人工智能导论

主讲：王 博

人工智能与自动化学院

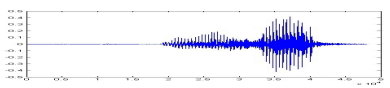
第2章 搜索与机器学习

第2节 机器学习简介

机器学习简介 – 目录

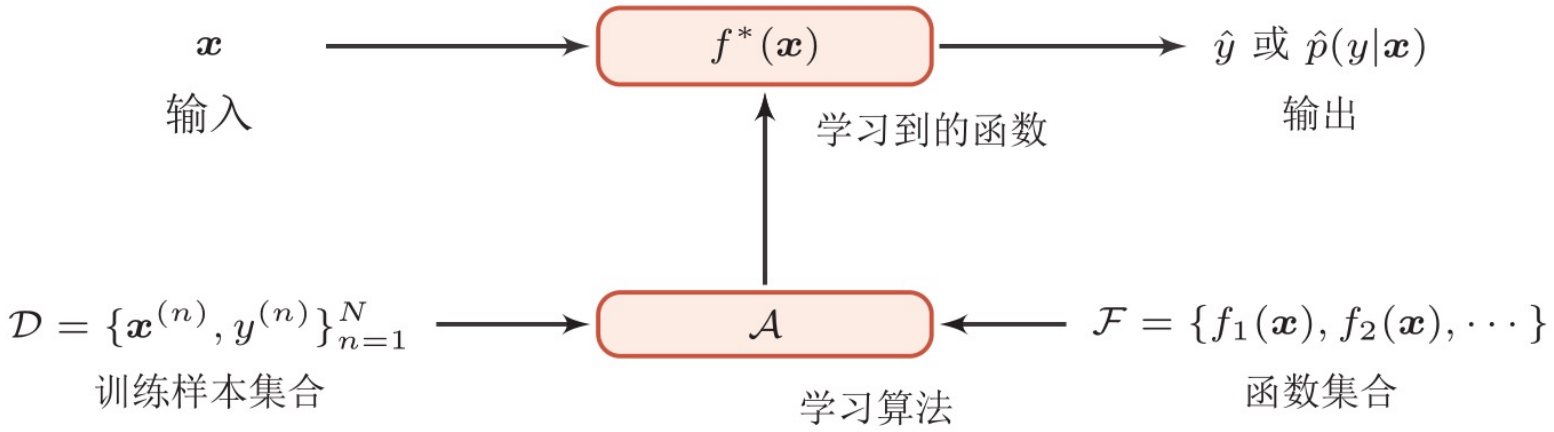
- 机器学习原理与概念
- 机器学习分类
- 机器学习关键思想
- 机器学习与人工智能

机器学习 ≈ 构建一个映射函数

- 语音识别
  - $f($  $)=$ “你好”
- 图像识别
  - $f($  $)=$ “猫”
- 围棋
  - $f($  $)=$ “5-5”
- 对话系统
  - $f($ “你好” $)=$ “今天天气真不错”

什么是机器学习？

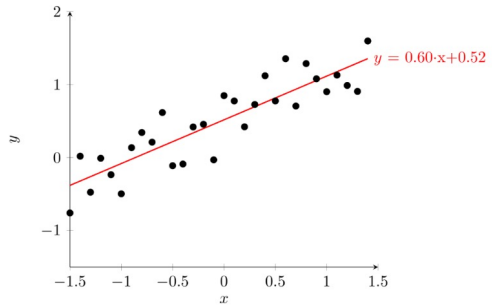
- 机器学习：通过算法使得机器能从大量数据中学习规律从而对新的样本做决策。
- 规律：决策（预测）函数



挑西瓜为例

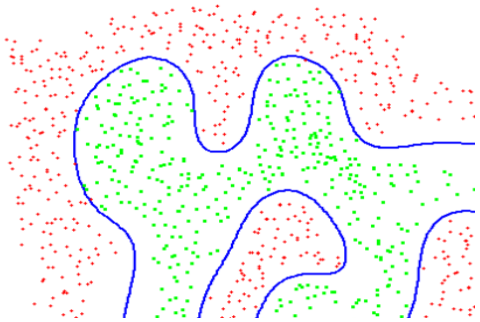
常见的机器学习问题

- 回归



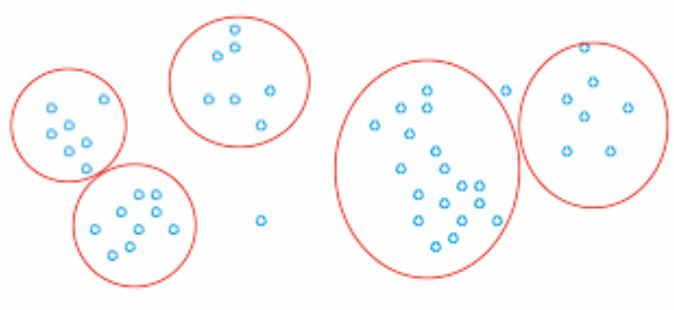
- 给西瓜打分

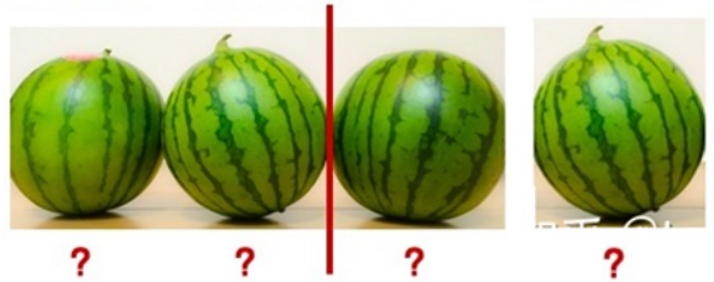
- 分类



- 看西瓜好坏

- 聚类





- 让相似西瓜抱团儿

机器学习基本概念

- 名词解释

- 特征(Feature)：西瓜的颜色，大小，形状，产地，品牌等
- 标签(Label)： 连续值，西瓜的甜度、水分、成熟度的综合打分；离散值，西瓜的“好”“坏”标签

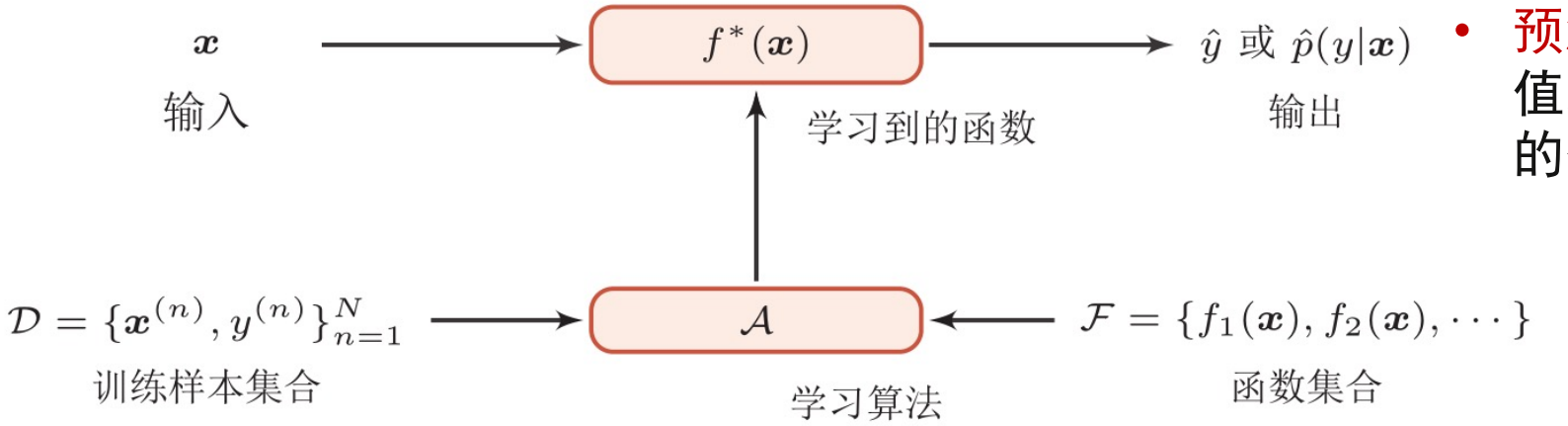
- 我们通常用一个 D 维向量$\boldsymbol{x} = [x_1, x_2, \cdots, x_D]^{\mathrm{T}}$表示一个西瓜的所有特征构成的向量，称为特征向量(Feature Vector)，其中每一维表示一个特征
  - 标签通常用标量y来表示

- 我们可以将一个标记好特征以及标签的西瓜看作一个样本(Sample)，也经常称为示例(Instance)

- 数据集(Data Set)：一组样本构成的集合。一般将数据集分为两部分:
- 训练集(Training Set)：用来训练模型的样本（训练样本）的集合
- 测试集(Test Set)：用来检验模型好坏的样本（测试样本）的集合

机器学习基本概念

- 机器学习的内涵
  - 我们希望让计算机从一个函数集合$F = \{f1(\boldsymbol{x}), f2(\boldsymbol{x}), \cdots\}$中自动寻找一个“最优”的函数$f^*(\boldsymbol{x})$来近似每个样本的特征向量$\boldsymbol{x}$和标签$y$之间的真实映射关系



- 预测标签的值或者标签的条件概率

- 独立同分布 (IID)
样本独立地从相同的数据分布$p(\boldsymbol{x}, y)$中抽取

- 寻找最优函数$f^*(\boldsymbol{x})$是机器学习的关键任务
  - 通过学习算法(Learning Algorithm)$\mathcal{A}$来完成
- 这个寻找过程通常称为学习(Learning)或训练(Training)

机器学习的三要素

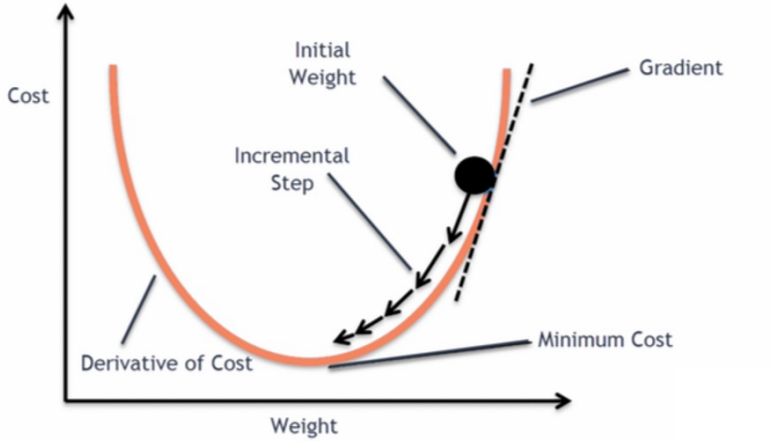
- 模型
  - 线性方法

    $f(\boldsymbol{x}, \theta) = \mathrm{w}^T \boldsymbol{x} + b$
  - 广义线性方法

    $f(\boldsymbol{x}, \theta) = \mathrm{w}^T \phi(\boldsymbol{x}) + b$
  - 解决学什么
- 学习准则
  - 期望风险

    $\mathcal{R}(f)$

    $= \mathbb{E}_{(\boldsymbol{x},y)\sim p_r(\boldsymbol{x},y)}[\mathcal{L}(f(\boldsymbol{x}), y)]$
  - 解决学成什么样
- 优化算法
  - 梯度下降



  - 解决怎么学

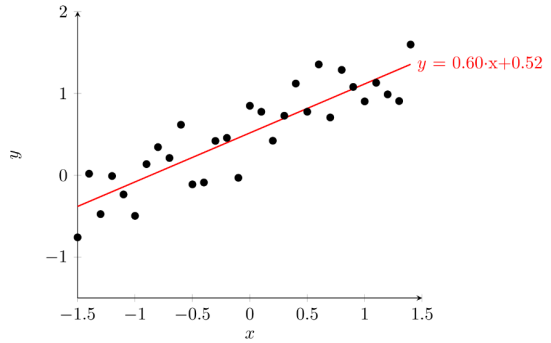
机器学习的三要素

- 模型

- 输入空间 $\mathcal{X}$ 和输出空间 $\mathcal{Y}$ 构成了一个样本空间
- 样本空间中的样本 $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$， $\boldsymbol{x}$ 和 y 之间的关系可以描述为：
  - 未知的真实映射函数 $y = g(\boldsymbol{x})$，或
  - 真实条件概率分布 $p_r(y|\boldsymbol{x})$
- 模型是 $g(\boldsymbol{x})$ 或 $p_r(y|\boldsymbol{x})$ 的近似
- 我们不知道 $g(\boldsymbol{x})$ 或 $p_r(y|\boldsymbol{x})$ 的具体形式，因而只能根据经验来假设一个函数集合 $\mathcal{F}$，称为假设空间(Hypothesis Space)
  - 选择一个理想的假设(Hypothesis) $f^* \in \mathcal{F}$
- 假设空间 $\mathcal{F}$ 通常为一个参数化的函数族

$$\mathcal{F} = \{f(\boldsymbol{x}; \theta) | \theta \in \mathbb{R}^D\}$$

其中 $f(\boldsymbol{x}; \theta)$ 是参数为 θ 的函数，也称为模型(Model)，D 为参数的数量



- 以线性回归(Linear Regression)为例

模型：

$$f(\mathrm{x}, \theta) = \mathrm{w}^T \mathrm{x} + b$$

机器学习的三要素

- 学习准则
  - 训练集 $\mathcal{D}=\{(\boldsymbol{x}^{(n)},y^{(n)})\}_{n=1}^{N}$ 应当由 N 个独立同分布(Independent and Identically Distributed，IID)的样本组成
    - 样本分布$p_r(\boldsymbol{x},y)$必须固定（可以未知）
    - 如果$p_r(\boldsymbol{x},y)$本身可变，无法通过这些数据学习

- 一个好的模型 $f(\boldsymbol{x},\theta^*)$应该在所有$(\boldsymbol{x},y)$的可能取值上都与真实映射函数$y=g(\boldsymbol{x})$ 一致
  - 衡量 $f(x,\theta^*)$ 与 y 分布相似性的常用方法：KL散度或交叉熵
  - 模型$f(\boldsymbol{x},\theta^*)$的好坏可以通过期望风险(Expected Risk) $\mathcal{R}(\theta)$来衡量

$$\mathcal{R}(f)=\mathbb{E}_{(\boldsymbol{x},y)\sim p_r(\boldsymbol{x},y)}[\mathcal{L}(f(\boldsymbol{x}),y)]$$

  - $\mathcal{L}(f(\boldsymbol{x}),y)$表示损失函数，用来量化两个变量间的差异

例：

$$\mathcal{L}(y,f(\boldsymbol{x};\theta))=\begin{cases}0 & \text{if } y=f(\boldsymbol{x};\theta)\\ 1 & \text{if } y\neq f(\boldsymbol{x};\theta)\end{cases}$$

$$=I(y\neq f(\boldsymbol{x};\theta)),$$

$$\mathcal{L}(y,f(\boldsymbol{x};\theta))=\frac{1}{2}\big(y-f(\boldsymbol{x};\theta)\big)^2.$$

机器学习的三要素

- 风险最小化准则
  - 期望风险未知，通过经验风险近似
    - 给定一个训练集$\mathcal{D}=\{(\boldsymbol{x}^{(n)},y^{(n)})\}_{n=1}^{N}$，我们可以计算的是经验风险(Empirical Risk)，即在训练集上 的平均损失：

$$\mathcal{R}_{\mathcal{D}}^{emp}(\theta)=\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(y^{(n)},f(\boldsymbol{x}^{(n)};\theta)).$$

    - 实践：寻找一个参数θ^*，使得经验风险函数最小化

$$\theta^*=\arg\min_{\theta}\mathcal{R}_{\mathcal{D}}^{emp}(\theta),$$

    - 称为经验风险最小化 (Empirical Risk Minimization，ERM)准则

- 经验风险最小化原则很容易导致模型在训练集上错误率很低，但是在未知数据上错误率很高
- 所谓的过拟合(Overfitting)

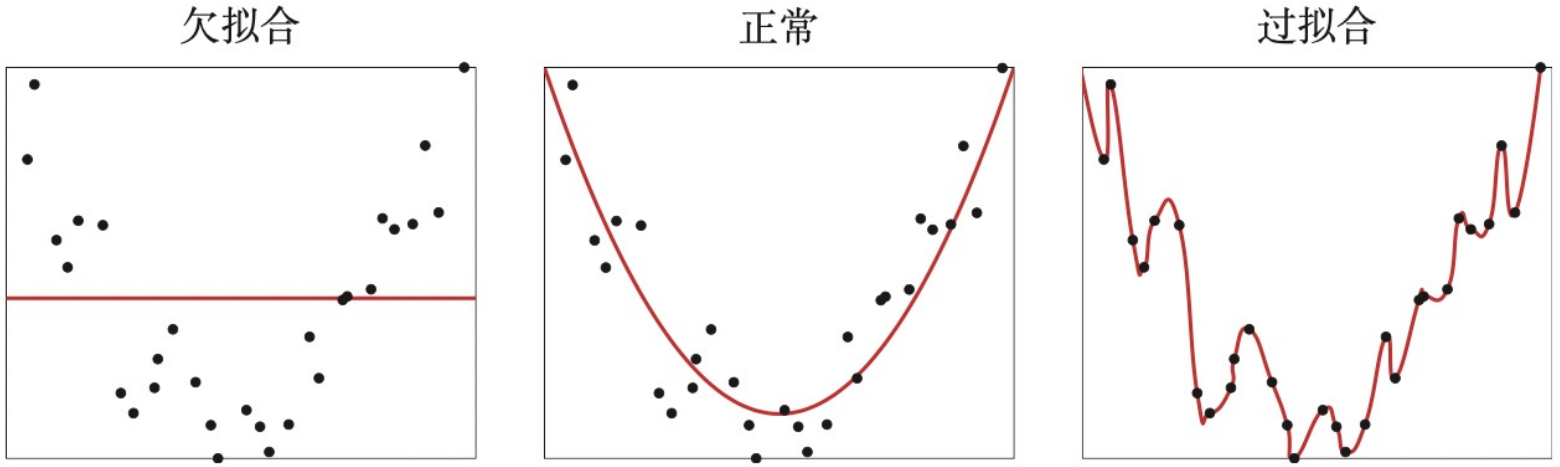
机器学习的三要素

- 过拟合与欠拟合

- 和过拟合相反的一个概念是欠拟合(Underfitting)
  - 模型不能很好拟合训练数据，训练集上错误率高
  - 模型能力不足造成

- 过拟合和欠拟合示例

定义 2.1 – 过拟合： 给定一个假设空间 $\mathcal{F}$，一个假设 f 属于 $\mathcal{F}$，如果存在其他的假设 f' 也属于 $\mathcal{F}$，使得在训练集上 f 的损失比 f' 的损失小，但在整个样本空间上 f' 的损失比 f 的损失小，那么就说假设 f 过度拟合训练数据 [Mitchell, 1997].

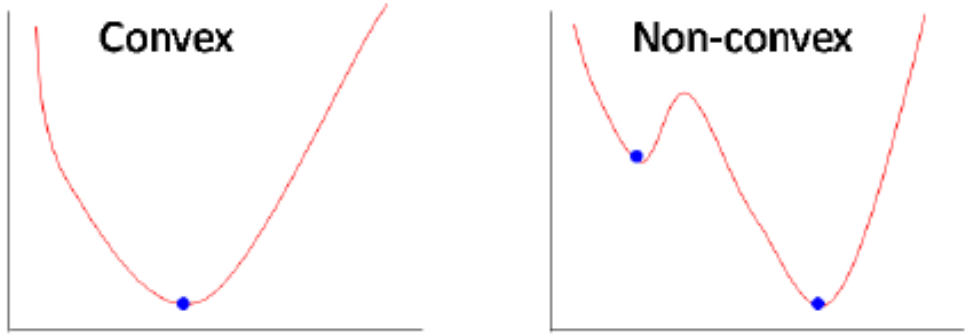


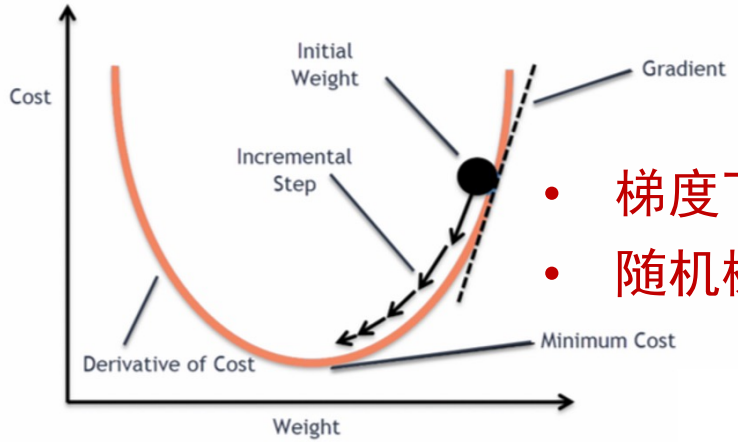
机器学习的三要素

- 优化算法
  - 机器学习问题转化成为一个最优化问题
    - 寻找参数θ^*使经验风险最小化

$$\theta^* = \arg\min_{\theta} \mathcal{R}_{\mathcal{D}}^{emp}(\theta),$$

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$





- 梯度下降法
- 随机梯度下降法

- 机器学习中的优化又可以分为参数优化和超参数优化
  - 参数：模型 $f(\boldsymbol{x};\theta)$ 中的参数 θ
  - 超参数(Hyper-Parameter)：用来定义模型结构或优化策略的参数

- 机器学习≠优化！
  - 例：最优化不考虑过拟合问题

机器学习简介 – 目录

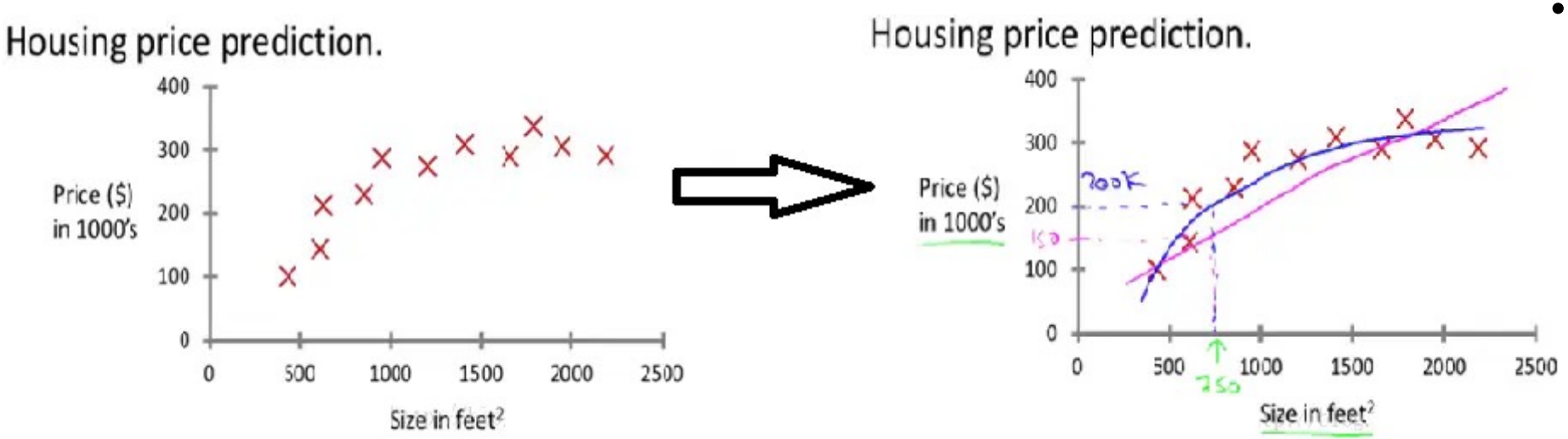
- 机器学习原理与概念
- 机器学习分类
- 机器学习关键思想
- 机器学习与人工智能

常见的机器学习类型

	监督学习	无监督学习	强化学习
训练样本	训练集 $\{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$	训练集 $\{\boldsymbol{x}^{n}\}_{n=1}^{N}$	智能体和环境交互的轨迹 τ 和累积奖励 G_τ
优化目标	$y = f(\boldsymbol{x})$ 或 $p(y\|\boldsymbol{x})$	$p(\boldsymbol{x})$ 或带隐变量 $\boldsymbol{z}$ 的 $p(\boldsymbol{x}\|\boldsymbol{z})$	期望总回报 $\mathbb{E}_\tau[G_\tau]$
学习准则	期望风险最小化 最大似然估计	最大似然估计 最小重构错误	策略评估 策略改进

监督学习(Supervised Learning)

- 已知输入和输出的情况下训练出一个模型，将输入映射到输出 $g: \mathcal{X} \to y$
- 通过学习标记的训练样本来构建预测模型，并依此模型推测新的实例
- 输出可以是一个连续的值（称为回归分析），或是预测一个分类标签（称作分类）



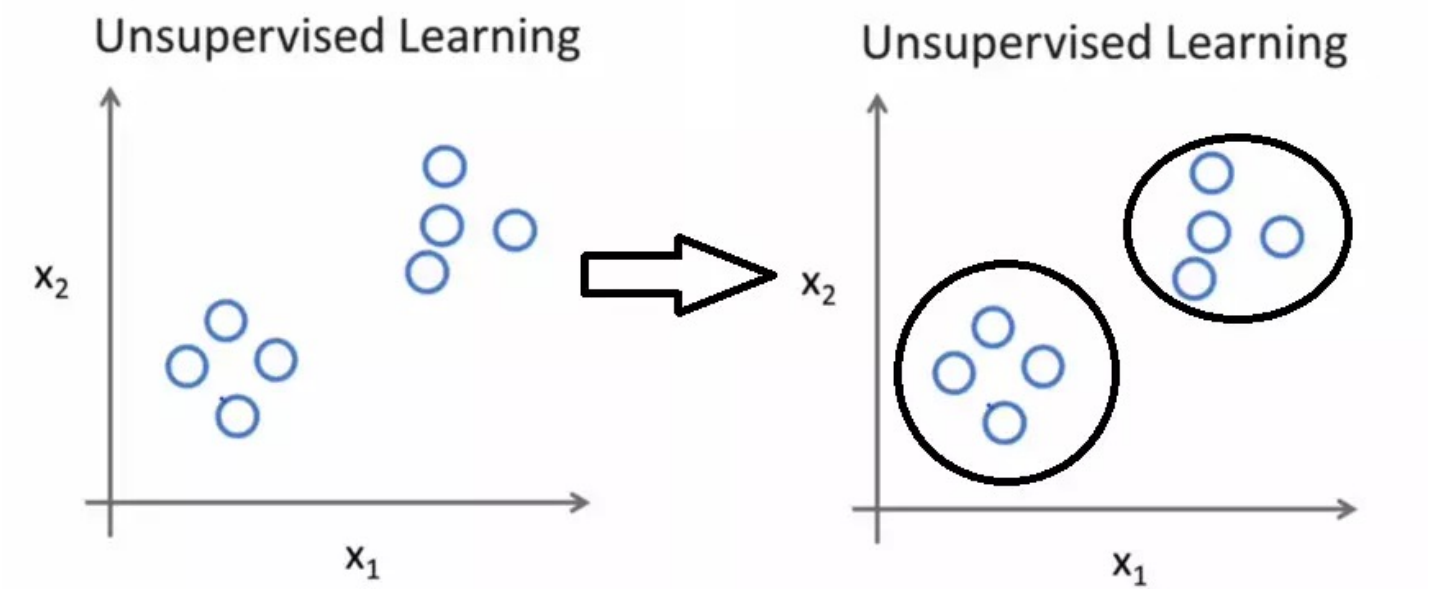
- 典型监督学习算法
  - 朴素贝叶斯
  - 决策树
  - 支持向量机
  - 线性回归
  - 神经网络
  - ……

- 主流监督学习算法

算法	类型	简介
朴素贝叶斯	分类	贝叶斯分类法是基于贝叶斯定定理的统计学分类方法。它通过预测一个给定的元组属于一个特定类的概率，来进行分类。朴素贝叶斯分类法假定一个属性值在给定类的影响独立于其他属性的 —— 类条件独立性。
决策树	分类	决策树是一种简单但广泛使用的分类器，它通过训练数据构建决策树，对未知的数据进行分类。
SVM	分类	支持向量机把分类问题转化为寻找分类平面的问题，并通过最大化分类边界点距离分类平面的距离来实现分类。
逻辑回归	分类	逻辑回归是用于处理因变量为分类变量的回归问题，常见的是二分类或二项分布问题，也可以处理多分类问题，它实际上是属于一种分类方法。
线性回归	回归	线性回归是处理回归任务最常用的算法之一。该算法的形式十分简单，它期望使用一个超平面拟合数据集（只有两个变量的时候就是一条直线）。
回归树	回归	回归树（决策树的一种）通过将数据集重复分割为不同的分支而实现分层学习，分割的标准是最大化每一次分离的信息增益。这种分支结构让回归树很自然地学习到非线性关系。
K邻近	分类+回归	通过搜索K个最相似的实例（邻居）的整个训练集并总结那些K个实例的输出变量，对新数据点进行预测。
Adaboosting	分类+回归	Adaboost目的就是从训练数据中学习一系列的弱分类器或基本分类器，然后将这些弱分类器组合成一个强分类器。
神经网络	分类+回归	它从信息处理角度对人脑神经元网络进行抽象， 建立某种简单模型，按不同的连接方式组成不同的网络。

无监督学习(Unsupervised Learning)

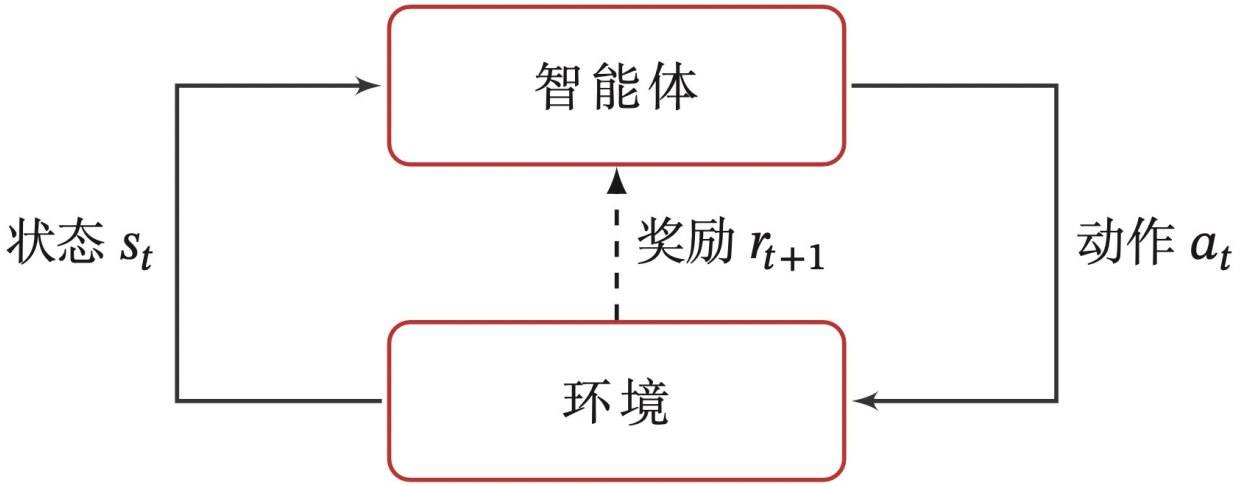
- 不给定事先标记过的训练示例，自动对输入的数据进行分析
- 不需要数据标注，对大数据分析很重要，但在实际应用中性能受限
- 包括聚类、降维等



- 典型无监督学习算法
  - 聚类：K-均值
  - 降维：PCA
  - 自编码器
  - ……

强化学习(Reinforcement Learning)

- 强化学习问题可以描述为一个智能体从与环境的交互(试错，Trial-and-Error)中不断学习以完成特定目标(比如取得最大奖励值)
- 强化学习就是智能体不断与环境进行交互，并根据经验调整其策略来最大化其长远的所有奖励的累积值



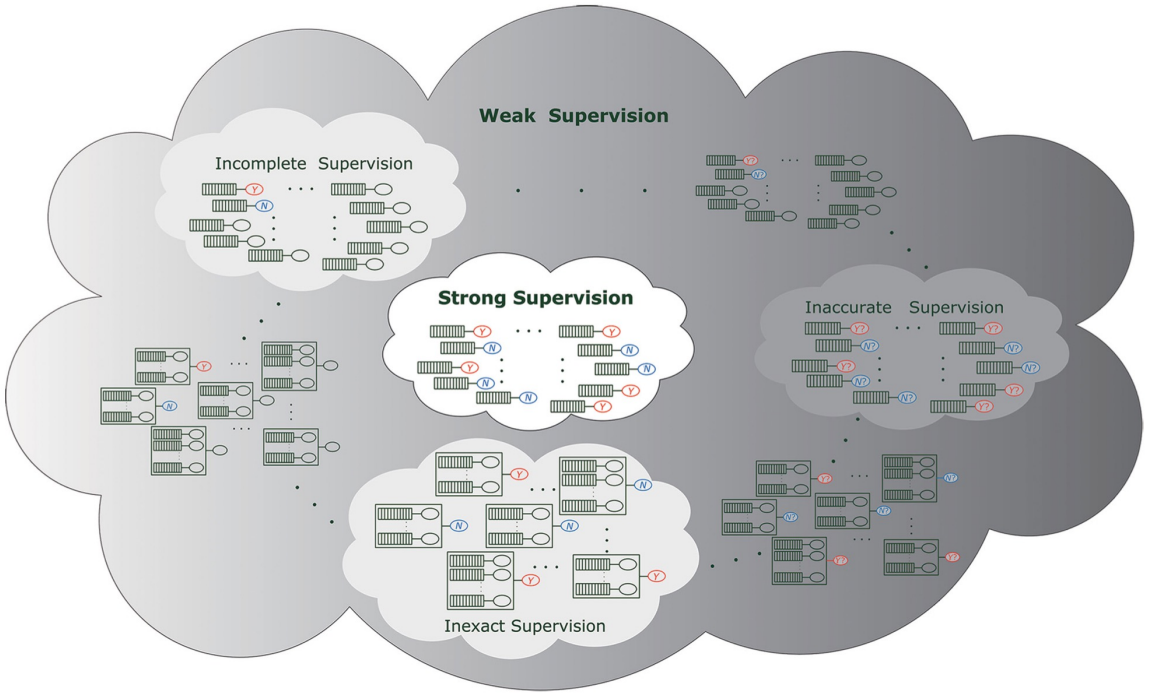
- 典型强化学习算法
  - 基于值函数
    - Q学习
    - 深度Q网络
  - 基于策略：
    - 策略梯度
    - 近端策略优化
  - ……

弱监督学习(Weakly Supervised Learning)

监督学习 → 数据标注成本太高；
无监督学习 → 学习过程困难、发展缓慢
↓
弱监督学习
数据标签允许是不完全的、不确切、不精确的

- 不完全监督 (Incomplete supervision)
- 不确切监督 (Inexact supervision)
- 不精确监督 (Inaccurate supervision)

Weakly supervised learning is an umbrella term covering a variety of studies that attempt to construct predictive models by learning with weak supervision.



[1] Zhi-Hua Zhou, A brief introduction to weakly supervised learning, *National Science Review*, Volume 5, Issue 1, January 2018, Pages 44–53,

弱监督学习：以图像分类为例

- incomplete

Image classification

It is easy to get a huge number of images from the Internet, but only a small subset of images can be annotated due to the human cost.

- inexact

Important target detection

Usually we only have image-level labels rather than object-level labels.

- inaccurate

Crowdsourcing data analysis

when the image annotator is careless or weary, or some images are difficult to categorize.

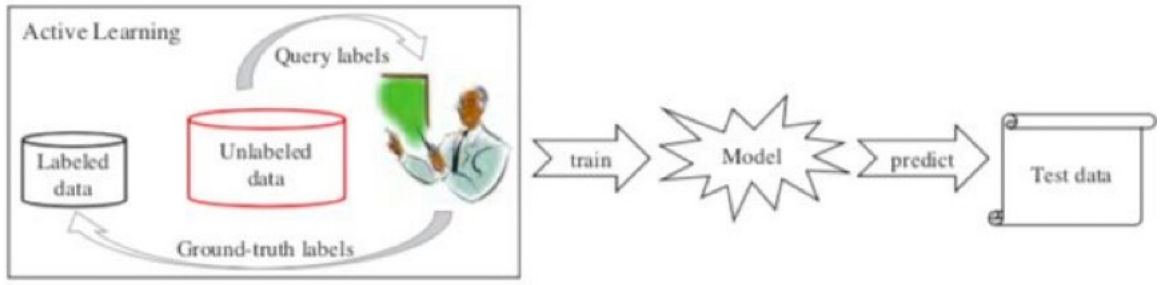
弱监督学习

- 为了解决不完全监督，我们可以考虑两种主要技术，主动学习和半监督学习。一种是有人类干预的，一种是没有人类干预的。
- 为了解决不确切监督，我们可以考虑多示例学习。
- 为了解决不精确监督，我们考虑带噪学习。

主动学习

- 在这些未标记数据中，主动学习(Active learning)尝试选择最有价值的未标记实例进行查询。最有价值指的是信息性和代表性。主动学习的目标是最小化查询的数量。

Active learning (Incomplete supervision)
With Human Intervention



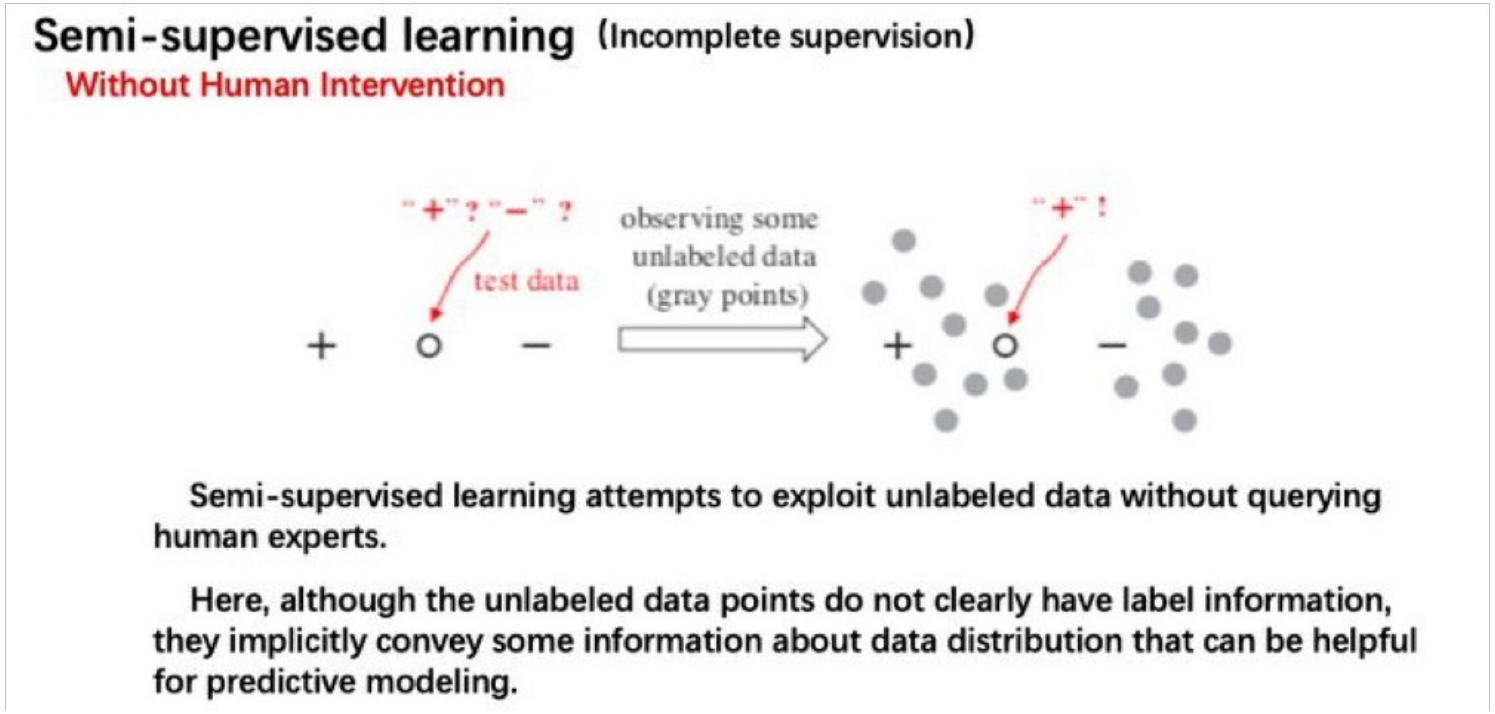
Active learning assumes that the ground-truth labels of unlabeled data can be queried from an oracle.

In these unlabeled data, active learning attempts to select the most valuable (informativeness and representativeness) unlabeled data to query.

Goal：minimize the number of queries

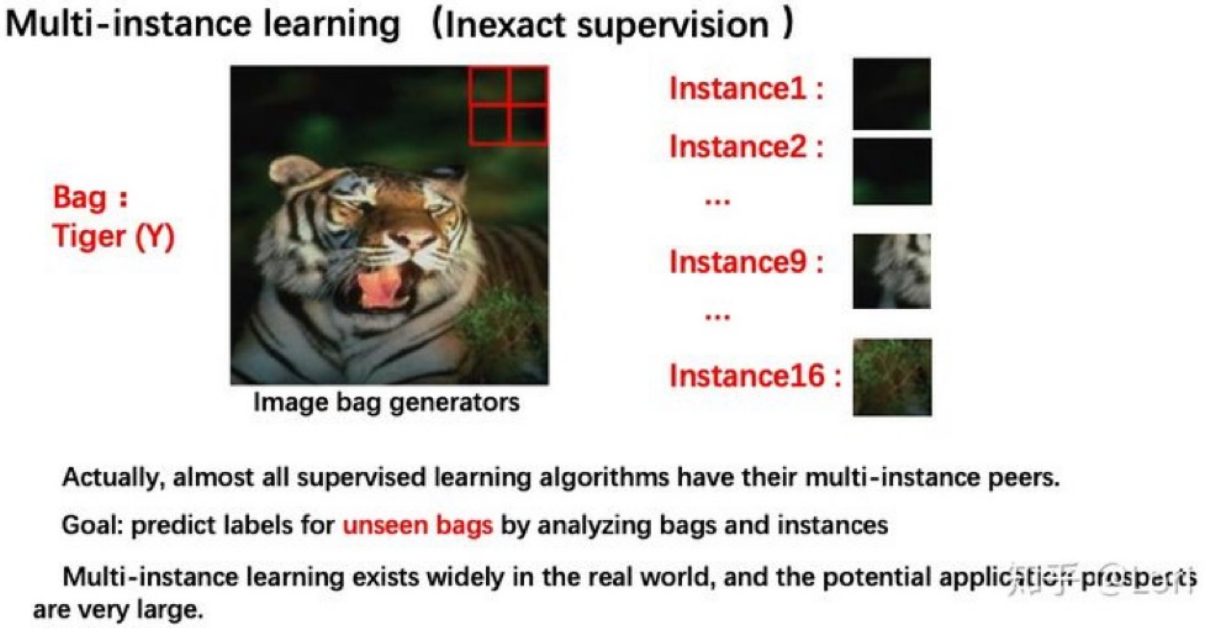
半监督学习

- 半监督学习(Semi-supervised learning)尝试在不查询人类专家的情况下利用未标注的数据

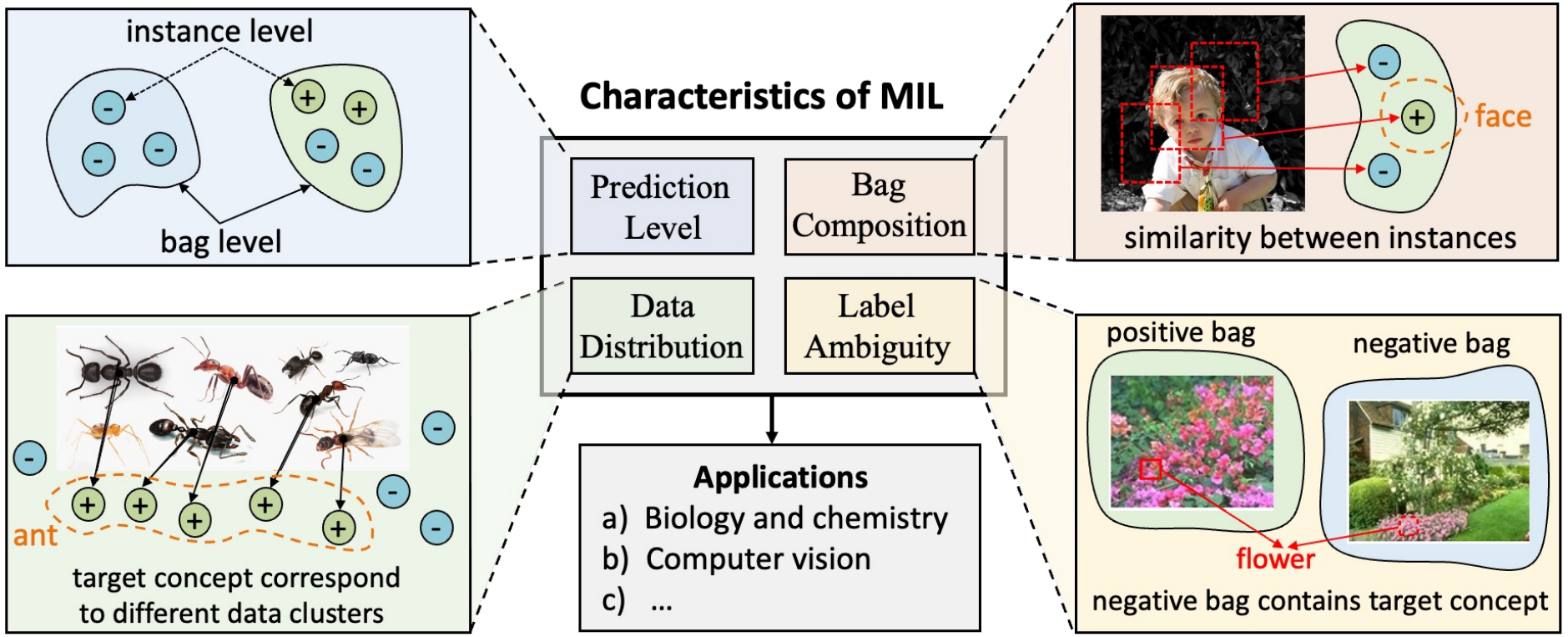


多示例学习(Multi-Instance Learning)

- 多示例学习训练数据集中每一个数据看做一个包(Bag)，每个包由多个示例(Instance)构成，每个包有一个可见的标签。
- 多示例学习假设每一个正包必须存在至少一个关键示例。
- 多示例学习的过程就是通过模型对包及其包含的多个实例进行分析预测得出包的标签。

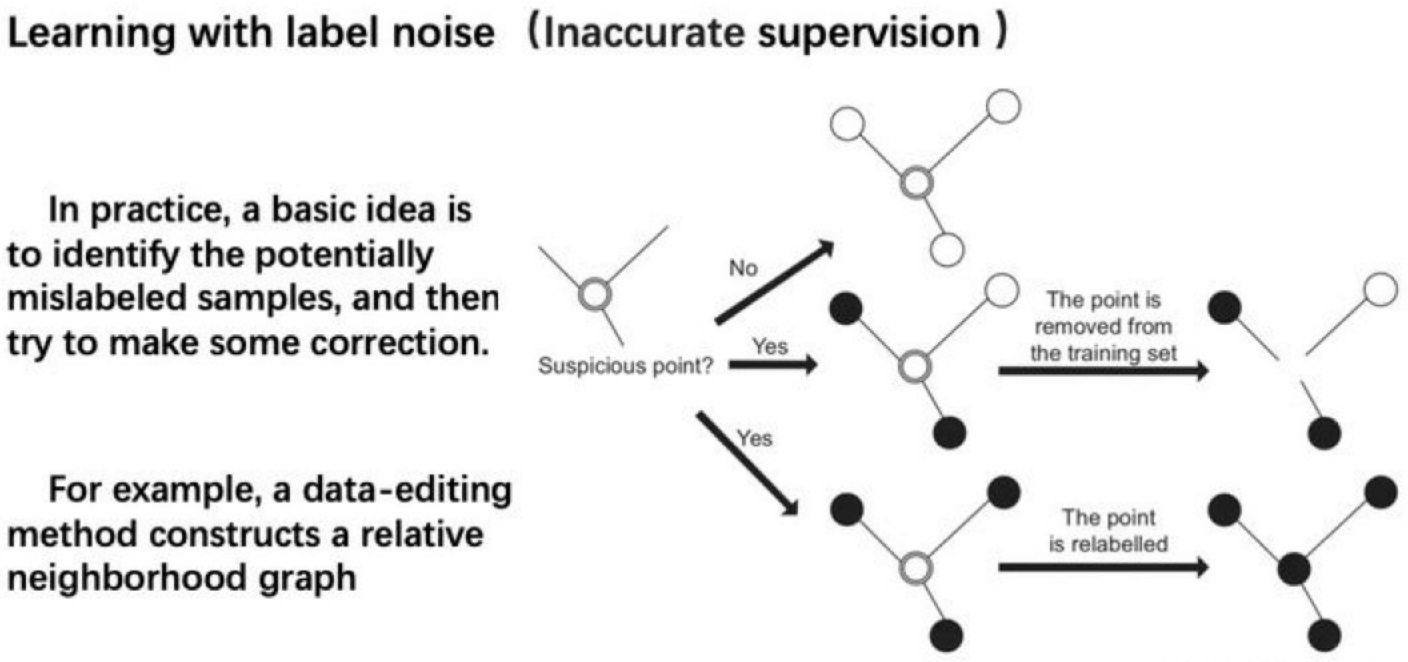


多示例学习的特点

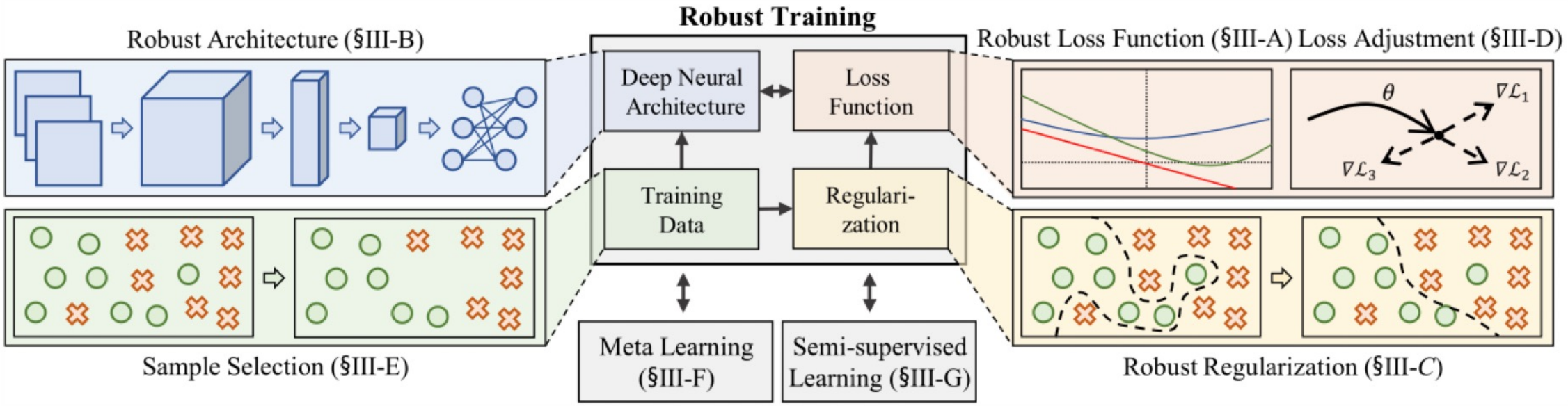

instance level
bag level
Characteristics of MIL
Prediction Level
Bag Composition
Data Distribution
Label Ambiguity
Applications
a) Biology and chemistry
b) Computer vision
c) ...
face
similarity between instances
ant
target concept correspond to different data clusters
positive bag
negative bag
flower
negative bag contains target concept

带噪学习

- 带噪学习（Learning with label noise）基本的思想是识别潜在的误分类样本，然后尝试进行修正。

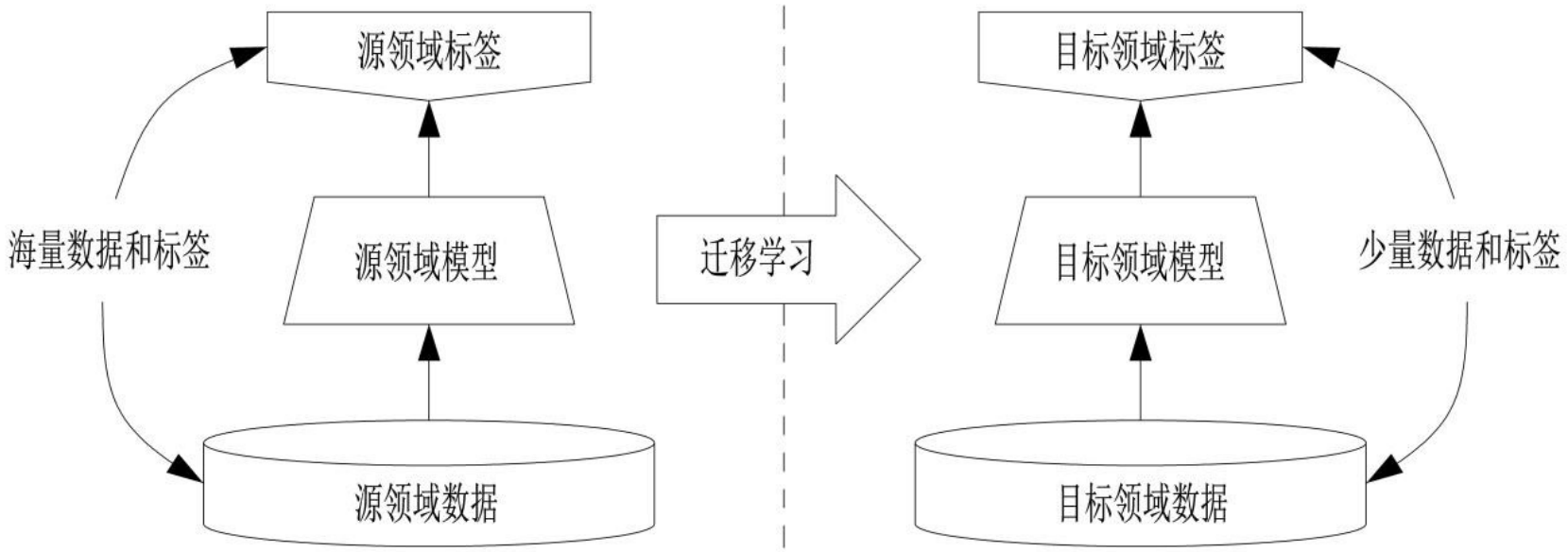


解决带噪学习的思路


Robust Training
Robust Architecture (§III-B)
Deep Neural Architecture
Loss Function
Training Data
Regulari-zation
Robust Loss Function (§III-A)
Loss Adjustment (§III-D)
θ
$\nabla\mathcal{L}_1$
$\nabla\mathcal{L}_3$
$\nabla\mathcal{L}_2$
Sample Selection (§III-E)
Meta Learning (§III-F)
Semi-supervised Learning (§III-G)
Robust Regularization (§III-C)

迁移学习(Transfer Learning)

- 迁移学习(Transfer Learning)：将已经学习过的知识迁移应用到新的问题中
  - 在数据独立同分布不成立的条件下



机器学习简介 – 目录

- 机器学习原理与概念
- 机器学习分类
- 机器学习关键思想
- 机器学习与人工智能

常用的定理

- 没有免费午餐定理(No Free Lunch Theorem，NFL)
  - 对于基于迭代的最优化算法，不存在某种算法对所有问题（有限的搜索空间内）都有效。如果一个算法对某些问题有效，那么它一定在另外一些问题上比纯随机搜索算法更差

不存在一种机器学习算法适合于任何领域或任务

没有免费午餐定理（No Free Lunch Theorem，NFL）是由Wolpert 和Macerday在最优化理论中提出的

常用的定理

- 丑小鸭定理(Ugly Duckling Theorem)：丑小鸭与白天鹅之间的区别和两只白天鹅之间的区别一样大

什么才是相似的？

世界上不存在相似性的客观标准，一切相似性的标准都是主观的。

分类结果取决于选择什么特征作为分类标准，而特征的选择又依赖于任务的目的。

这里的“丑小鸭”是指白天鹅的幼雏，而不是“丑陋的小鸭子”

1969 年由渡边慧提出

常用的定理

- 奥卡姆剃刀原理(Occam's Razor)
  - 如无必要，勿增实体

机器学习中的正则化思想：简单的模型泛化能力更好。如果有两个性能相近的模型，我们应该选择更简单的模型

由14世纪逻辑学家William of Occam提出的一个解决问题的法则

归纳偏置(Inductive Bias)

- 很多学习算法经常会对学习的问题做一些假设，这些假设就称为归纳偏置
  - 在最近邻分类器中，我们会假设在特征空间中，一个小的局部区域中的大部分样本都同属一类。
  - 在朴素贝叶斯分类器中，我们会假设每个特征的条件概率是互相独立的。
  - 归纳偏置在贝叶斯学习中也经常称为先验 (Prior)。

E.g., 深度神经网络就认为层次化处理信息会有好的效果；卷积神经网络则认为信息具有空间局部性，可以用滑动卷积共享权重的方法降低参数空间；循环神经网络则将时序信息考虑进来，强调顺序的重要性；图网络则认为中心节点于邻居节点的相似性会更好引导信息流。

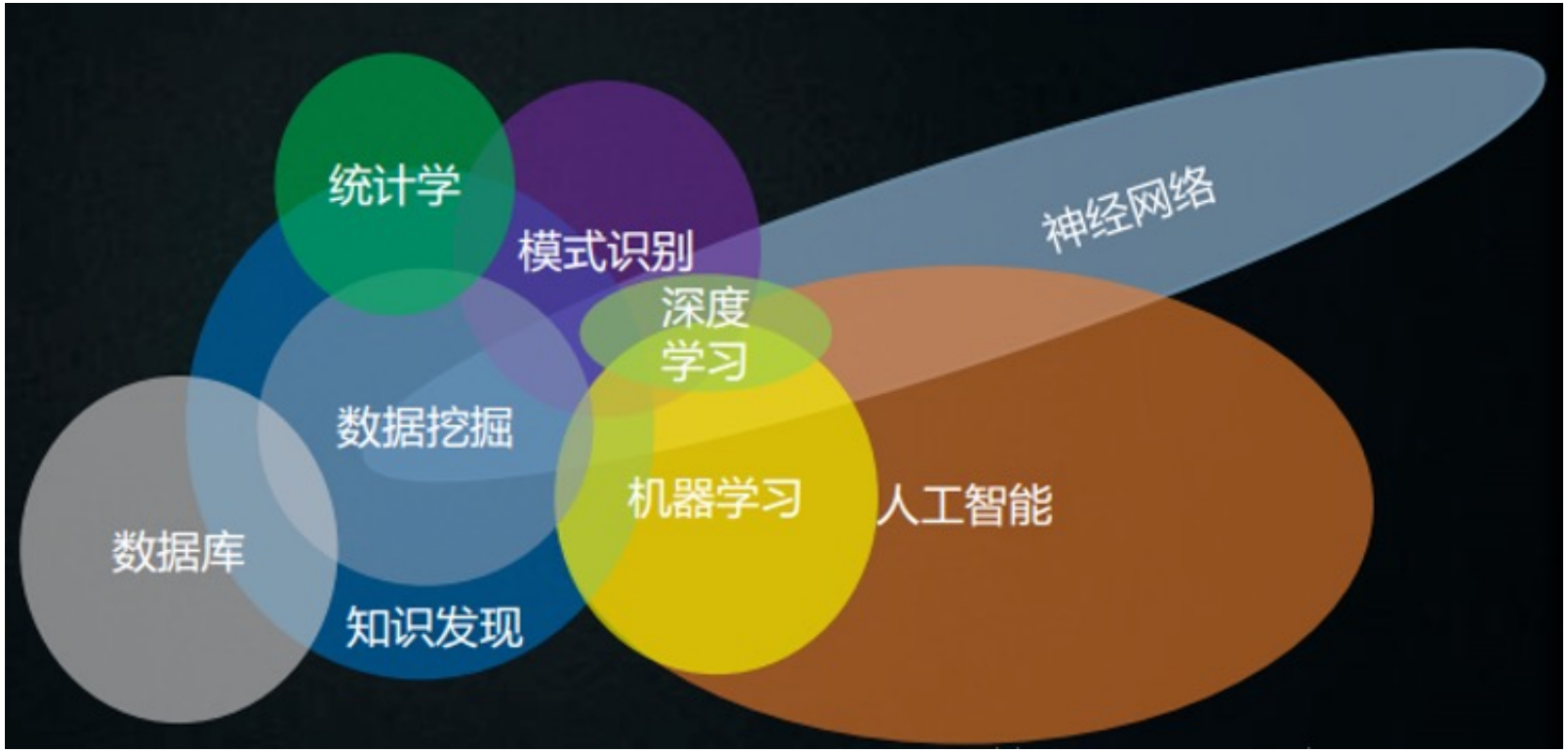
偏置/偏好：在学习算法之初，就人为地认为某一种解决方案优先于其他，这种偏好既可以是在底层数据分布的假设上，也可以体现在模型设计上。

机器学习简介 – 目录

- 机器学习原理与概念
- 机器学习分类
- 机器学习关键思想
- 机器学习与人工智能

机器学习与人工智能

- 机器学习
- 模式识别
- 人工智能
- 深度学习
- 数据挖掘
- ……



机器学习与人工智能

- 模式识别：自己建立模型刻画已有的特征，样本是用于估计模型中的参数。模式识别的落脚点是感知
- 机器学习：根据样本训练模型，如训练好的神经网络是一个针对特定分类问题的模型；重点在于“学习”，训练模型的过程就是学习；机器学习的落脚点是思考，是一种实现人工智能的方法
- 深度学习：深度学习本来并不是一种独立的学习方法，其本身也会用到有监督和无监督的学习方法来训练深度神经网络，是一种实现机器学习的技术。

机器学习是人工智能的基础