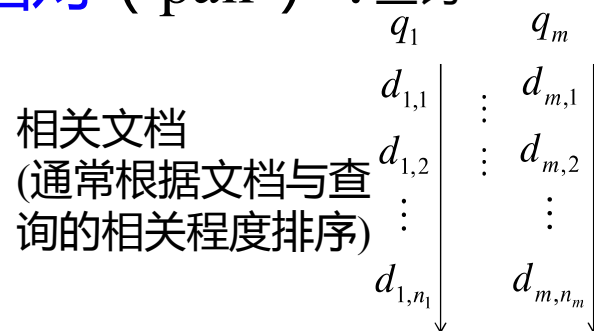
- 需要预测概率： $f(\mathbf{x}, c) = p(y = c | \mathbf{x}, \mathcal{D}, M)$
 - 如靠近分类的边界的样本（蓝色所示样本）
 - 有歧义时返回概率/可能性 $p(y = c | \mathbf{x}, \mathcal{D})$ ，即给定训练数据 \mathcal{D} 和输入 \mathbf{x} 的情况下，输出为 c 的条件概率
- 预测：最大后验估计（Maximum a Posteriori, MAP）

$$\hat{y} = \arg \max_c p(y = c | \mathbf{x}, \mathcal{D})$$



► 排序 (Rank)

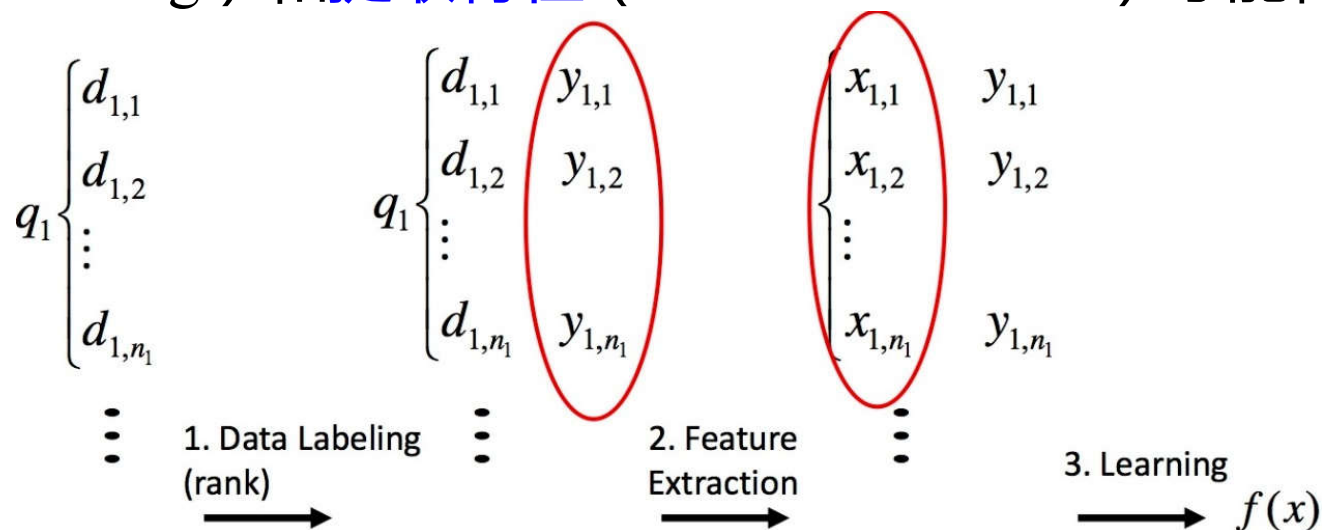
- 排序学习是推荐、搜索、广告的核心方法。
- 以信息检索为例，训练时我们给定文档集合 $D = \{d_1, d_2, \dots, d_N\}$ 和查询 - 文档对 (pair) : 查询



- 排序学习根据训练学习一个排序模型 $f(q, d)$ ，然后利用该模型对新的查询 q_{m+1} ，给出每个文档的排序: $f(q_{m+1}, d_1)$ 、...、 $f(q_{m+1}, d_{nm+1})$

► 排序 (Rank)

- 和一般监督学习直接给定训练数据 $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ 不同，排序学习中需要首先根据查询 q 及其文档集合进行标注 (data labeling) 和提取特征 (feature extraction) 才能得到 $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$



► 非监督学习

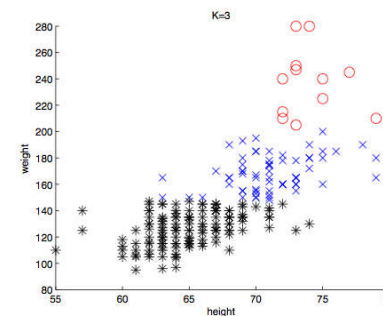
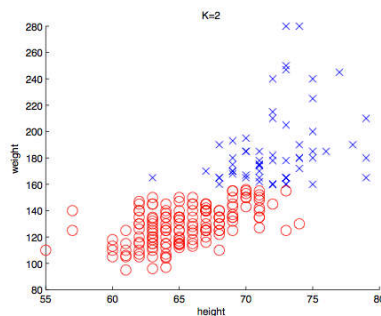
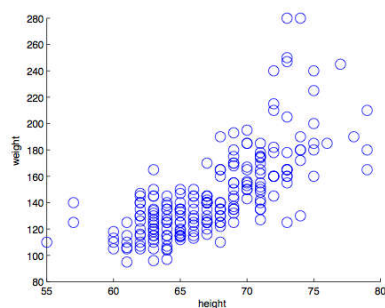
- 非监督学习：发现数据中的“有意义的模式”，亦被称为知识发现
 - 训练数据不包含标签
 - 标签在训练数据中为隐含变量

$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$$

► 聚类

- 例：人的“类型”

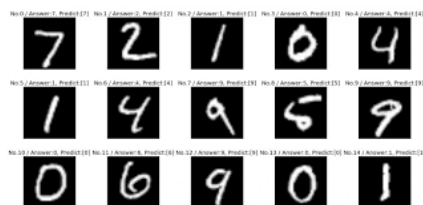
$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$$



- 分多少类？模型选择 $K^* = \arg \max_K p(K | \mathcal{D})$
- 某个样本属于哪个类？ $z_i \in \{1, \dots, K\}$ 表示第*i*个数据点所属类别，为隐含变量 $z_i^* = \arg \max_k p(z_i = k | \mathbf{x}_i, \mathcal{D})$

► 降维

- 样本 \mathbf{x} 通常有多维特征，有些特征之间会相关而存在冗余。
 - 如图像中相邻像素的值通常相同或差异很小

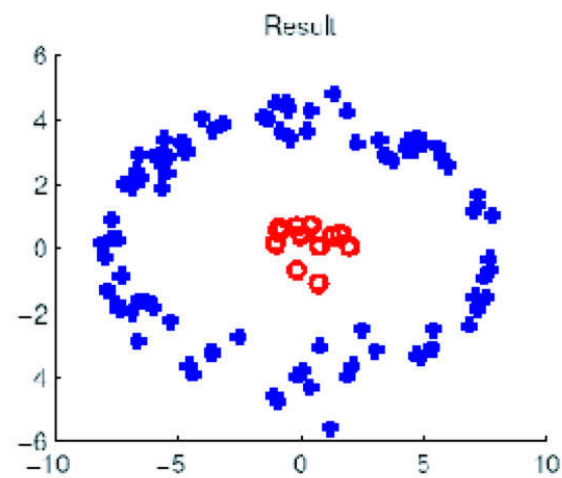
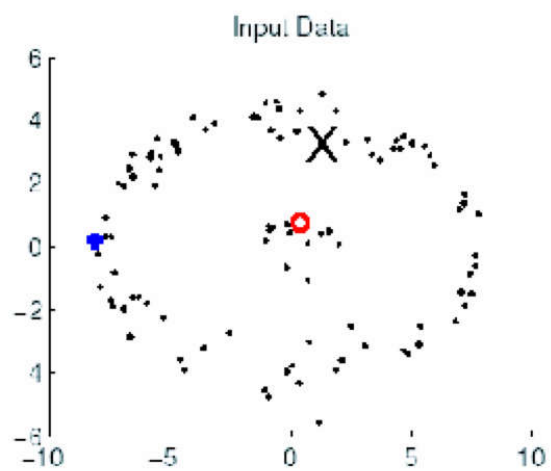


- 降维是一种将原高维空间中的数据点映射到低维度空间的技术。其本质是学习一个映射函数 $f: \mathbf{x} \rightarrow \mathbf{x}'$ ，其中 \mathbf{x} 是原始数据点的表达， \mathbf{x}' 是数据点映射后的低维向量表达。
- 在很多算法中，降维算法成为了数据预处理的一部分，如主成分分析（Principal Components Analysis, PCA）。

► 半监督学习 (Semisupervised Learning)

- 根据带标签数据+不带标签数据进行学习
- 监督学习+非监督学习 的组合
- 当标注数据 “昂贵” 时有用
 - 如：标注3D 姿态、蛋白质功能等等

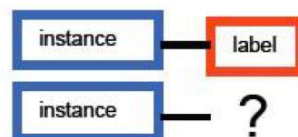
► 半监督学习



► 其他类型的学习任务



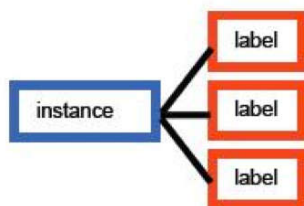
standard supervised learning



semi-supervised learning

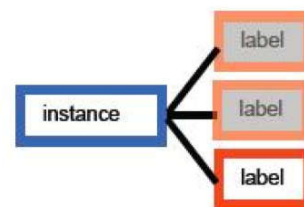


unsupervised learning



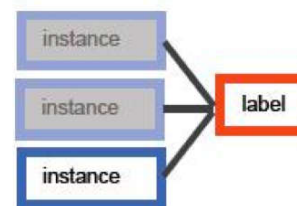
multi-label learning

all correct labels



ambiguous-label learning

only 1 correct label



multi-instance learning

at least 1 instance has label

- 增强学习：从行为的反馈（奖励或惩罚）中学习
 - 设计一个回报函数（reward function），如果learning agent（如机器人、回棋AI程序）在决定一步后，获得了较好的结果，那么我们给agent一些回报（比如回报函数结果为正），得到较差的结果，那么回报函数为负
 - 增强学习的任务：找到一条回报值最大的路径

► 小结：机器学习任务类型

- 监督学习(Supervised Learning)
 - 分类 (Classification)
 - 回归 (Regression)
 - 排序 (Ranking)
- 非监督学习 (Unsupervised Learning)
 - 聚类 (Clustering)
 - 降维 (Dimensionality Reduction)
 - 概率密度估计 (density estimation)
- 增强学习 (Reinforcement Learning)
- 半监督学习 (Semi-supervised Learning)
- 迁移学习 (Transfer Learning)
- ...