人工智能基础

北京交通大学 软件学院

王方石

Email: fshwang@bjtu.edu.cn

第4章　机器学习

4.1 机器学习概述

4.2 监督学习

4.3 无监督学习

4.4 弱监督学习

本章学习目标

- 理解机器学习的定义、基本术语及三个视角。
- 掌握监督学习、无监督学习的基础知识和典型算法。
- 了解弱监督学习的三种方法和应用场景。

4.1 机器学习概述

4.1.1 机器学习的定义

机器学习主要有以下几种定义。

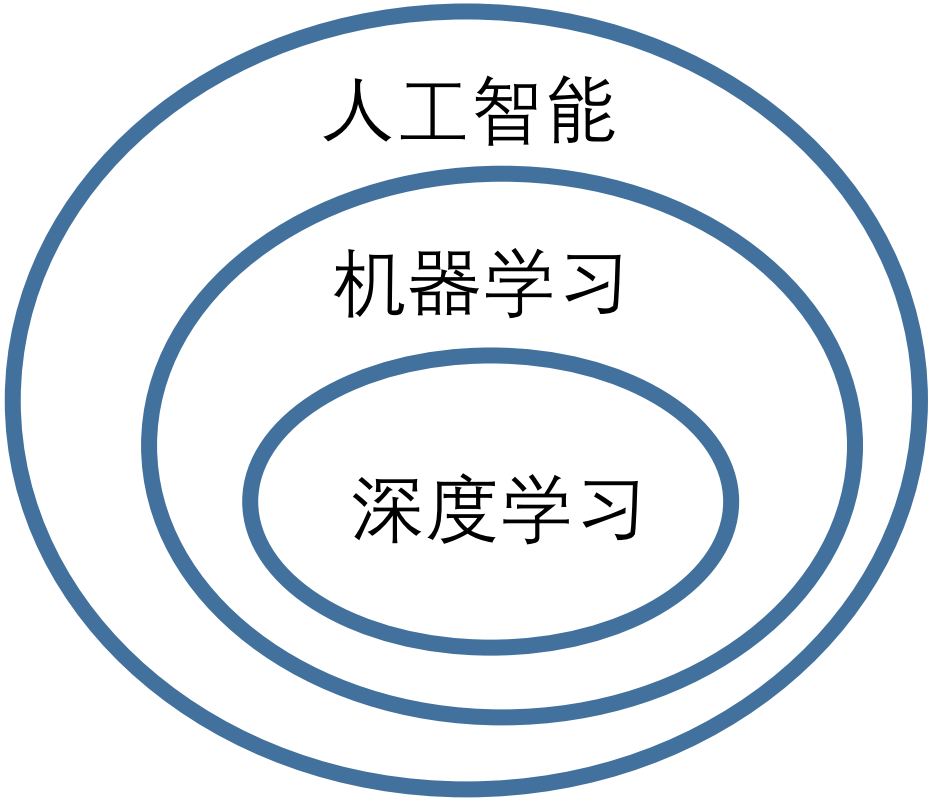
（1）机器学习主要研究如何在经验学习中改善计算机算法的性能。

（2）机器学习是研究用数据或以往的经验来优化计算机程序性能的科学。

（3）机器学习是一门研究机器获取新知识和新技能，并识别现有知识的学问。

（4）机器学习研究算法和数学模型，用以逐步提高计算机系统在特定任务上的性能。

AI、Machine Learning (ML)、Deep Learning (DL)的关联



成功实现AI应用的三要素:

- ◆算法（菜谱）
- ◆算力（厨具）
- ◆数据（食材）

深度学习=大**数据**+高性能**计算**+灵巧的**算法**

4.1.1 机器学习的定义

目前为止，尚未有一个公认的机器学习定义。而作者认为：

- 机器学习是研究如何使机器模拟或实现人类的学习行为，以获取知识和技能，并不断改善系统自身性能的学科。
- 机器学习是一门多领域交叉学科，涵盖概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科的知识、理论和方法。
- 它的**根本任务**是数据的智能分析与建模，并从数据中挖掘出有价值的信息。
- 其**目标：**要构建可以从数据中学习、并对数据进行预测的系统。

4.1.1 机器学习的定义

- 机器学习是AI的一个分支，是其中最具智能特征、最前沿的研究领域之一，是AI的核心和研究热点。
- 机器学习是实现人工智能的关键和重要途径。
- 机器学习理论和方法在基于知识的系统、自然语言理解、语音识别、计算机视觉、机器人、模式识别、生物信息学等许多领域得到了广泛应用。
- 一个系统是否具有学习能力已成为评判其是否具有“智能”的一个标准。
- 机器学习的研究主要分为**两大类**：
  - **基于统计学的传统机器学习**：主要研究学习机制，注重探索模拟人的学习机制，研究方向包括：支持向量机、决策树、随机森林等。
  - **基于大数据和人工神经网络的深度学习**：主要研究如何充分利用大数据时代下的海量数据，采用深度学习技术构建深度神经网络，从中获取隐藏的、有效的、可理解的知识。

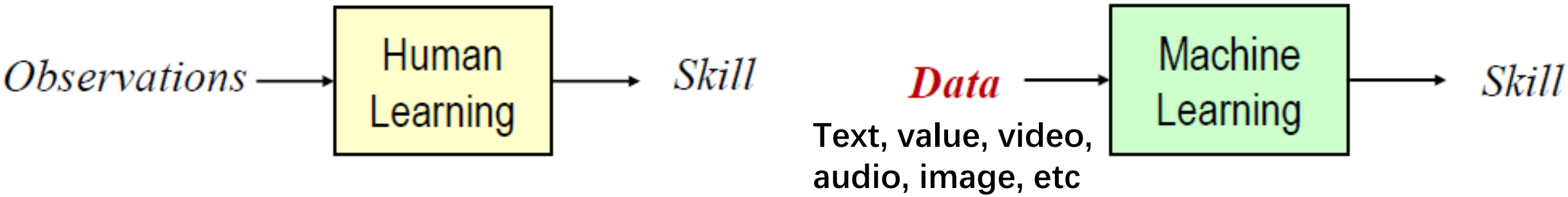
人工智能与机器学习

◆ **人类学习**

人类是从**观察**中积累**经验**来获取**技能**。

◆ **机器学习**

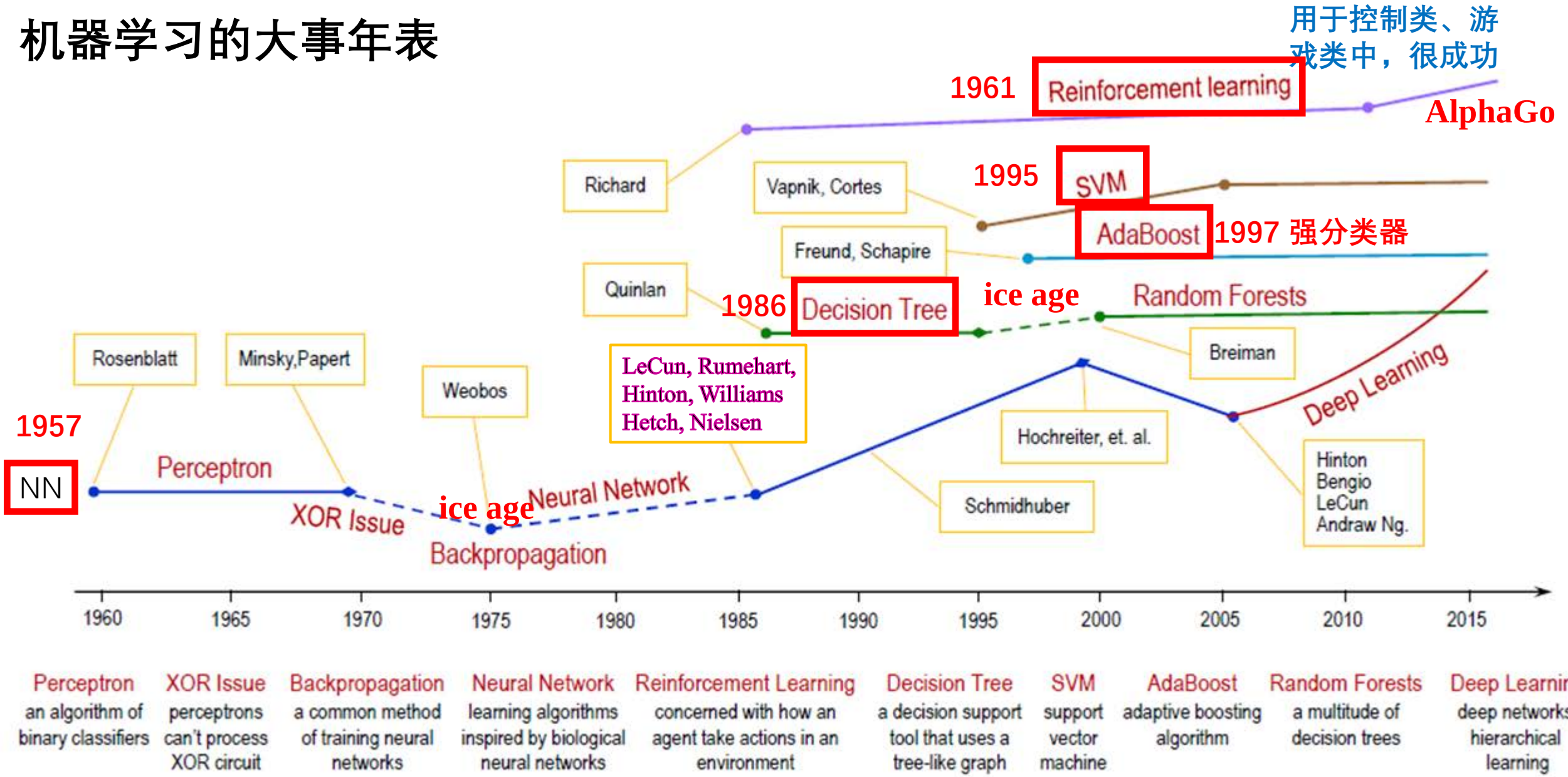
机器是从**数据**中积累或者计算的**经验中**获取技能。



机器模拟人类的学习行为.

机器学习的发展历史

机器学习的大事年表



4.1.2 机器学习的基本术语

1. 数据集（Dataset）

数据集是指数据的集合。例如（20301001，张三，175cm，70kg）。

2. 样本（Sample）

- 样本也称为实例（Instance），指待研究对象的个体，包括属性已知或未知的个体。
- 例如，每个学生所对应的一条记录就是一个“样本”。数据集即为若干样本的集合。

3. 标签（Label）

- 标签是为样本指定的数值或类别。
- 在**分类**问题中，标签是样本被指定的特定类别;
- 在**回归**问题中，标签是样本所对应的实数值。
- **已知样本**是指标签已知的样本，**未知样本**是指标签未知的样本。

4.1.2 机器学习的基本术语

4. 特征（Feature）

- 特征是指样本的一个独立可观测的属性或特性。
- 它反映样本在某方面的表现或性质，
- 例如“姓名”“身高”是“特征”或“属性”。
- 特征的取值，例如“张三”“175cm”是“特征值”或“属性值”。

5. 特征向量（Feature Vector）

- 特征向量是由样本的n个属性组成的n维向量，第i个样本X_i表示为：$(x_{i1}, x_{i2}, \ldots, x_{in})$
- **特征分为**
  - **手工式特征**也称为**设计式特征**,是指由学者构思或设计出来的特征, 如SIFT、HOG等。
  - **学习式特征**是指由机器从原始数据中自动生成的特征。

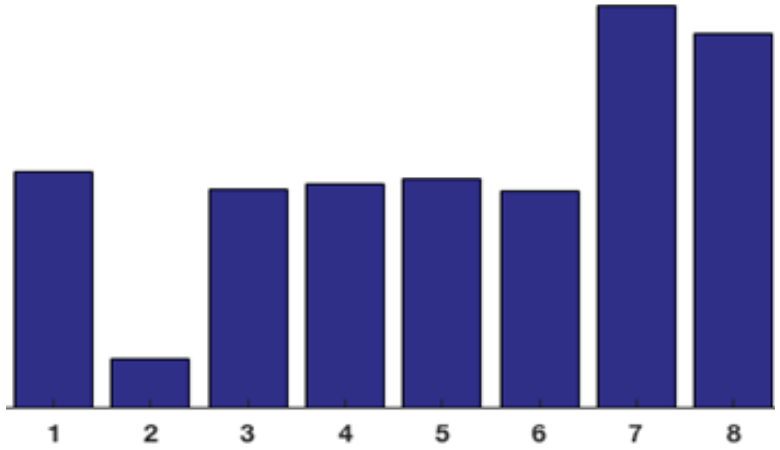
  例如，通过卷积神经网络获得的特征就属于学习式特征。

手工式特征 / 设计的特征

例如:

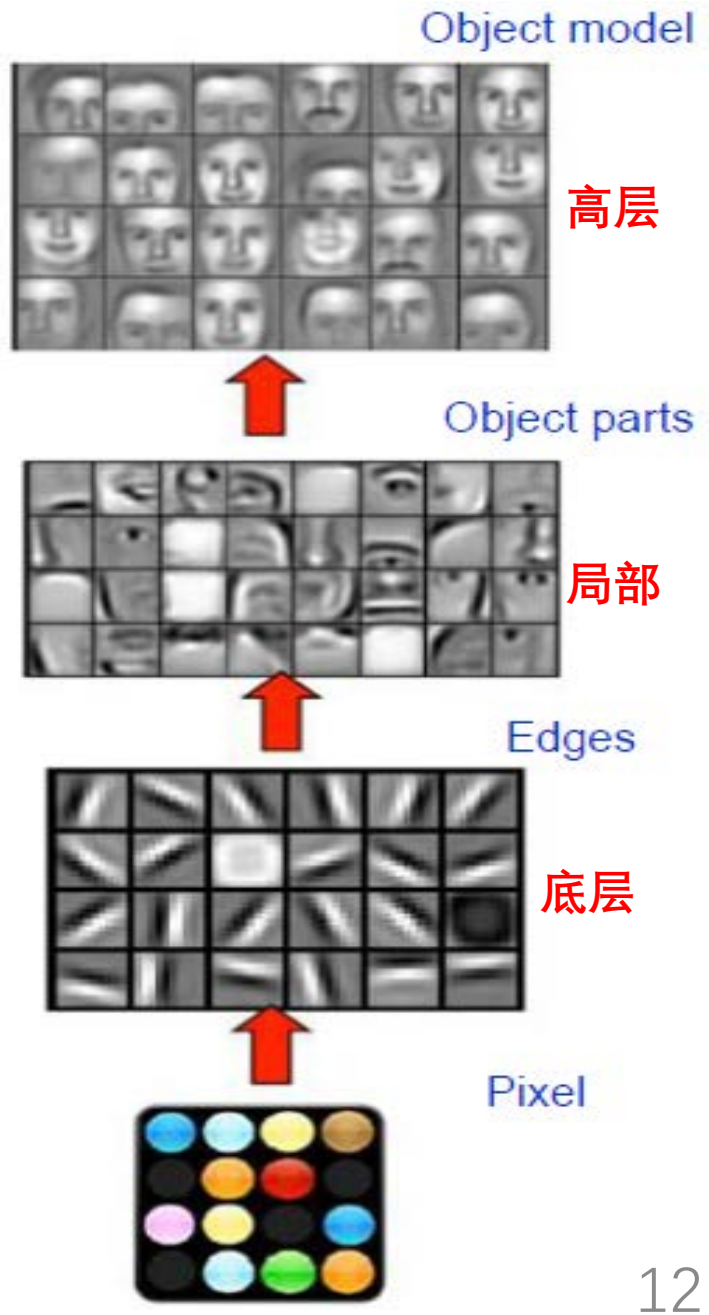
- HOG (Histogram of Oriented Gradients, 定向梯度直方图)
- 用边缘检测算子提取的边缘特征。

边缘特征



HOG特征

学习式特征



4.1.2 机器学习的基本术语

6. 特征空间（Feature Space）

- 特征空间是指特征向量所在的p维空间，每一个样本是该空间中的一个点。**特征空间**也称为**样本空间**。
- 例如，将“身高”“体重”作为两个坐标轴，它们就形成了用于描述学生体态的二维空间，每个学生在此特征空间中都能找到自己的位置坐标。

4.1.2 机器学习的基本术语

通常将数据集分成训练集、验证集和测试集，需要保证这三个集合是不相交的。

（1）训练集（Training Dataset）

- 训练过程中使用的数据称为“**训练数据**”，其中每个训练数据称为一个“**训练样本**”；
- 每个训练样本都有一个已知标签，由所有训练样本及其标签组成的集合称为“**训练集**”。
- 训练集包括**一个样本集**和**一个对应的标签集**，用于学习得到拟合样本的模型。
- 一般地，训练集中的标签都是正确的，称为**真实标签**（Ground-Truth）。
- 例如，在图像分类任务中，训练集包括
  - 一个由特定图像组成的样本集合
  - 一组由语义概念（如山、水、楼等）组成的标签集合，标签即为Ground-Truth。

4.1.2 机器学习的基本术语

（2）验证集（Validation Dataset）

- 在实际训练中，有时模型在训练集上的结果很好，但对于训练集之外的数据的结果并不好。
- 此时，可单独留出一部分样本，不参加训练，而是用于调整模型的超参数，并对模型的能力进行初步评估，这部分数据称为**验证集**。
- **超参数**（hyperparameter）是指模型中人为设定的、无法通过训练得到的参数，如NN的层数、卷积的尺寸、滤波器的个数、KNN和K-Means算法中的K值等。

（3）测试集（Test Dataset）：

- 测试过程中使用的数据称为“**测试数据**”，被预测的样本称为“**测试样本**”，测试样本的集合称为“**测试集**”。
- 测试集不参与模型的训练过程，仅用于评估最终模型的**泛化能力**。

4.1.2 机器学习的基本术语

7. 泛化能力（Generalization Ability）

泛化能力是指训练得到的模型对未知样本正确处理的能力，即模型对新样本的**适应能力**，亦称为**推广能力**或**预测能力**。

8. 模型参数

- 给定训练集，希望能够拟合一个函数 $f(x,\theta)$来完成从输入的特征向量到标签的映射。
- 对于连续的标签或非概率模型，通常会采用拟合函数来表示从输入空间（样本集X）到输出空间（标签集Y）的映射： $Y' = f(x,\theta)$

  其中，Y' 是样本x的**预测标签**，θ为模型中可训练得到的参数，即**模型参数**，也称为**学习参数**，**并非是由人为设置的超参数**。

4.1.2 机器学习的基本术语

9.学习算法（learning algorithm）

- 希望为每个样本x**预测的标签**与其所对应的**真实标签**都相同，这就需要有一组好的模型参数θ。
- 为了获得这样的参数θ，则需要有一套学习算法来优化函数 f，此优化过程称为**学习**(Learning)或者**训练**(Training)，拟合函数 f 称为**模型**（Model）。

10. 假设空间（hypothesis space）

- 从输入空间至输出空间的映射可以有多个，它们组成的映射集合称为**假设空间**。
- **学习的目的：**在此假设空间中选取最好的映射，即**最优的模型**。
- 用训练好的最优模型对测试样本进行预测的过程称为**测试**。

4.1.2 机器学习的基本术语

11. **损失函数（Loss function），**也称为**代价函数**（Cost Function），用于度量**预测标签**和**真实标签**之间差异或损失。

真实标签集表示为Y，**预测标签集**表示为$Y'=f(X)$，则损失函数记为$L(Y,f(X))$，是一个**非负**的实值函数。常用的损失函数包括：

- **0-1**损失函数公式为：$L(Y,f(X))=\begin{cases}0, & \text{if } Y=f(X)\\ 1, & \text{if } Y\neq f(X)\end{cases}$
- **平方**损失函数公式为：$L(Y,f(X))=\frac{1}{2}(Y-f(X))^2$ **平方误差损失也称为L2损失。**
- **绝对**损失函数公式为：$L(Y,f(X))=|Y-f(X)|$，**绝对误差也称为L1损失。**
- **对数**损失函数公式为：$L(Y,f(X))=-logP(Y|X)$
- **交叉熵**损失函数公式为：$L(Y,f(X))=-\sum_{c=1}^{C}Y_c\,logf(X_c)$

4.1.2 机器学习的基本术语

12.风险函数（Risk Function）

- 风险函数又称**期望损失**（Expected Loss）或**期望风险**（Expected Risk），是所有数据集（包括训练集和预测集）上损失函数的期望值，用于度量平均意义下模型预测的好坏。
- **机器学习的目标**是选择风险函数最小的模型。

4.1.2 机器学习的基本术语

13.优化算法

- 在获得了数据集、确定了假设空间以及选定了损失函数之后，需要解决**最优化**问题。
- 机器学习的训练和学习的过程，就是求解最优化问题的过程，寻找**全局最优解**。
- 若最优化问题存在显式的解析解，则可以很容易求得它的解；
- 但通常不存在解析解，则只能通过数值计算的方法来不断逼近它的解。
- 最简单也最常用的优化算法是**梯度下降法**（Gradient Descent，GD）。
- 梯度下降法通过不断迭代的方式来降低风险函数的值，公式为：$\theta_{t+1} = \theta_t - \eta \frac{\partial R(\theta)}{\partial \theta}$

  其中，θ_t为第t次选代时的参数值，R(θ)为风险函数，η为优化的步长，又称为**学习率**。
  - 学习率过小，会导致学习速度太慢，还可能导致陷入局部最优；
  - 学习率过大，又会出现震荡，严重时会导致发散。

4.1.2 机器学习的基本术语

14. 机器学习的基本流程是：数据预处理→模型学习→模型评估→新样本预测。

① 数据预处理

收集并处理数据，有时还需要完成数据增强、裁剪等工作，划分训练集、验证集、测试集。

② 模型学习，即模型训练

- 在训练集上运行学习算法，利用损失函数和优化算法求解一组模型参数，得到风险函数最小的最优模型。
- 一般在训练集上会反复训练多轮，即训练样本被多次利用。

4.1.2 机器学习的基本术语

③ 模型评估

- 将验证集样本输入到学习获得的模型中，用以评估模型性能，还可以进一步调节模型的超参数，找到最合适的模型配置。
- 常用的模型评估方法为K折交叉验证。
- 通常所说的“模型调参”一般指的是调节超参数，而不是模型参数。

④ 新样本预测

- 将测试集中的样本输入到训练好的模型中，对比预测的结果与真实值，计算出各种评价指标，以此来**评价模型的泛化能力**。
- 例如，图像分类任务有精确率（Precision）和召回率（Recall）等评价指标。

4.1.2 机器学习术语

从机器学习的基本流程可知，**学习算法有三个基本要素**：

- **模型**（哪一类模型：线性模型、概率模型、非线性模型、网络模型）
- **损失函数（学习准则、学习策略）**：选出什么损失函数来衡量错误的代价，才能找到**最优的模型参数**。
- **优化算法，也称为优化器，**最简单也最常用的优化算法是**梯度下降法。**

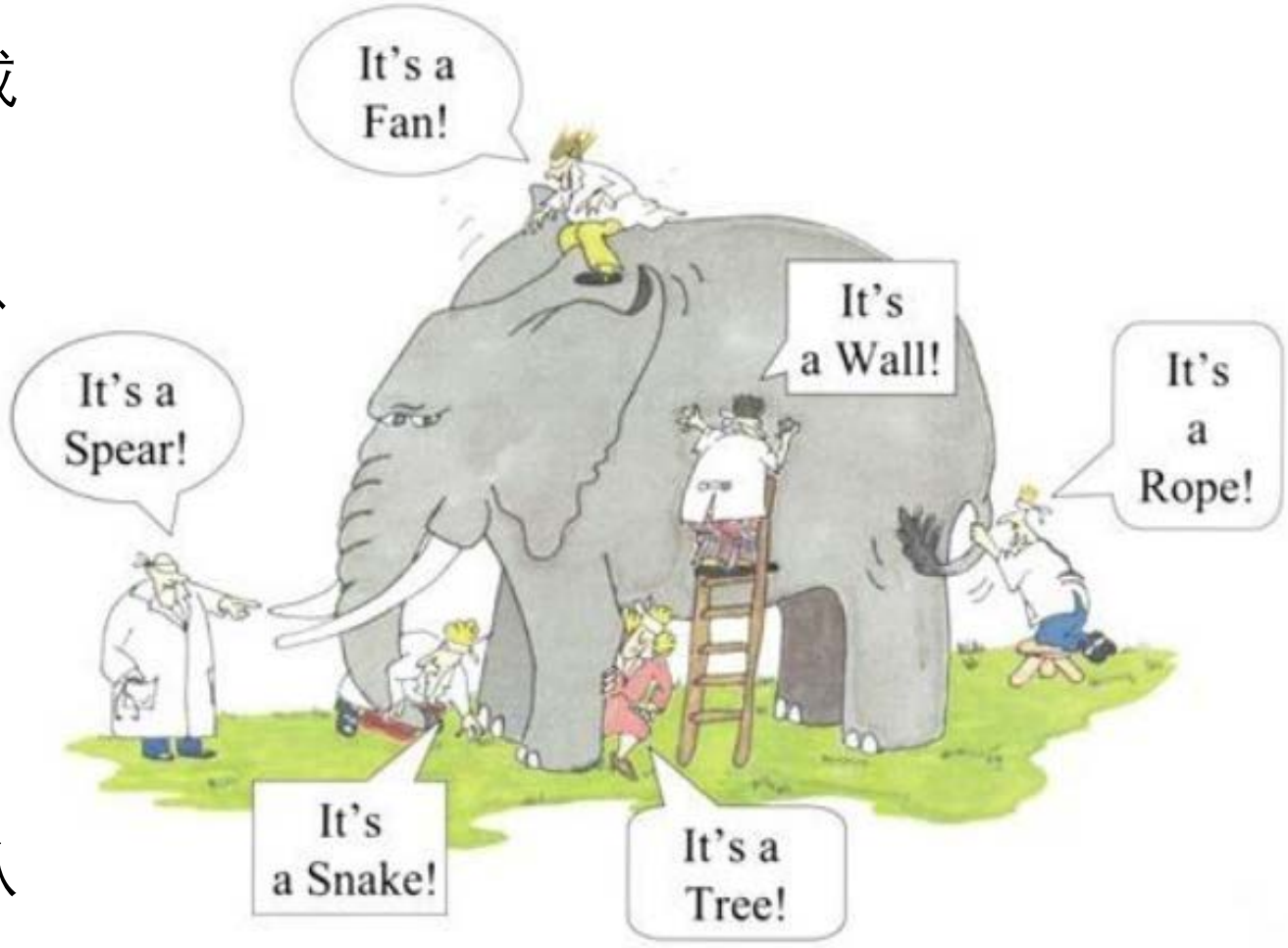
这三个要素都需要学者根据经验人为确定。

4.1.3 机器学习的三个视角

◆机器学习算法众多，且应用广泛，可完成很多任务：

- **图像分类**的方法有：SVM、KNN、朴素贝叶斯分类、决策树、随机森林等
- **预测某一地区房价**的方法有：多元线性回归、多项式回归、马尔可夫预测模型、遗传算法等。

◆**需要对机器学习算法进行全面的了解，**从三个视角来介绍机器学习算法：**学习任务、学习范式**和**学习模型**。



Maybe "Blind Men and an Elephant"

Three Perspectives of Machine Learning
机器学习的三个视角

Perspectives	Description 描述
Learning Tasks 学习任务	表示可以用机器学习解决的**通用问题（分类、回归、聚类、排名、降维）**。
Learning Paradigms 学习范式	表示机器学习的**典型场景（有无数据、环境互动？）**。
Learning Models 学习模型	表示可以完成一个学习任务的**方法（SVM，KNN）**。

1. 学习任务

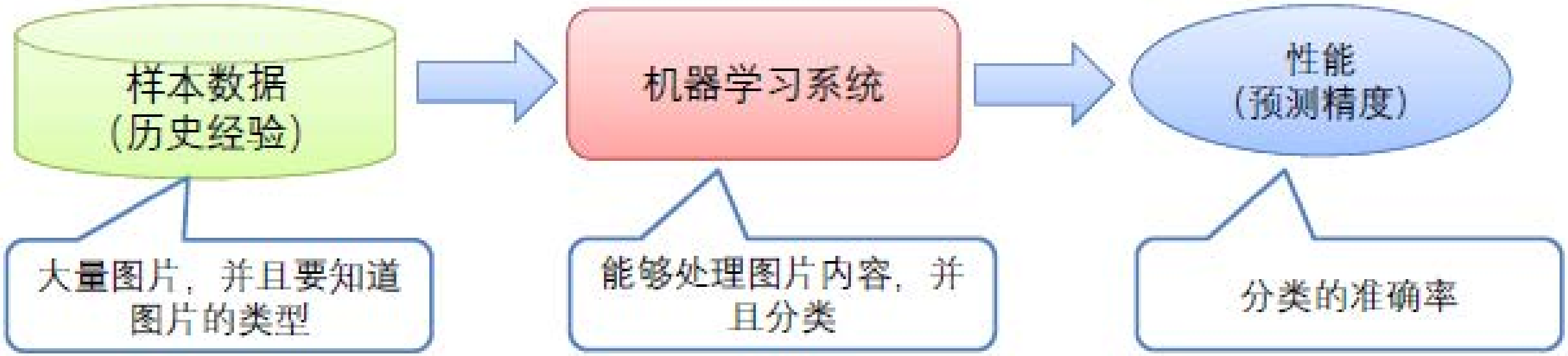
◆ **学习任务**（Learning Tasks）是指可以用机器学习方法解决的通用问题。

◆ 机器学习中的**典型任务**包括:

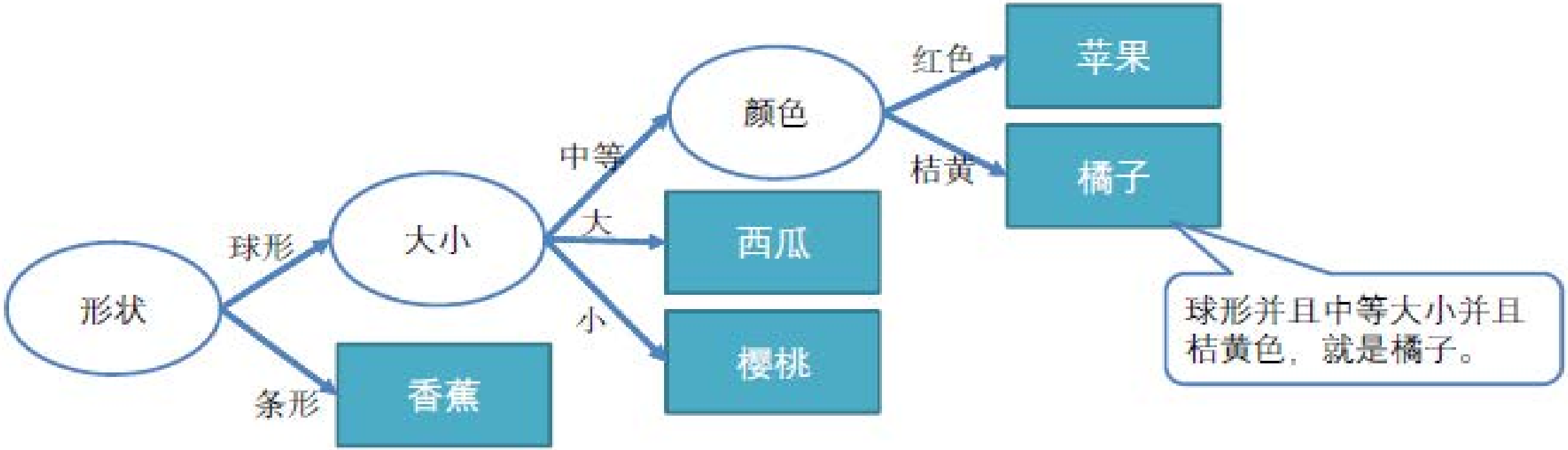
（1）分类（Classification）是将输入数据划分成两个或多个类别。

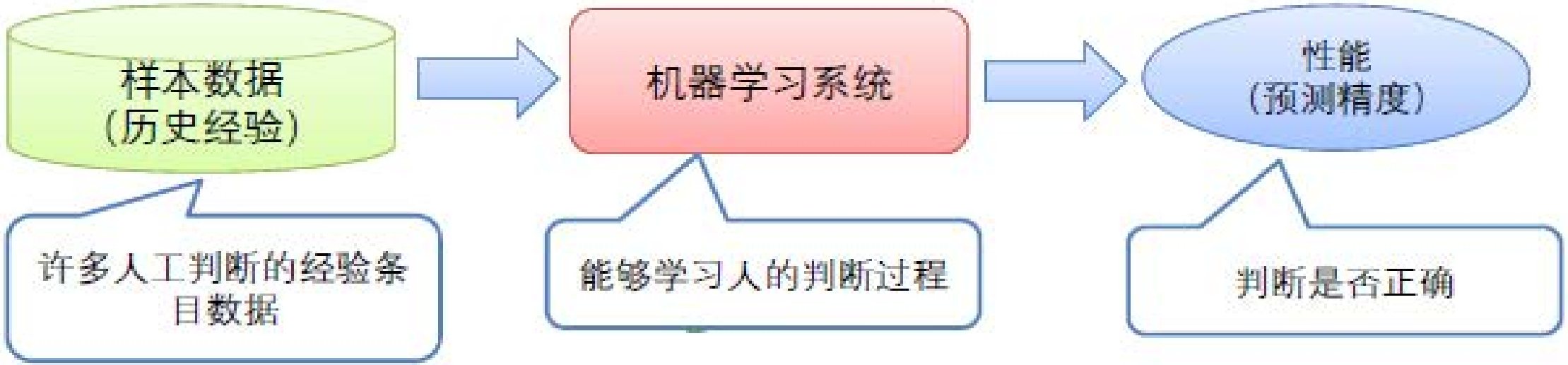
- 输出值是离散的。
- 例如，垃圾邮件过滤、人脸识别、银行用户信用评级、手写体字符和数字识别等。
- 解决此类任务的典型算法有：支持向量机、K-近邻、朴素贝叶斯、决策树、逻辑回归算法等。

例4.1 图像分类的机器学习



例4.2 归纳规律（预测判断）

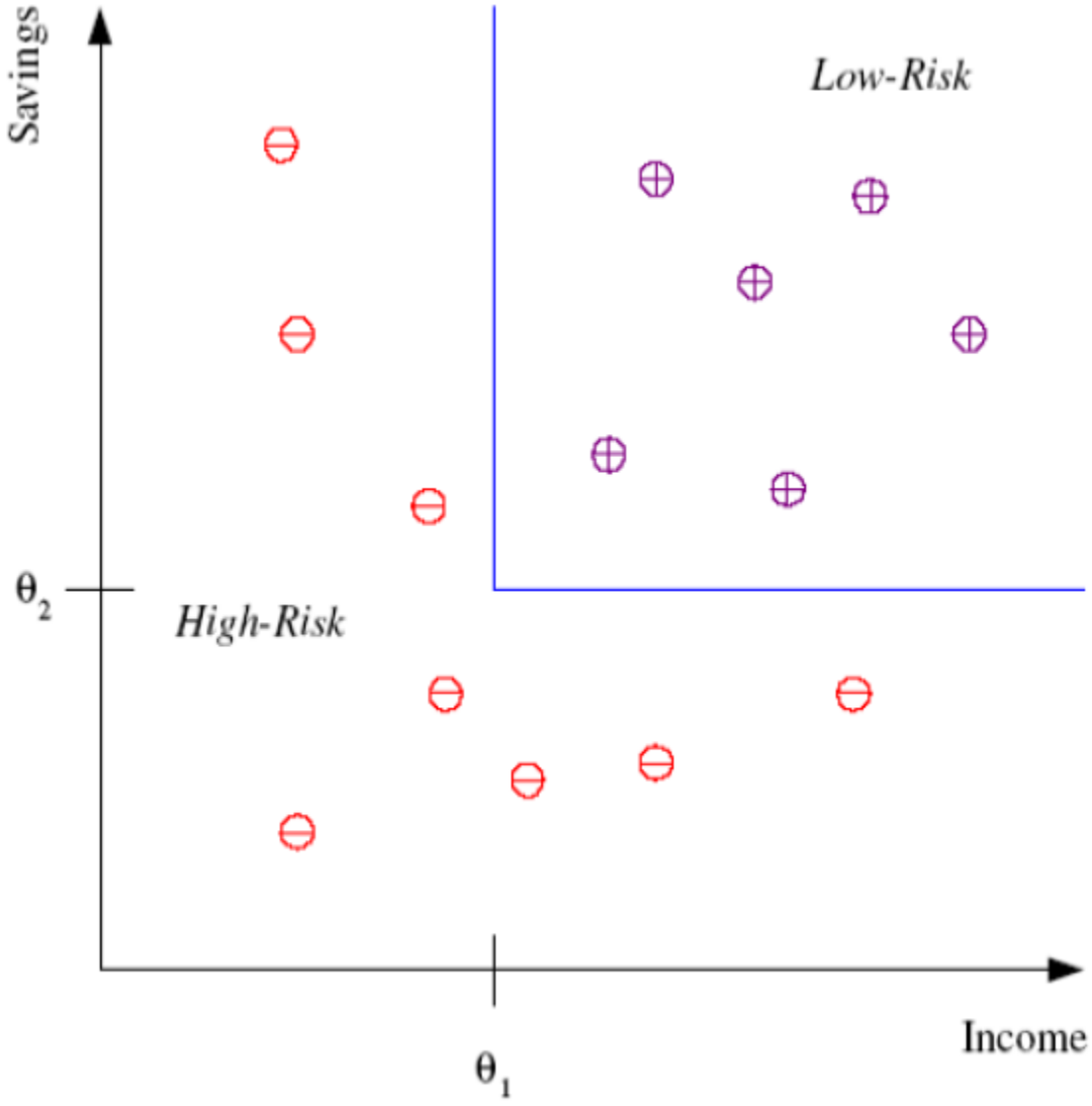




例4.3 信用评分

- **二分类**：低风险和高风险客户。
- 根据客户信息，将其归为二类中的一类。
- 用过去的数据训练之后，可以学习得到如下分类规则：

IF *income* > θ_1 AND *savings* > θ_2
THEN *low-risk*
ELSE *high-risk*



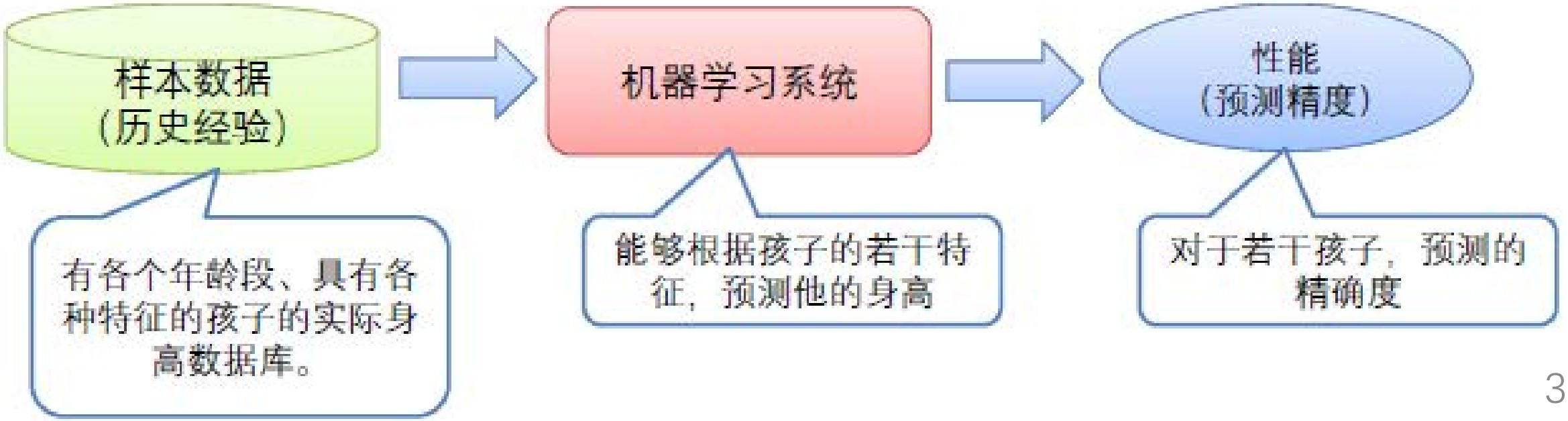
1. 学习任务

（2）回归（Regression）是确定某些变量之间定量关系的一种统计分析方法，即建立数学模型并估计未知参数。

- 回归就是找到一个函数，给定一个输入特征值X，便能输出一个与之相对应的连续数值（不是离散的）。
- 常用于预测股票行情、二手车价格、身高、体重、医学诊断等。
- 与分类问题不同，回归预测的是数值而不是类别。
- 解决此类任务的典型算法有：多元线性回归、贝叶斯线性回归（Bayesian Linear Regression）、多项式回归等算法。

例4.4 身高预测（预测数值---回归）

如何根据年龄、性别、体重等等特征，来预测小孩的身高？



1. 学习任务

（3）聚类（Clustering）是指将具体的或抽象的对象的集合分成由相似对象组成的多个**不知名称的组**（Group）或**簇**（Cluster）的过程。

- 此处，不用“类别”一词代替“组”或“簇”，以示与“分类”任务的区别。
- 聚类也称为**聚类分析**，在日常生活和工作中已有广泛的应用。
- 例如，可以采用聚类方法，根据用户行为、销售渠道、商品等原始数据，将相似的市场和用户聚集在一起，以便找准潜在的目标客户和市场。
- 解决此类任务的典型算法有：K均值聚类（也有人称其为C均值聚类）、层次聚类、模糊C均值聚类、基于密度的聚类等。

1. 学习任务

（4）排名（Ranking）是指依据某个准则对项目进行排序。

- 主要应用场景是各大搜索引擎对基于关键词的查询结果的条目进行排序。
- 解决此类任务的典型算法有网页排名（PageRank）算法，它是一种利用网页（节点）之间的超链接数据进行计算的技术，用于对搜索到的结果列表进行评估和排名，以体现网页与特定查询的相关性和重要性。

（5）降维（Dimensionality Reduction）是指通过将输入数据从高维特征空间映射到低维特征空间，去除无用、冗余的特征，降低学习的时间复杂度和空间复杂度。

- 具体应用有特征工程中的特征选择（选择最有效的特征子集）、数据可视化（低维数据易于可视化）。
- 解决降维的典型算法有主成分分析法（PCA）、线性判别分析（LDA，又称为Fisher线性判别，FDA）、多维缩放（MDS）等。

2. 学习范式

2. 学习范式（Learning Paradigms）是指机器学习的场景或模式。

根据机器学习模型训练时所使用的数据集的完整性和质量，通常将机器学习分为：

（1）监督学习（Supervised Learning）。

- 监督学习是指采用一组有标注的数据样本对模型进行训练，再用训练好的模型对未知样本做出预测。
- 也可以理解为：利用有标注的数据学习到一个模型，用以建立从输入到输出的一种映射关系，再用该模型对测试数据集中的样本进行预测。
- 监督学习的训练数据由两部分组成，即描述事件/对象的特征向量（x）和真实标签（y），**有训练模型的过程**。
- 需要采用监督学习方法完成的学习任务主要包括：**分类、回归和排名**。
- 典型的监督学习方法有：SVM、KNN、线性回归、决策树、隐马尔可夫模型等。

2. 学习范式

（2）无监督学习（Unsupervised Learning）。

- 由于缺乏足够的先验知识，因此难以人工标注数据类别，或者人工标注的成本太高，导致数据缺少标注信息，即缺少真实标签。
- 在此情况下，利用未标记（类别未知）的数据样本解决模式识别中的各种问题，称为无监督学习。
- 相比于监督学习的训练数据，无监督学习的数据只是其中的一个部分，即只有描述事件/对象的特征向量（x），但是没有标签（y），且没有训练模型的过程。
- **无监督学习的效果一般比较差**。
- 需要采用无监督学习方法完成的学习任务主要包括：**聚类**和**降维**。
- 典型的无监督学习方法有：**K-Means聚类**、**主成分分析法**等。

2. 学习范式

（3）弱监督学习（Weakly Supervised Learning）。

- 弱监督学习介于监督学习和无监督学习之间，它利用**带有弱标签的训练数据集**进行监督学习，同时利用**大量无标签数据**进行无监督学习。
- **弱标签**是指**标注质量不高的标签**，即标签信息可能**不完全、不确切、不准确**。
- 根据训练时所使用数据的质量，弱监督学习分为
  - 不完全监督学习
  - 不确切监督学习
  - 不准确监督学习
- 虽然将弱监督学习分为了上述三种类别，但在实际操作中，它们经常同时发生。
- **不完全监督学习**又包括：**主动学习、半监督学习、迁移学习、强化学习**。
- 以上各个学习范式的分类并不是严格互斥的。

机器学习的范式

4.2 监督学习

4.3 无监督学习

4.4 弱监督学习

4.4.1 不确切监督学习

4.4.2 不准确监督学习

4.4.3 不完全监督学习

4.4.3.1 主动学习

4.4.3.2 半监督学习

4.4.3.3 迁移学习（样本、特征、模型）

4.4.3.4 强化学习

3. 学习模型

3. 学习模型（Learning Models）用于表示可以完成一个学习任务的方法。

（1）几何模型。

可采用线、面或距离等几何图形模型来构建学习算法。用于学习线性模型的算法有线性回归，用于学习二维平面模型的算法有支持向量机，用于学习距离模型的算法有KNN。

（2）逻辑模型。

用逻辑模型来构建学习算法，其**典型算法**包括：归纳逻辑编程和关联规则算法。

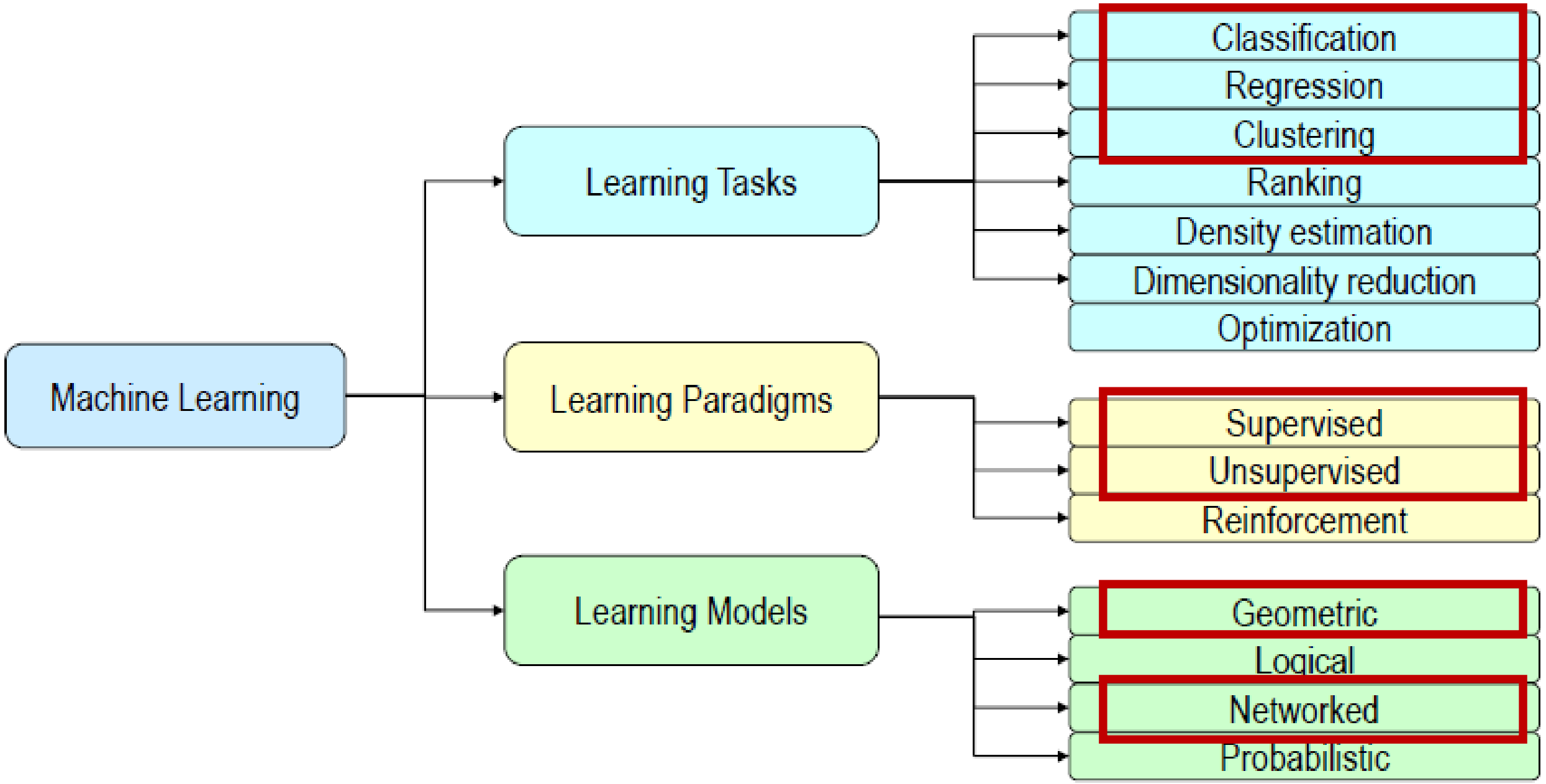
（3）概率模型。

采用概率模型来表示随机变量之间的条件相关性，其**典型算法**包括：贝叶斯网络、概率规划和线性回归等方法。

（4）网络模型。

采用网络模型构建机器学习算法，**典型的**浅层网络有感知机，深层网络有各种深度CNN。

三个视角


Machine Learning
Learning Tasks
Learning Paradigms
Learning Models
Classification
Regression
Clustering
Ranking
Density estimation
Dimensionality reduction
Optimization
Supervised
Unsupervised
Reinforcement
Geometric
Logical
Networked
Probabilistic

4.2 监督学习

4.2.1 监督学习的步骤

4.2.2 监督学习的主要任务：

4.2.2.1 分类

4.2.2.2 回归

4.2.3 监督学习的典型算法

4.2.3.1 KNN

4.2.3.2 SVM

4.2 监督学习

- 监督学习是在机器学习算法中占据绝大部分的一种十分重要的方法。
- **监督学习**是在已知标签（即监督信息）的训练集上学习出一个模型，当向此模型输入新的样本（也称为未知数据，即测试集中的样本）时，可以预测出其所对应的输出值。

期望输出： **cat**

期望输出： **dog**

4.2.1 监督学习的步骤（1）

整理好**训练集**和**测试集，保证两个数据集不相交。**

已知：训练集中有N个数据样本，$X=\{X_1, X_2,\ldots, X_N\}$，M个标签的结果集

$Y=\{y_1, y_2,\ldots, y_M\}$，每个样本$X_i$有对应的标签为$y_i$。

假设指定一个学习模型model，当输入样本X_i $(x_{i1},x_{i2}, \ldots,x_{in})$ 时，

可输出其**预测结果**y_i'；令**期望结果**为y_i。

- 若$y_i'=y_i$，说明实际输出的预测结果正确，与期望结果y_i一致，无需修正模型；
- 若$y_i'\neq y_i$，说明实际输出的预测结果不正确，需要改进模型。

4.2.1 监督学习的步骤（2）

采用监督学习建立学习模型的过程如下。

（1）采用指定的初始模型model，初始化 $i=1$；

（2）若$i>N$（样本数），学习过程结束，得到最终模型；

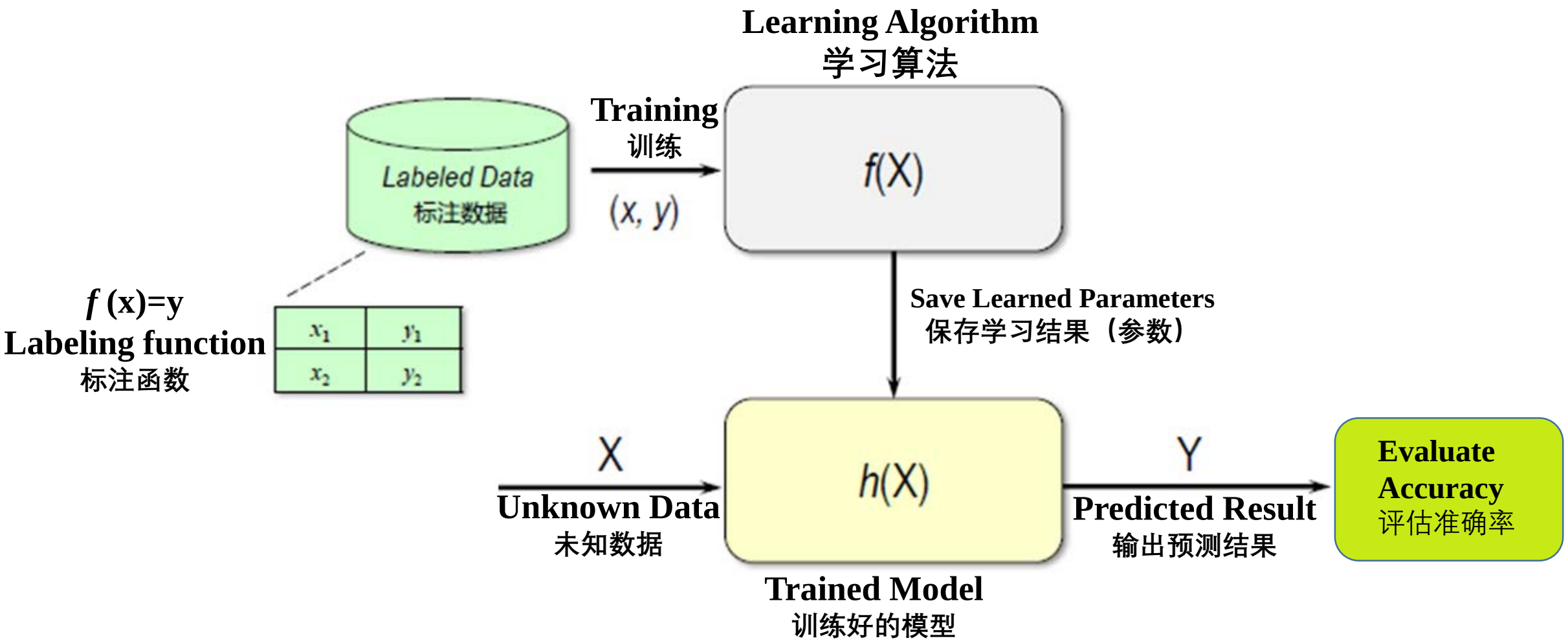
否则，向学习模型model输入**样本X_i**的n维特征向量$(x_{i1},x_{i2},\ldots,x_{in})$，

计算输出结果，记为y_i'；

（3）若$y_i'\neq y_i$（**真实结果**），则将错误结果 y_i' 与**期望结果** y_i 之间的误差作为纠正信号，传回到模型model，用以更新model的参数，改进模型。

（4）令$i=i+1$，返回第（2）步。

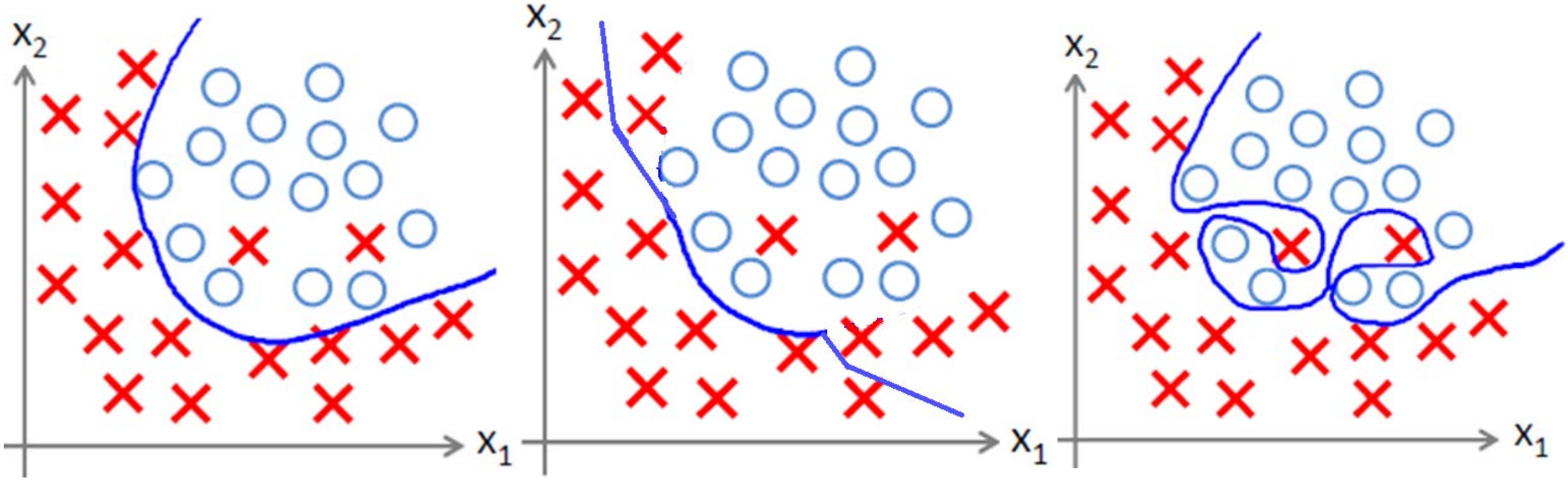
监督学习的步骤


Learning Algorithm
学习算法
Training
训练
Labeled Data
标注数据
(x, y)
f(X)
f (x)=y
Labeling function
标注函数
x_1 y_1
x_2 y_2
Save Learned Parameters
保存学习结果（参数）
X
Unknown Data
未知数据
h(X)
Y
Predicted Result
输出预测结果
Evaluate
Accuracy
评估准确率
Trained Model
训练好的模型

评价监督学习的性能

- 若所学习到的模型能很好地拟合训练集，这种**拟合能力**称为**学习能力**或**训练能力**或**逼近能力**。
- 希望：训练好的模型对测试集也能输出正确的结果。这种适应新样本的能力称为**预测能力**或**泛化能力**或**推广能力**。
- 若所构造的模型不能很好地拟合（逼近）训练数据，称之为“**欠拟合**”。
- 若模型能比较好地拟合（逼近）训练数据，称之为“**良拟合**”。
- 一般情况下，随着训练能力的提高，预测能力也会提高。但这种趋势并不是固定的，有时当达到某个极限时，随着训练能力的提高，预测能力反而会下降，这种现象称为“**过拟合**”，即训练误差变小，测试误差也会随之减小，然而减小到某个值后，测试误差却反而开始增大。
- 通常，在**训练数据不足**或**模型过于复杂**时，会导致模型在训练集上过拟合。

分类问题中的拟合



(a) Good fit 良拟合　　**(b) under-fit 欠拟合**　　**(c) Over-fit 过拟合**

解决策略:

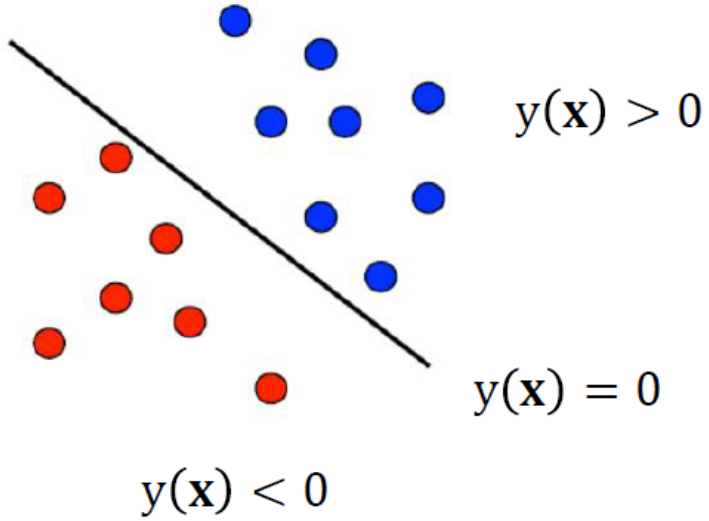
- Early-stopping（早停法）
- Dropout(0.3)(随机失活/抛弃30% 的神经元)
- Data enhancement (数据增强)
- Weight regularization (权重正则化，可以降低模型的复杂度)

4.2.2 监督学习的主要任务

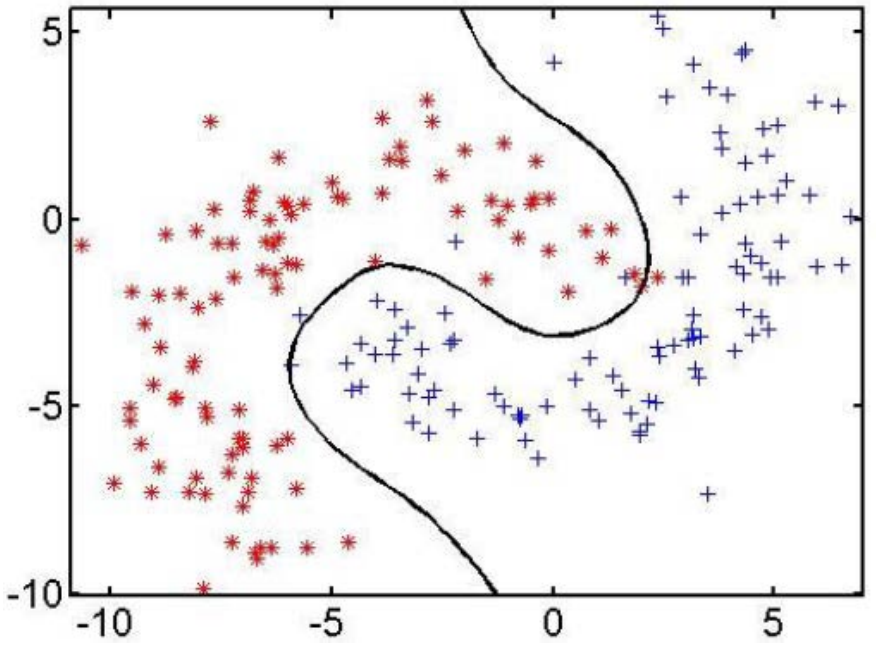
监督学习可以完成的主要任务包括**分类**和**回归**。

1. 分类：输出离散值。可分为：

- 二分类任务：如电子邮件中的垃圾邮件过滤
- 多分类任务：手写体数字识别则



线性分类器：线性函数 $y(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b$



非线性分类器：非线性函数

4.2.2 监督学习的主要任务

2. **回归**：回归主要用于预测数值型数据，它输出的是连续值，而非离散值。

- 回归被广泛应用于各个领域的预测和预报，例如，传染病学中的发病趋势、在经济领域中预测消费支出、固定资产投资支出、持有流动资产需求等。
- 回归又分为：

（1）线性回归

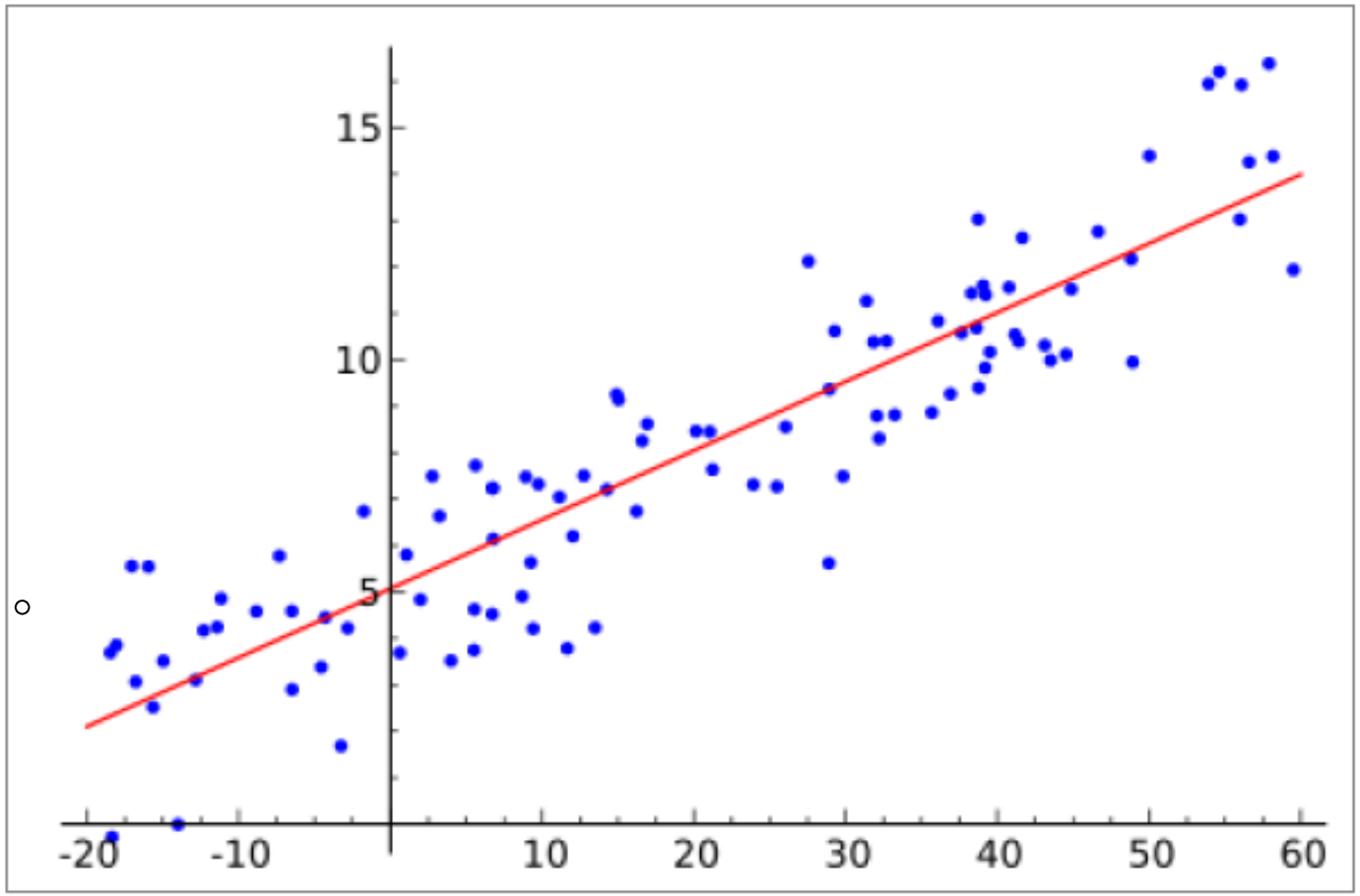
（2）逻辑回归

（1）线性回归

- 线性回归是指采用线性函数来建模，根据已知数据来估计未知的模型参数。
- 线性回归模型：$y = w^T x+e$，其中w^T是模型参数向量的转置，e表示误差。
- 研究一个因变量与一个或多个自变量间多项式的回归分析方法，称为**多项式回归**。
- 如果自变量只有一个时，称为**一元多项式回归**；
- 如果自变量有两个或两个以上时，称为**多元多项式回归**。
- 若一个因变量与一个或多个自变量间是非线性关系，例如，$y(x)= w_2x^2+w_1x+e$，则称为**非线性回归**。
- 在一元多项式回归分析中，若一个自变量和一个因变量的关系可用一条直线近似表示，这种回归称为**一元线性回归**，即找一条直线来拟合数据。
- 如果在多元多项式回归分析中，一个因变量和多个自变量之间是线性关系，则称为**多元线性回归**。

线性回归

- 线性回归中，采用具有如下特征的函数对观测数据进行建模：
  - 该函数是模型参数的线性组合；
  - 该函数取决于一个（一元线性回归）或多个独立自变量（多元线性回归）。

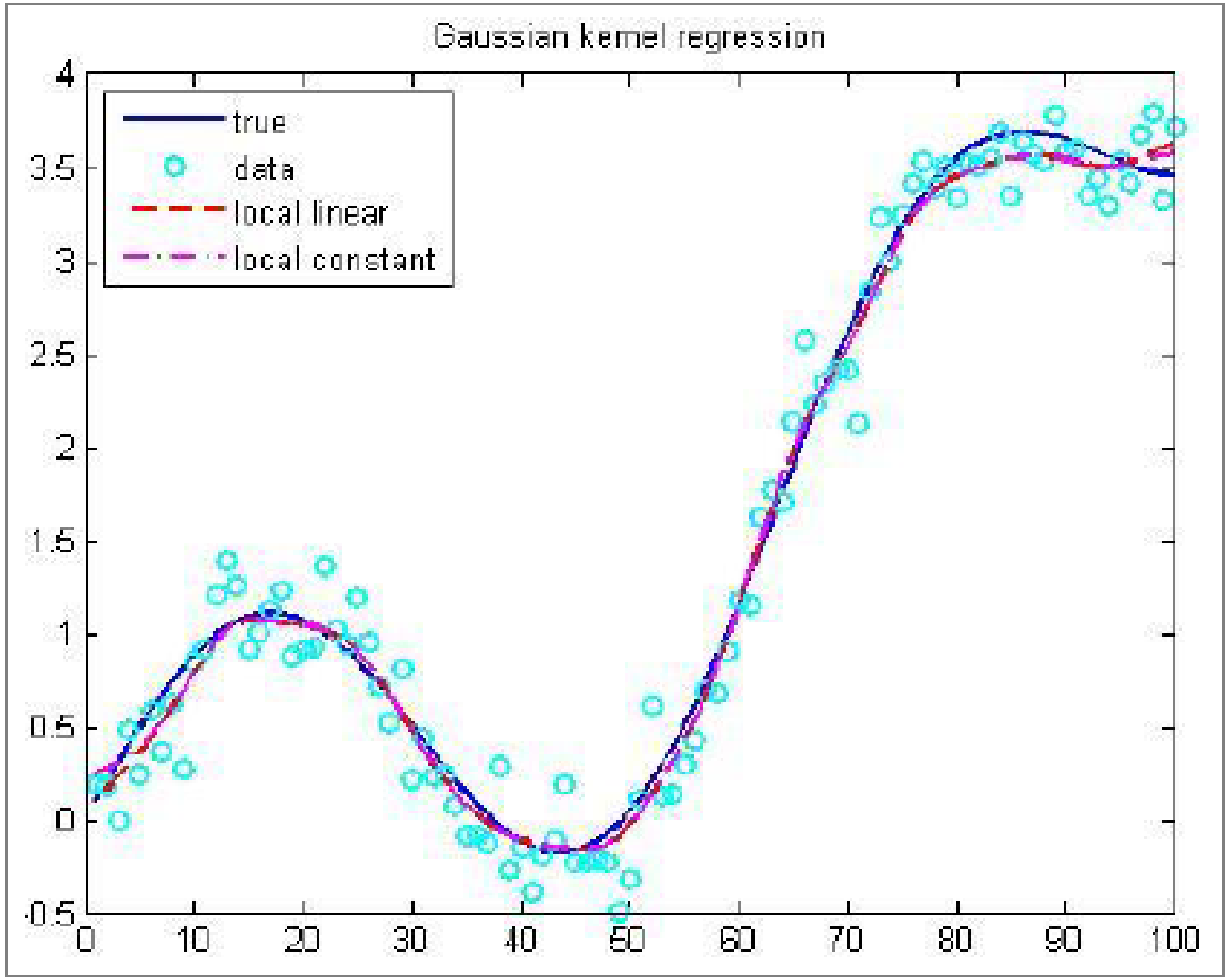


$$y(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b$$

模型表达：$y(x,w)=w_1x_1+\ldots+w_nx_n+b$

非线性回归

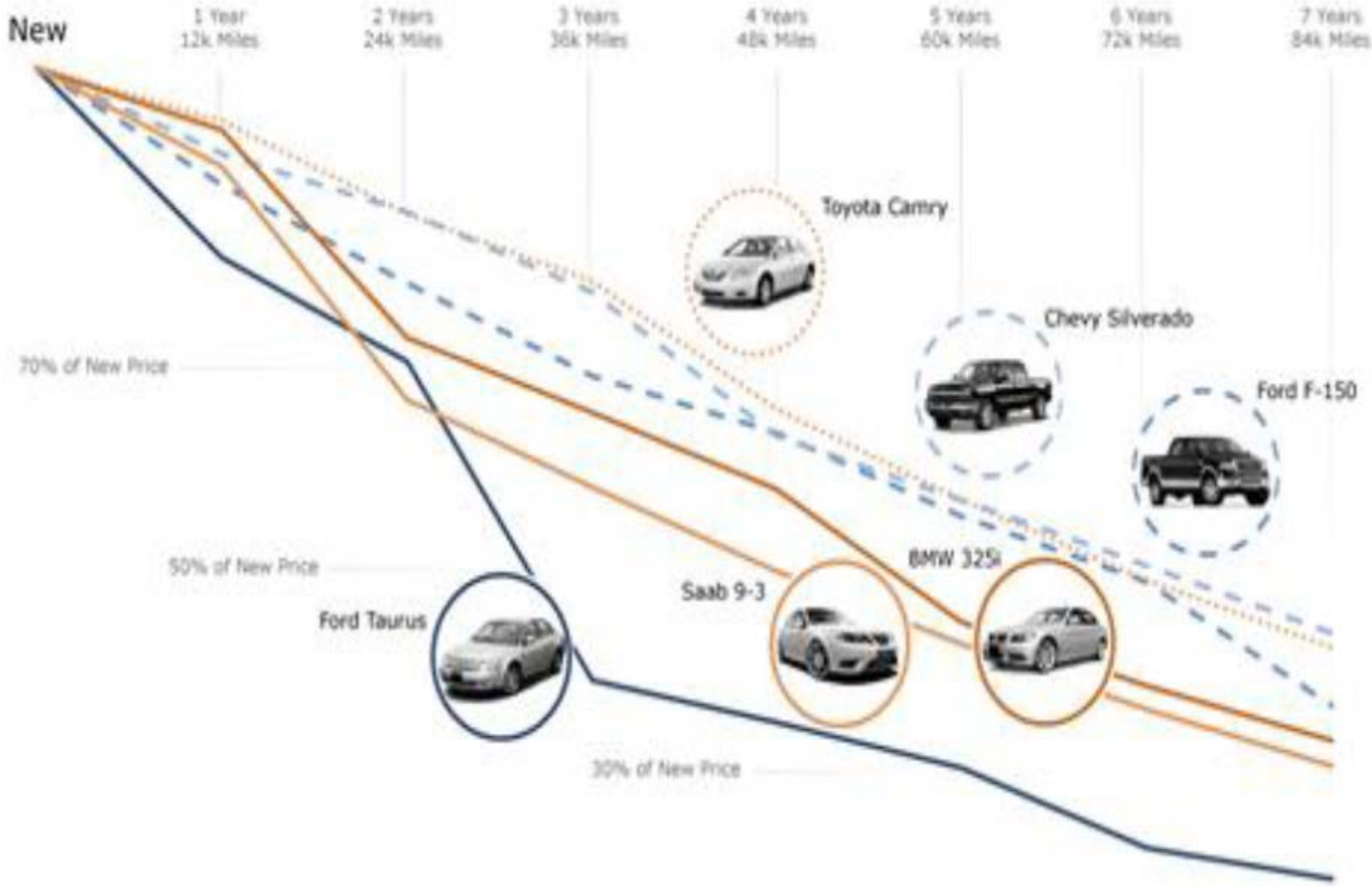
- 非线性回归中，采用具有如下特征的函数对观测数据进行建模：
  - 该函数是模型参数的非线性组合；
  - 该函数取决于一个或多个独立变量。



$$y(x) = w_2 \cdot x^2 + w_1 \cdot x + b$$

线性回归应用：二手车价格预测

- 构建一个预测二手车价格的线性模型。
- 输入是**车的属性**：品牌、年式、引擎功率、里程、以及其它信息。
- 输出是**车的价格**。



（2）逻辑回归

- 逻辑回归又称为**逻辑回归分析**，是通过历史数据的表现对未来结果发生的概率进行预测。
- 尽管逻辑回归输出的是实数值，但本质上它**是一种分类方法，而不是回归方法**。
- 逻辑回归的自变量可以有一个，也可以有多个。
- 有一个自变量的，称为**一元回归分析**;
- 有两个或两个以上自变量的，称为**多元回归分析**。
- 逻辑回归的因变量可以是二分类，也可以是多分类。二分类更为常用，也更容易解释。
- 若采用**sigmoid函数**计算概率，令阈值为0.5，则完成**二分类**任务。

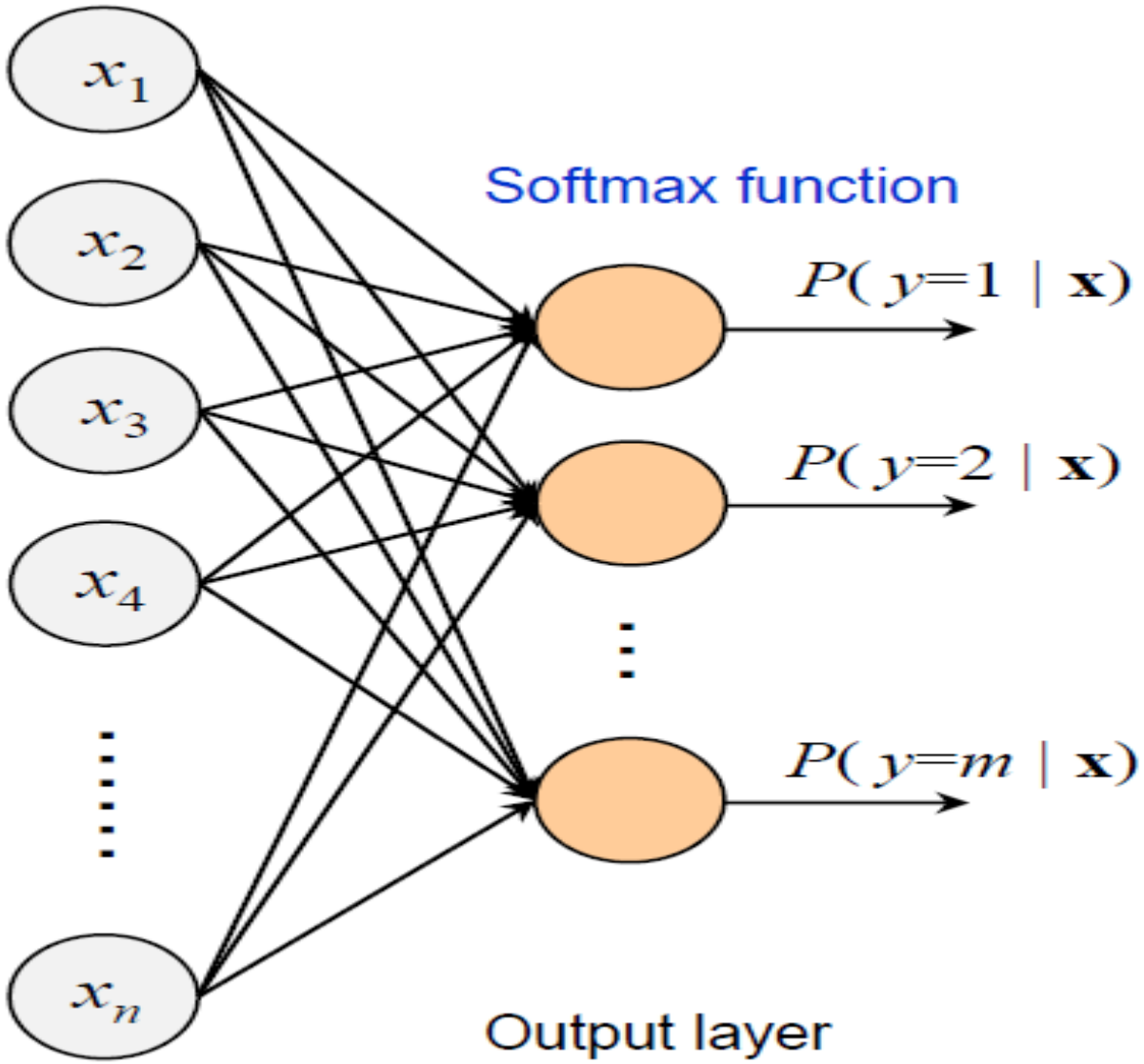
$$f(x)=\frac{1}{1+e^{-x}}$$

- 若采用**softmax函数**计算概率，则逻辑回归可完成**多分类**任务。

$$f(x_i)=\frac{e^{x_i}}{\sum_{k=1}^{M} e^{x_k}} \qquad i=1,\ldots,M \quad ; \quad M\text{为类别数}$$

输出的实数表示未知样本 x 属于某一类别的概率.

Softmax 分类器



Softmax 函数在ANN的最后一层用于多元分类。

线性回归与逻辑回归的异同点

- 线性回归与逻辑回归的**区别**在于：
  - 线性回归用于**预测连续值**，其输出的值域是实数集，其模型是线性的；
  - 逻辑回归主要用于**解决分类问题**，其输出的值域为[0,1]，其模型是非线性的。
- 线性回归与逻辑回归的**共同点**在于：

  两者的输入数据既可以是连续的值，也可以是离散的值。

4.2.3 监督学习的典型算法

4.2.3.1 K-近邻算法

◆ K-近邻算法：理论上比较成熟、设计思想直观、最简单的机器学习算法之一。

◆ KNN算法**既可用于分类问题，也可用于回归问题**。

◆ KNN算法完成**分类**任务时的流程：

对数据集中每个未知类别的样本依次执行以下操作。

①准备好已知类别标签的数据集D，对数据进行预处理，假设D中共有N个数据样本。

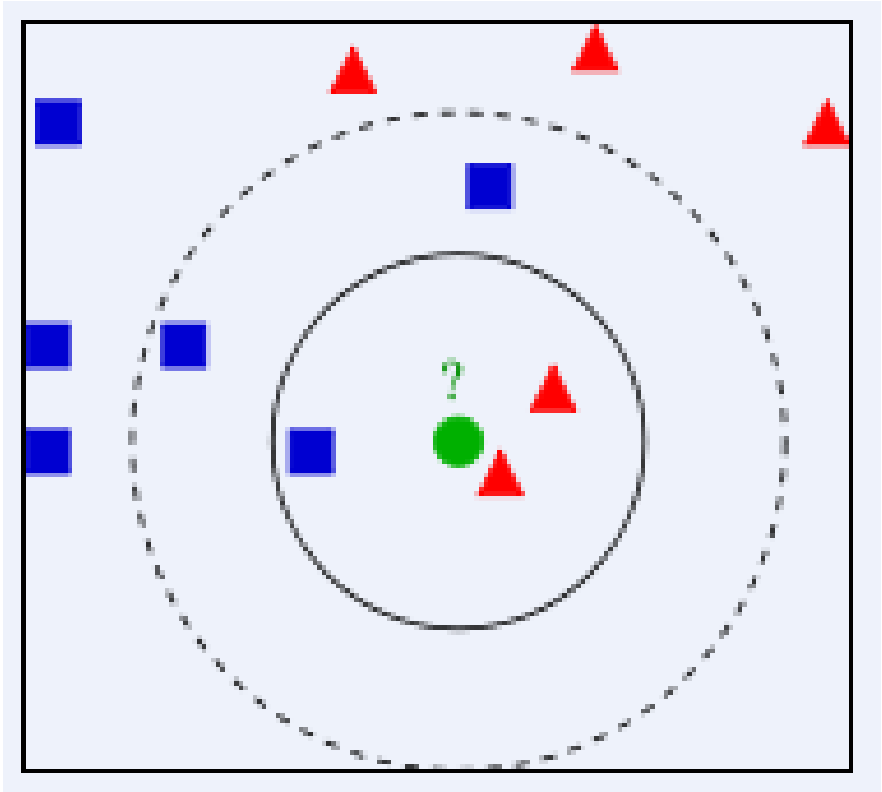
②计算测试样本x（即待分类样本）到D中每个样本的距离，存入数组Dist[1..N]。

③对Dist[1..N]中元素进行增序排序，找出其中距离最小的K个样本。

④统计这K个样本所属类别出现的频率。

⑤找出出现频率最高的类别，记为c，作为测试样本x的预测类别。

K-Nearest Neighbor （KNN） 算法



问题：如何确定 k 值？这是个经验值

K-近邻算法

◆KNN算法完成回归任务时的流程。

对数据集中每个未知属性值的样本依次执行以下操作。

①准备好已知属性值的数据集D，对数据进行预处理，假设D中共有N个数据样本。

②计算测试样本x（即待预测样本）到D中每个样本的距离，存入Dist[1..N]。

③对Dist[1..N]中元素进行增序排序，找出其中距离最小的K个样本。

④计算这K个样本的属性的平均值Y。

⑤将Y赋予测试样本x，作为其属性值。

KNN算法的优缺点

◆ **KNN算法的优点**有：

- 思路简单，易于理解，易于实现，无需训练过程，不必估计参数，可直接用训练数据来实现分类。
- 只需人为确定两个参数，即K的值和距离函数。KNN算法支持多分类。

◆ **KNN算法有以下4点不足**：

① 对超参数K的选择十分敏感，针对同一个数据集和同一个待测样本选择不同K值，会导致得到完全不同的分类结果。

② 当样本不平衡时，例如一个类中样本数很多，而其他类中样本数很少，有可能导致总是将新样本归入大容量类别中，产生错误分类。

③ 当不同类别的样本数接近时或有噪声时，会增加决策失误的风险。

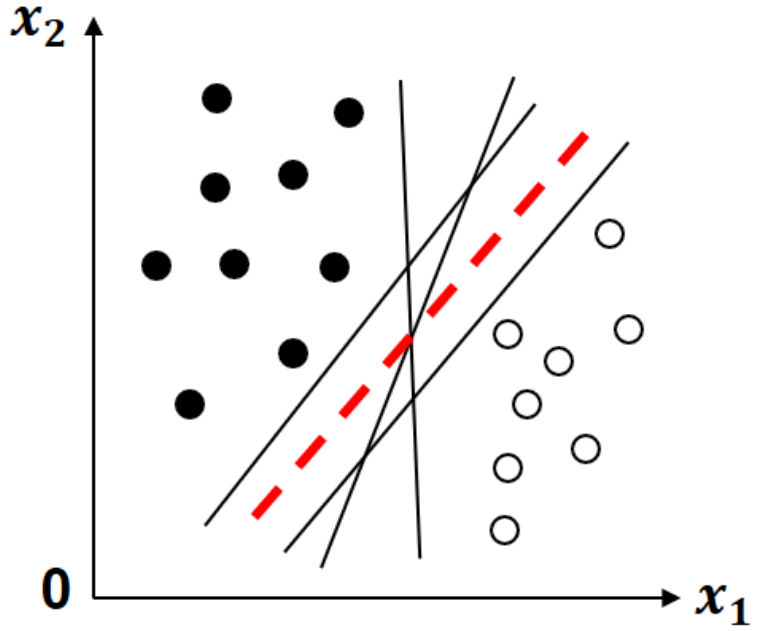
④ 当样本的特征维度很高或训练数据量很大时，计算量较大，KNN算法效率会降低，因为对每一个待分类样本都要计算它与全体已知样本之间的距离，才能找到它的K个最近邻点。目前常用的解决方法是事先对已知样本进行剪枝，去除对分类作用不大的样本；另外，可以对训练数据进行快速K近邻搜索。

4.2.3.2 支持向量机

◆ SVM是机器学习中的一个经典算法，具有严格的理论基础，尤其**在样本量小的情况下表现出色**。

◆ **SVM既可用于分类问题，也可用于回归问题**。

◆ SVM的**设计思路**

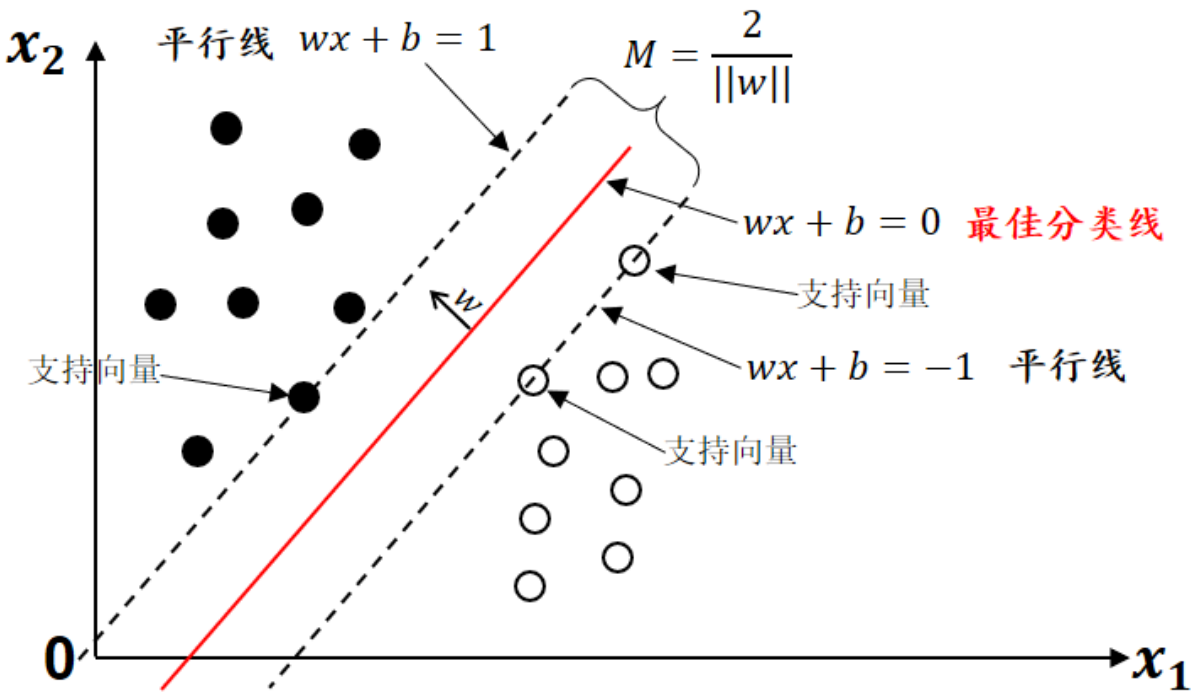


- 给定训练数据集，支持向量机将每个训练样本的特征向量表示为空间中的点，支持向量机想要求解一个分类超平面，使得不同类别的样本被尽可能大间隔地分开。
- 然后将新的样本映射到同一空间，根据它落在分类超平面的哪一侧来预测其所属的类别。

◆ SVM是一种二分类模型，其基本模型定义为特征空间上间隔最大的线性分类器，即SVM的学习策略就是使间隔最大化，最终可转化为一个凸二次规划问题的求解。

4.2.3.2 支持向量机

- 一般而言，一个样本点距离超平面的远近可以表示分类预测的置信度或准确程度。
- 当一个样本点距离超平面越远时，分类的置信度越大。
- 对于一个包含n个样本点的数据集，自然认为：其分类间隔是n个点中离超平面最近的距离。
- 为了提高分类的准确程度，希望所选择的超平面能够最大化该间隔。此为 SVM算法最朴素的思路，即**最大间隔**（Max-Margin）**准则**。
- 距离超平面最近的若干个训练样本被称为“**支持向量**”，两个异类支持向量到超平面的距离之和被称为“**间隔**”。



4.2.3.2 支持向量机

- 支持向量机的目标就是找到具有“最大间隔”的超平面。
- 在2D空间中，分类线可用线性方程wx+b=0来描述，其中，w和b都是待优化的参数。
- 分类线 wx+b=0 应该将所有训练样本都正确分类，则有如下表达式：

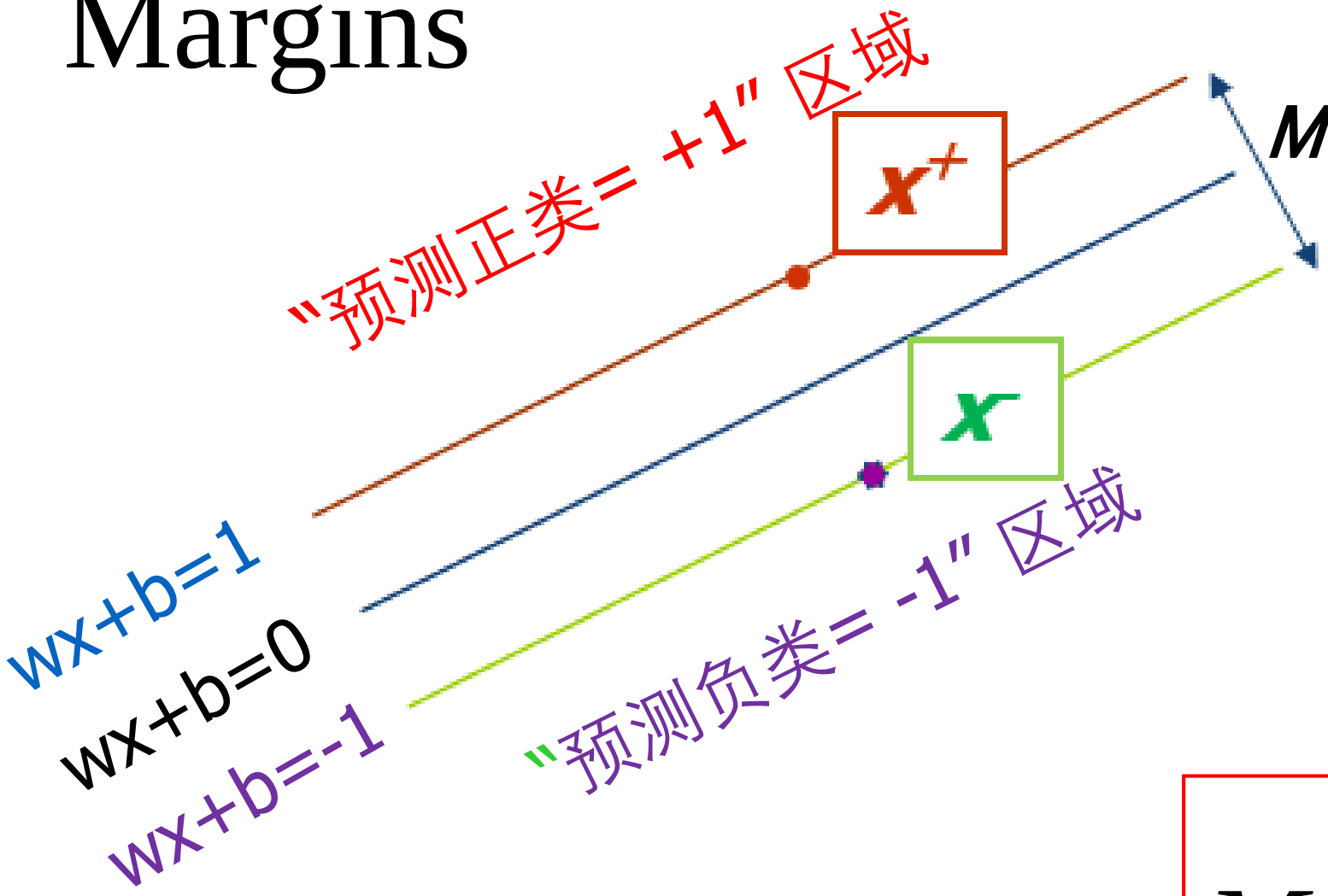
$$\begin{cases} wx_i + b \geq +1, & if \ y_i = +1 \\ wx_i + b \leq -1, & if \ y_i = -1 \end{cases}$$ 合并，写为： $$y_i(wx_i + b) - 1 \geq 0$$

  - 两条与分类线平行的虚线的方程分别为： wx+b=1 和 wx+b=-1
  - 这**两条平行虚线之间的距离**称为“**间隔**”，其计算公式为： M=2/||w||
  - 希望间隔M取最大值，即有：$\max \ M = \frac{2}{\|w\|} \Rightarrow \min_{w,b} \frac{1}{2}\|w\|^2$

$$s.t. \quad y_i(wx_i + b) \geq 1, \quad i = 1,2,\cdots,N \ ; \ N\text{为样本数}$$

  - 找“最大间隔”，转化为求解二次优化问题，求参数w和b，使得M最大。

Margins



M=间隔得宽度

$$\| x \|_\infty = \max_i | x_i |, (l_\infty 范数)$$

$$\| x \|_1 = \sum_{i=1}^{n} | x_i |, (l_1 范数)$$

$$\| x \|_2 = (\sum_{i=1}^{n} x_i^2)^{1/2}, (l_2 范数)$$

$$\| x \|_p = (\sum_{i=1}^{n} | x_i |^p)^{1/p}, (l_p 范数, 1 \le p < \infty)$$

间隔的计算公式 $$M = \frac{2}{\| w \|}$$

两条平行线之间的距离公式

两条平行线：$Ax+By+c1=0$

$Ax+By+c2=0$

其距离公式：$\frac{|C_1-C_2|}{\sqrt{A^2+B^2}}$

两条平行线公式：$wx+b-1=0$ and $wx+b+1=0$

距离：（c1=b-1，c2= b+1）：

$$M=\frac{2}{\|w\|}$$

目标函数

◆ 分类平面将所有数据都正确分类:

$$w \cdot x_i + b \geq 1 \quad if \ \ y_i = +1$$

$$w \cdot x_i + b \leq -1 \quad if \ \ y_i = -1$$

$$y_i (w \cdot x_i + b) - 1 \geq 0$$

◆ 最大化间隔:

$$\max M = \frac{2}{\|w\|} \Rightarrow \min \frac{1}{2} w^T w$$

◆ **二次优化问题**

- Minimize

$$\Phi(w) = \frac{1}{2} w^T w$$ **求使之最小的向量w**

- Subject to

$$y_i (w \cdot x_i + b) \geq 1$$ **前提条件是必须分类正确**

SVM的优缺点

◆ SVM的**优点**如下。

① 解决高维特征的分类问题和回归问题很有效，在特征维度大于样本数时，依然有很好的效果。

② 仅仅利用为数不多的支持向量便可确定超平面，无需依赖全部数据，因此适用于小样本集的应用。

③ 有大量的核函数可以使用，从而可灵活地解决各种非线性的分类和回归问题。

④ 在样本量不是海量数据时，分类准确率高，泛化能力强。

◆ SVM的**缺点**如下。

① 如果特征维度远远大于样本数，则SVM表现一般。

② 在样本量巨大、核函数映射维度非常高时，SVM的计算量过大。

③ 针对非线性问题的“核函数选择”问题，没有通用标准，难以选择一个合适的核函数。

④ SVM对缺失数据敏感。

4.3 无监督学习

4.3.1 无监督学习的基本原理

4.3.2 无监督学习的主要任务：

1. 聚类（Clustering）

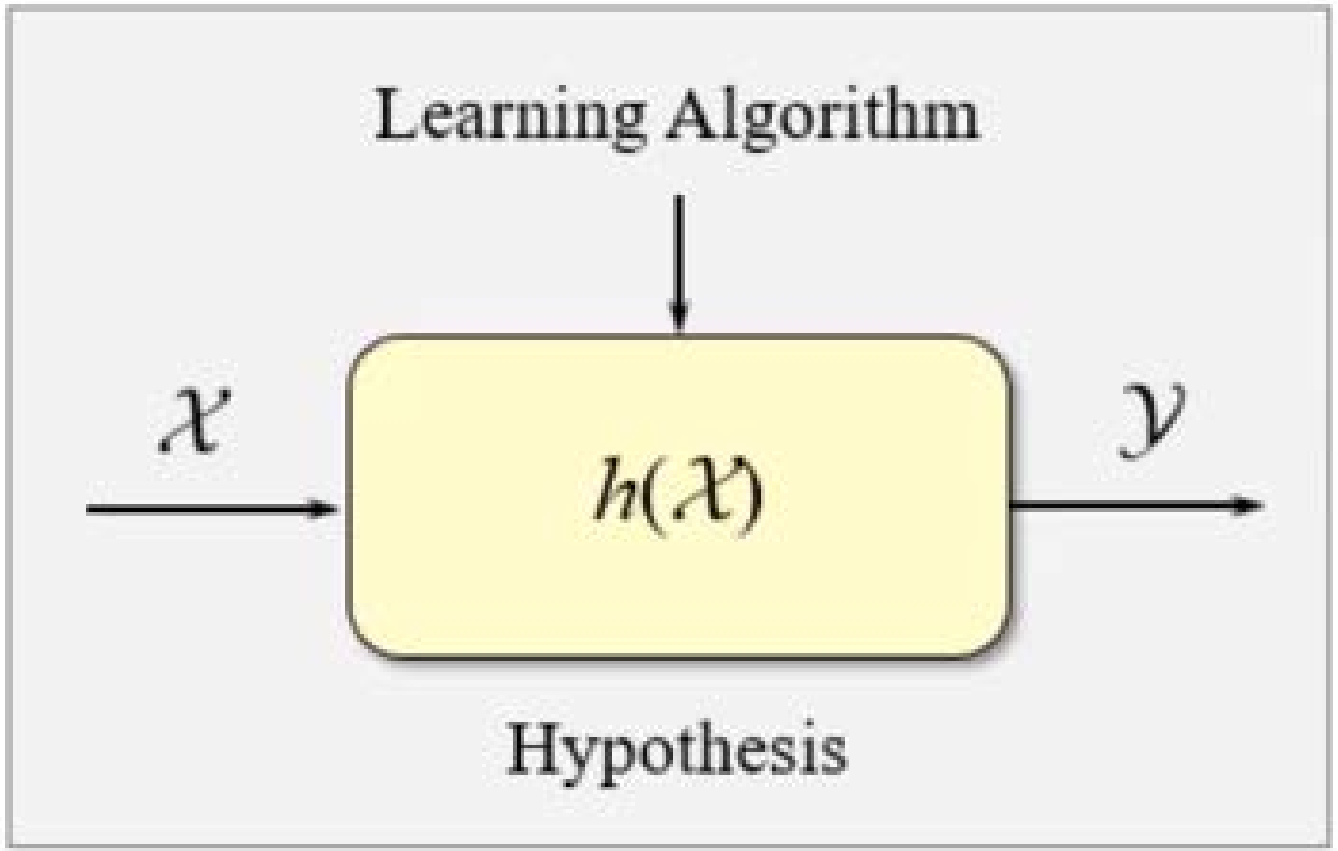
2. 降维（Dimensionality Reduction）

4.3.3 无监督学习的典型算法

4.3.3.1 K-Means

4.3.3.2 主成分分析（PCA）

无监督学习示意图



4.3 无监督学习

- 无监督学习，顾名思义，即没有监督的学习。
- 与监督学习的**不同之处**在于：无监督学习方法在学习时，没有事先指定的标签，也不需要人类提供标注数据，更不接收监督式信息（即告诉它何种操作是正确的）。
- 无监督学习是通过模型不断地自我认知、自我巩固、自我归纳来实现其学习过程。

It is a way of "teaching by itself", without a "teacher".

这是一种"自学"的方式，没有"老师"。**无师自通**

4.3.1 无监督学习的基本原理

- 无监督学习算法仅接收环境提供的未标注数据，从中学习**数据的统计规律**或者**内在结构**，调节自身的参数或结构，以发现外部输入的某种固有特性。
- **无监督学习没有训练过程**，其学习的**基本原理**如下。
  - 已知N个数据样本$X=\{X_1, X_2,\ldots, X_N\}$，第$i$个样本$X_i$表示为$(x_{i1}, x_{i2}, \cdots, x_{in})$。
  - 将无标注的样本输入到一个指定的学习模型中，学习得到用于聚类、降维或概率估计的决策函数 $f_\theta(x)$ 或条件概率分布 $P_\theta(y|x)$ 或者 $P_\theta(x|y)$ ，
  - 继而可以使用决策函数或条件概率分布式对**未知数据**进行预测和概率估计。
- 无监督学习的**主要任务**包括：**聚类、降维、密度估计**。
- 无监督的**学习模型**则为一组描述聚类、降维或概率估计的计算方法。
- 无监督学习的过程就是从模型集合中选择出最优预测模型的过程，预测过程是获取聚类、降维或概率估计结果的过程。

4.3.2 无监督学习的主要任务

无监督学习可以完成的主要**任务**：**聚类**和**降维**。

1. **聚类**：聚类是研究样本分组问题的一种统计分析方法。

- ◆ 聚类起源于分类学，但却不等于分类。
- ◆ 聚类与分类的**不同**在于：聚类所划分的组是未知的，是从样本中通过学习自动发现的，各个组的标签是未知的，但组的个数需要事先人为给定。
- ◆ 给定一个由无标注样本组成的数据集，**聚类**是根据数据特征的差异将样本划分为若干个组（簇），使得同组内的样本非常相似，不同组的样本不相似。
- ◆ **聚类的目的**是将相似的对象聚在一起，却不知道组中的对象是什么，更不知道组的名称。
- ◆ 因此，只需要确定样本相似度的计算方法，便可执行一个聚类算法了。

4.3.2 无监督学习的主要任务

- 聚类是将样本集分为若干互不相交的子集，即样本组。
- 聚类算法**划分组的原则**为：
  - 使划分在同一组的样本尽可能地彼此相似，即**类内的相似度高**；
  - 同时使划分在不同组的样本尽可能地不同，即**组间的相似度低**。
- 聚类的典型应用包括：商业营销、图像分割、经济区域分类。
- **传统的聚类分析计算方法**主要有如下几类：划分方法、层次聚类方法或基于连接的聚类方法、基于密度的聚类方法、基于网格的方法、基于模型的方法。

4.3.2 无监督学习的主要任务

2. 降维：通过数学变换，将原始高维特征空间转变为一个低维子空间的过程，称为降维。

- 降维也称为维数约简，即降低特征的维度，得到一组冗余度很低的特征。
- 若原特征空间是D维的，现在希望降至D-1维甚至更低维的。
- 特征维数由1000D降至100D用，用最具代表性的100D特征代替原来1000D特征。
- 在从高维空间转换到低维空间的过程中，低维空间不是事先给定的，而是从样本数据中自动发现的，但低维空间的维度通常是事先给定的。
- 降维后，新产生的特征的物理含义也需由学者自己去发现和总结。
- **需要降维的原因**在于：在原始高维数据空间中，样本特征可能包含冗余信息及噪声信息，在实际应用中会造成存储空间的浪费、计算量过大、维度灾难、无法获取本质特征等问题。

降维主要应用于3方面

(1) **数据压缩**，压缩后的图像、视频、音频数据不仅减少了占用计算机内存或磁盘的空间，还加速了各种算法的运行；

(2) **数据可视化**，通过降维可以得到更直观的数据视图。例如，将四维甚至更高维的数据降至二维或三维空间上，使得对数据的结构有更直观的理解与认识；

(3) **特征工程**，高维数据中的冗余特征和噪声会对模式识别造成误差，降低模型的准确率，增加模型的复杂度，导致模型出现过拟合现象。通过特征降维，可以去除部分不相关或冗余的特征，确保特征之间是相互独立的，以达到提高模型精确度、降低算法时间复杂度、提升模型泛化能力的目的。

特征工程中的降维

通常有两种策略，分别是**特征选择**和**特征提取**。

（1）特征选择。

- 特征选择也称为**特征子集选择，或属性选择、变量选择或变量子集选择**，它是指从原始特征集（n维特征）中选择出k维（k<n）最有效的、最具代表性的**特征子集**，舍弃冗余或无关的n-k个特征，使得系统的特定指标最优化。
- **特征选择的目的**是去除原始特征集中与目标无关的特征，保留与目标相关的特征。
- **特征选择的输出**可能是原始特征集的一个子集，也可能是原始特征的加权子集，保留了原始特征的物理含义。
- 常用的特征选择方法：高相关性滤波、随机森林、过滤方法等。

特征工程中的降维

（2）特征提取

- 图像处理等领域中的特征提取与降维中的特征提取是完全不同的概念。
- 降维中的特征提取是指通过数学变换方法，将高维特征向量空间映射到低维特征向量空间。
- 实际上，特征提取就是一个对已有特征进行某种变换，以**获取约简特征**的过程。
- 其**思路**是：将原始高维特征空间中的数据点向一个低维空间做投影（例如，从3D向2D做投影），以减少维数。
- 在此映射过程中，特征发生了根本性变化，**原始特征消失了，取而代之的是尽可能多地保留了相关特性的新特征，这些新特征不在原始特征集中**。
- 特征提取的主要**经典方法**：主成分分析法和线性判别分析法等。

特征选择与特征提取的异同点

◆ 特征选择与特征提取的**共同点**：

（1）两者都是在尽可能多地保留原始特征集中有用信息的情况下降低特征向量的维度。

（2）两者都能提高模型的学习性能，降低计算开销，并提升模型的泛化能力。

◆ 特征选择与特征提取的**区别**：

（1）特征选择是从原始特征集中选择出子集，两个特征集之间是包含关系；而特征提取则是将特征向量从原始空间映射到新的低维空间，创建了新特征，两个特征集之间是一种映射关系。

（2）特征选择的子集未改变原始的特征空间，且保留了原始特征的物理意义和数值；而特征提取获得的新特征没有了物理含义，其值也发生了改变。故特征选择获得的特征具有更好的可读性和可解释性。

（3）两种降维策略所采用的方法不同。特征选择常用的方法包括高相关性滤波、随机森林、过滤方法等；特征提取常用的方法包括主成分分析法和线性判别分析法等。

4.3.3 无监督学习的典型算法

4.3.3.1 K-Means聚类

◆ K-Means（即K均值）聚类是无监督学习中使用最广泛的聚类算法，无论是思想还是实现都比较简单。

◆ K-Means聚类算法的**基本思想**是:

- 针对给定的样本集合D，通过迭代寻找K个簇的一种划分方案，其**目标是使事先定义的损失函数最小**。
- 损失函数往往定义为各簇内各个样本与所属**簇中心点**的距离平方之和， 其公式:

$$E = \sum_{i=1}^{K} \sum_{X \in c_i} \|X - \mu_i\|_2^2$$

其中，K是簇的个数，μ_i是簇c_i的中心点；X是簇c_i中某个样本的特征向量。

E值越小，表示簇内样本相似度越高。

K-Means 算法的实现过程

(1) 从数据集D中随机选择K个样本作为初始簇的中心。

(2) 计算每个样本与K个簇中心的距离，并将该样本划分到距离其最近的簇中。

(3) 重新计算K个新簇的中心（即该簇内所有数据点的平均值）。

(4) 重复执行第(2)、(3)步，直到E值不再变小，即所有簇中的样本不再发生变化。

从上述过程可见，K-Means算法时间复杂度近于线性，适合于大规模数据集上的聚类。

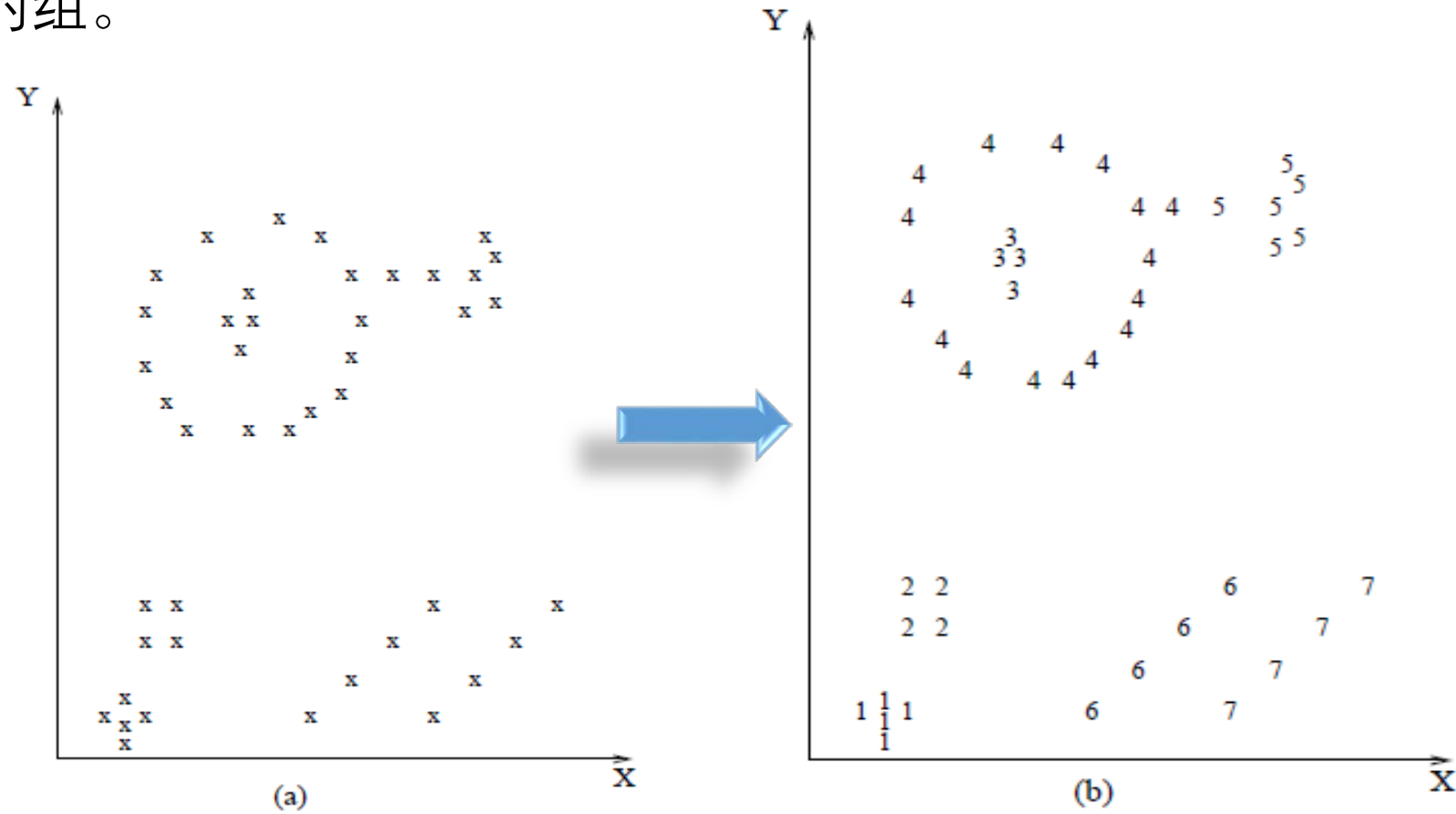
聚类分析的示意图

◆对样本进行分组

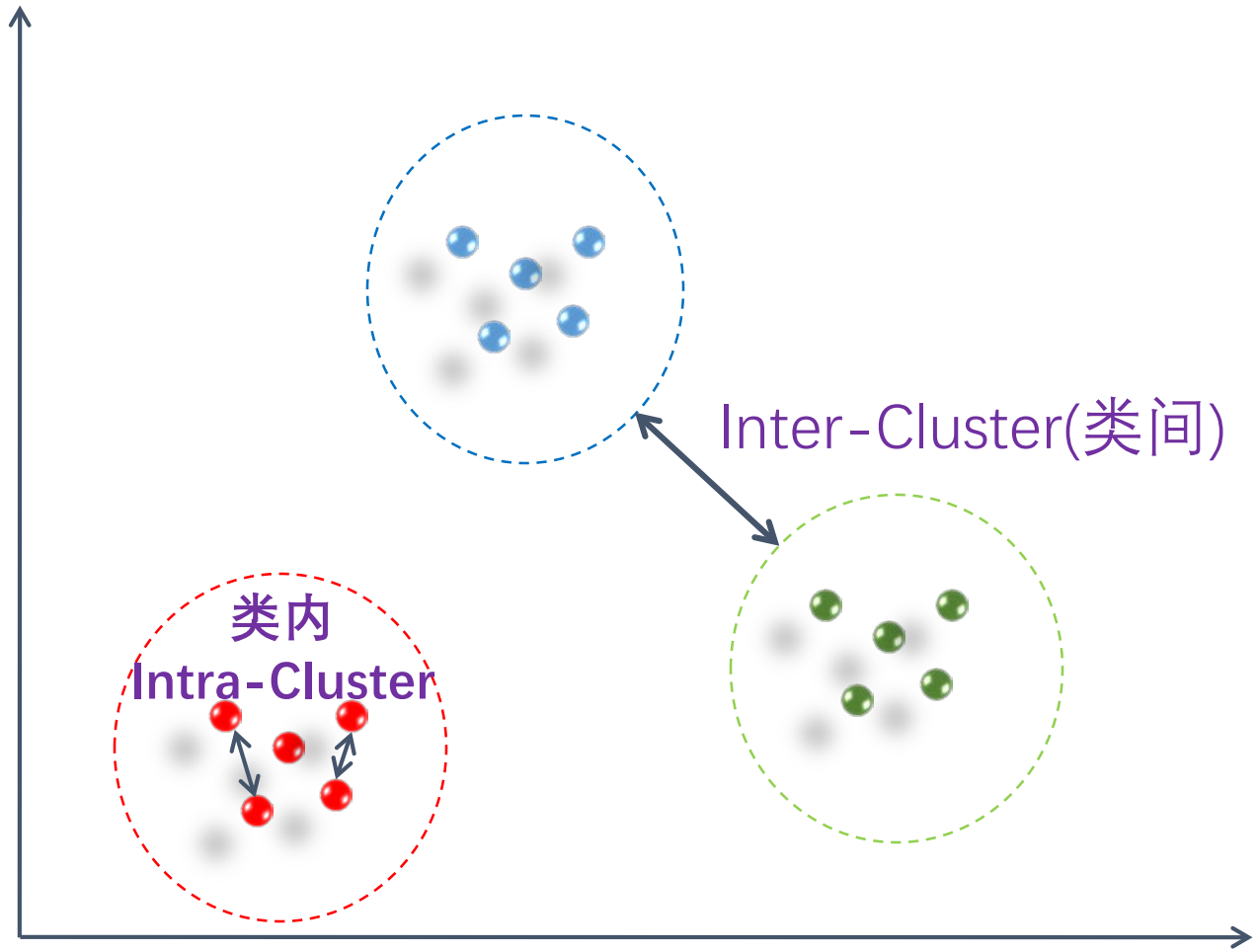
- 相似的样本被分在同一组。
- 不同的样本被分在不同的组。

◆无监督

- 样本没有标签
- 由数据驱动的

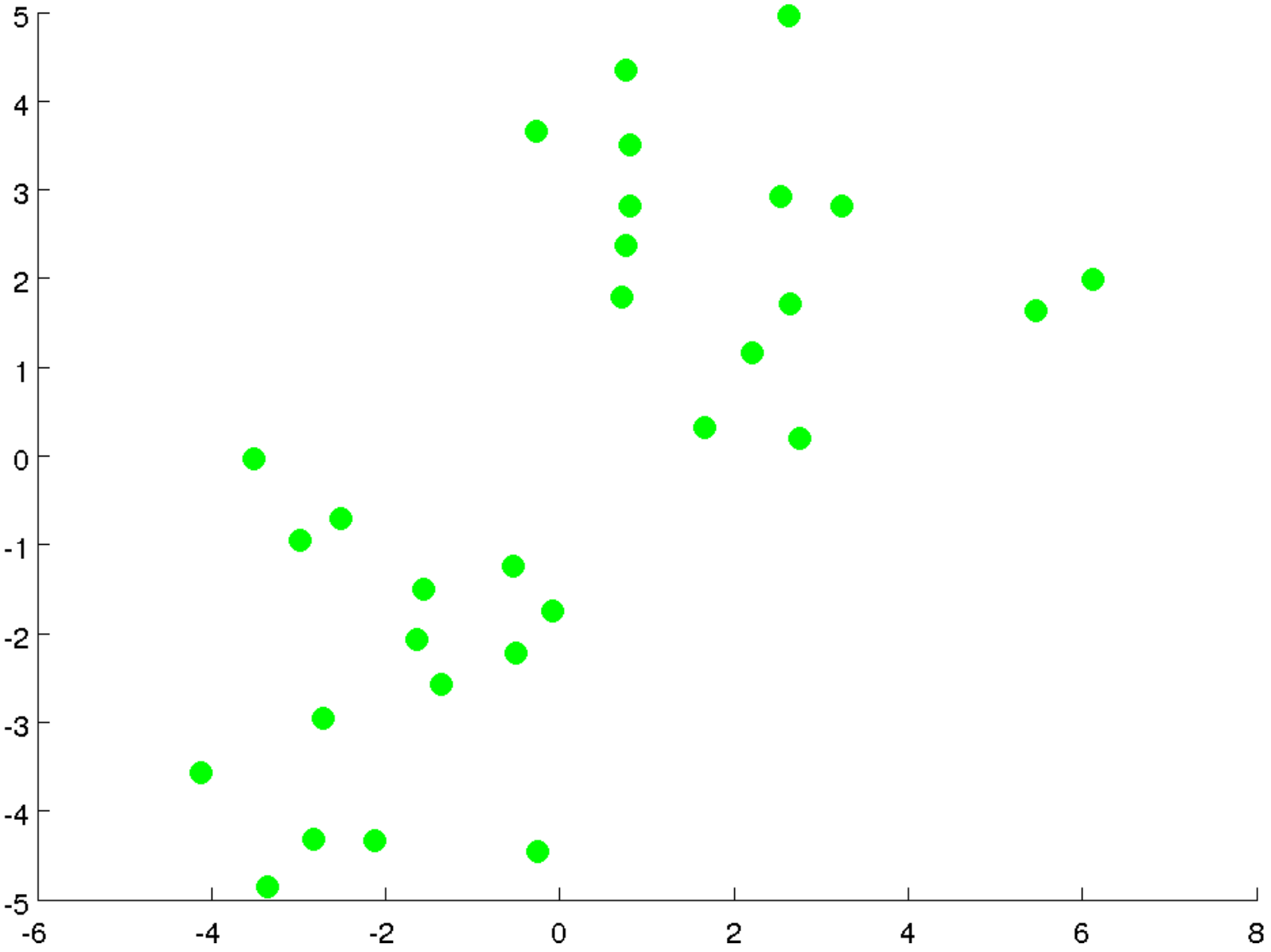


Clusters

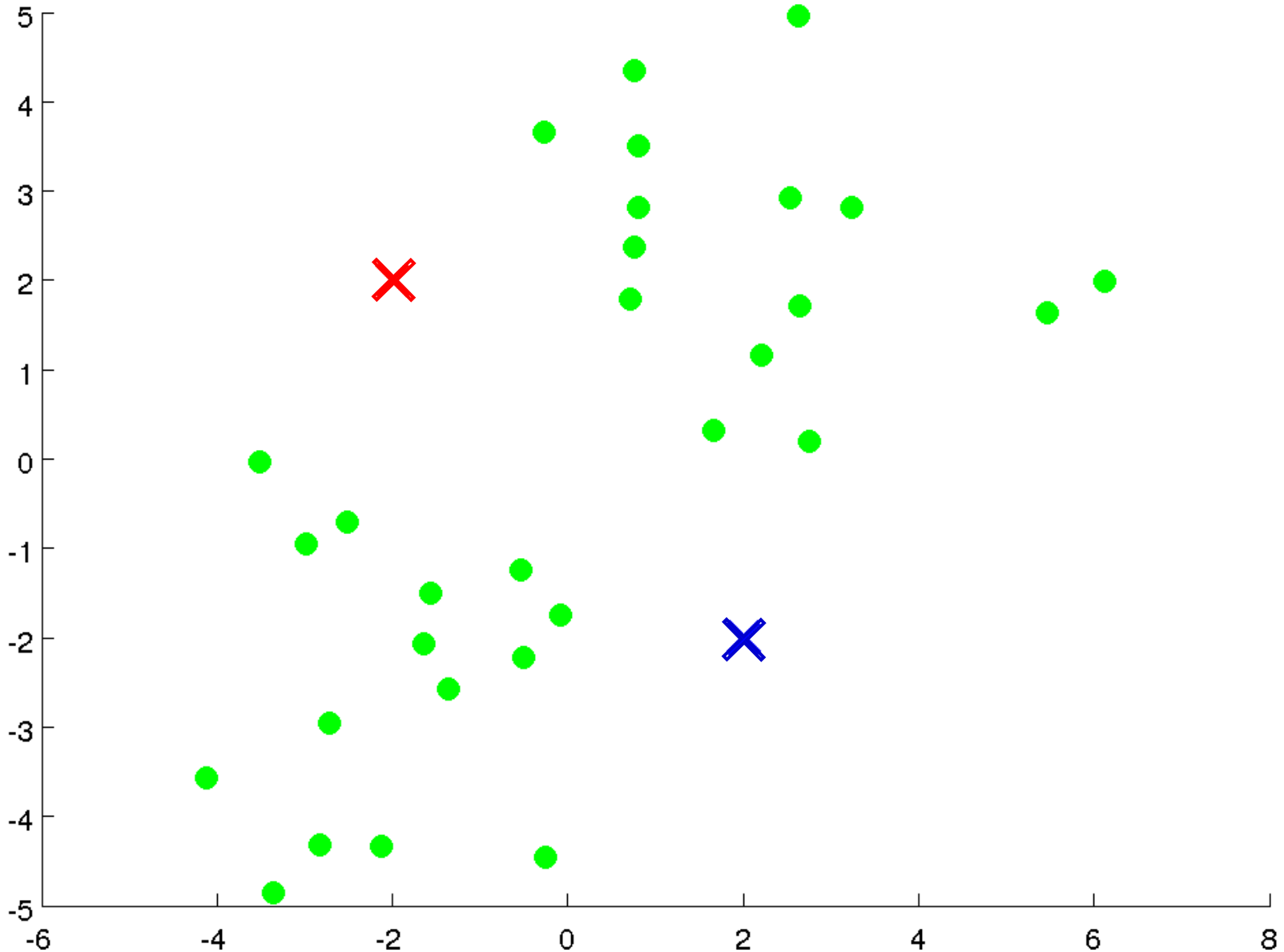


K-means算法执行过程示意图

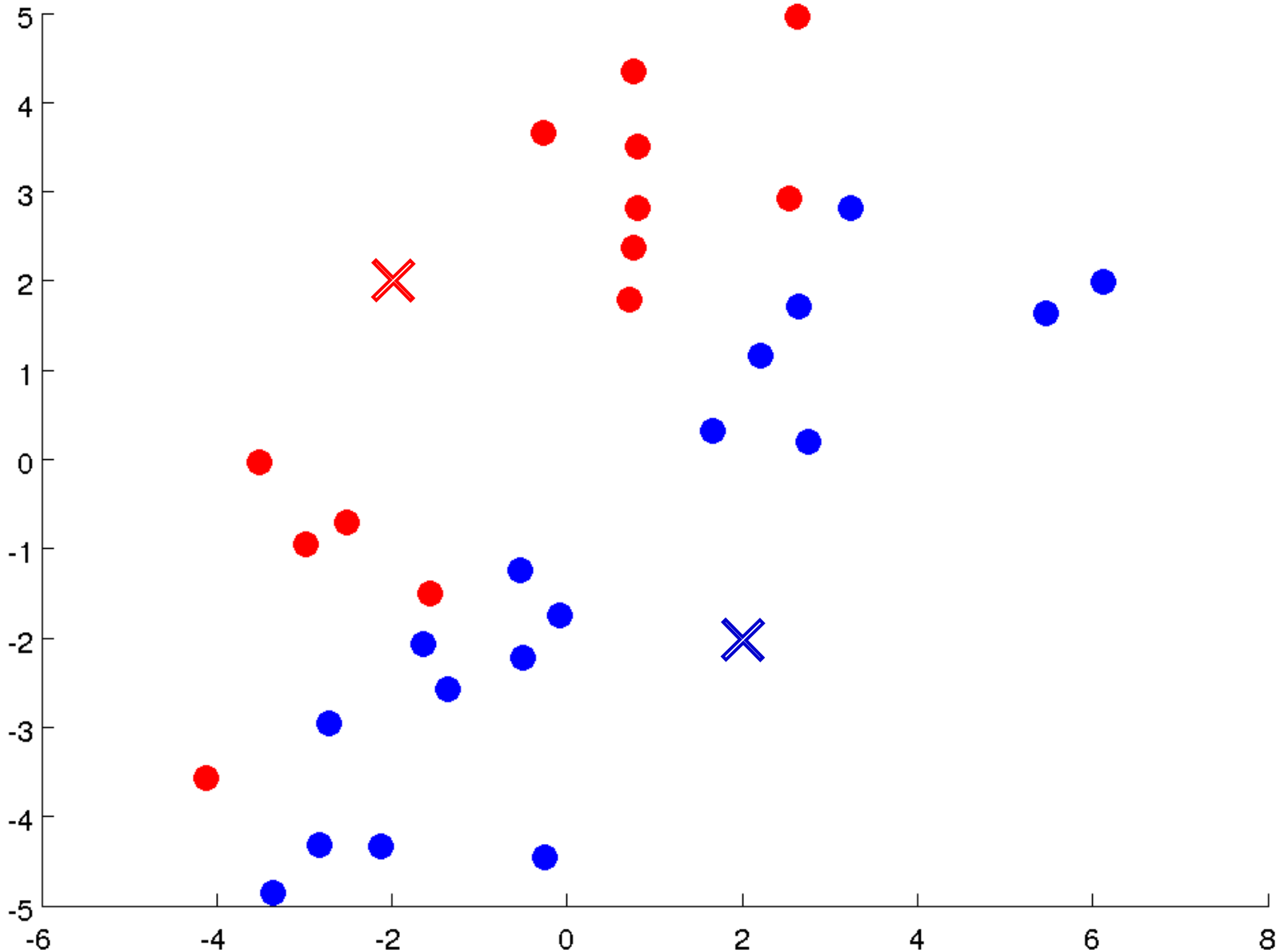
K=2



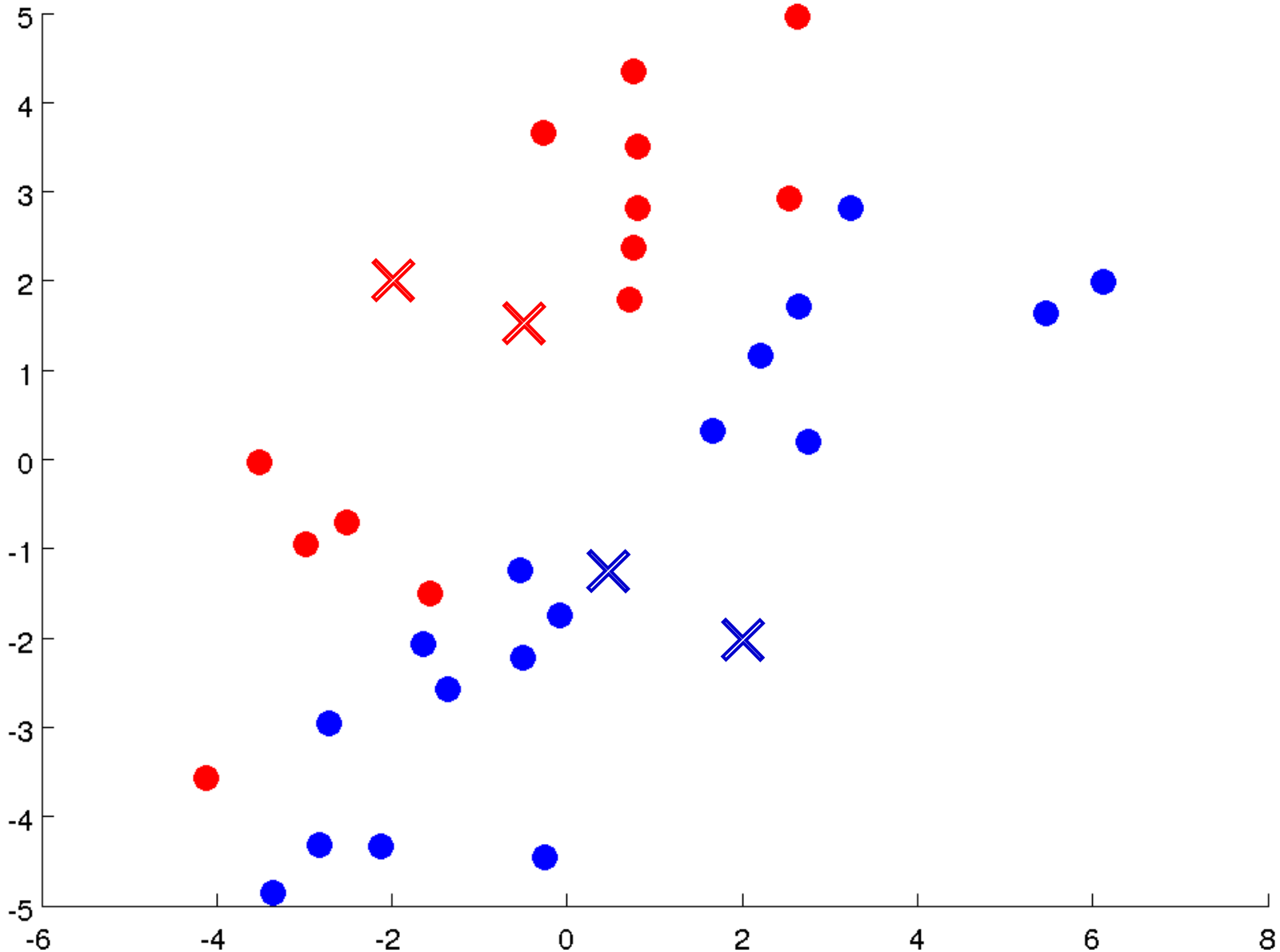
K=2



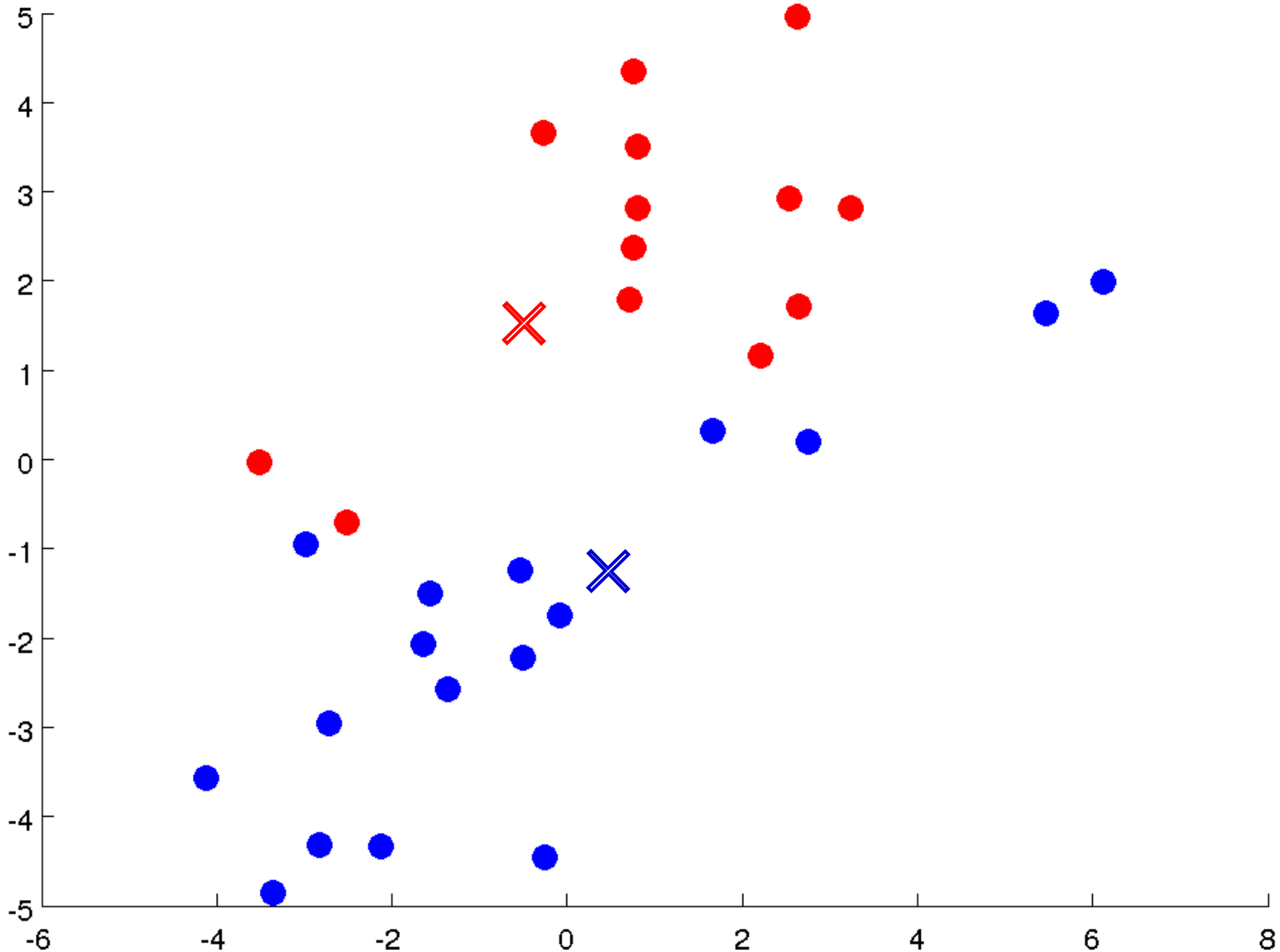
K=2



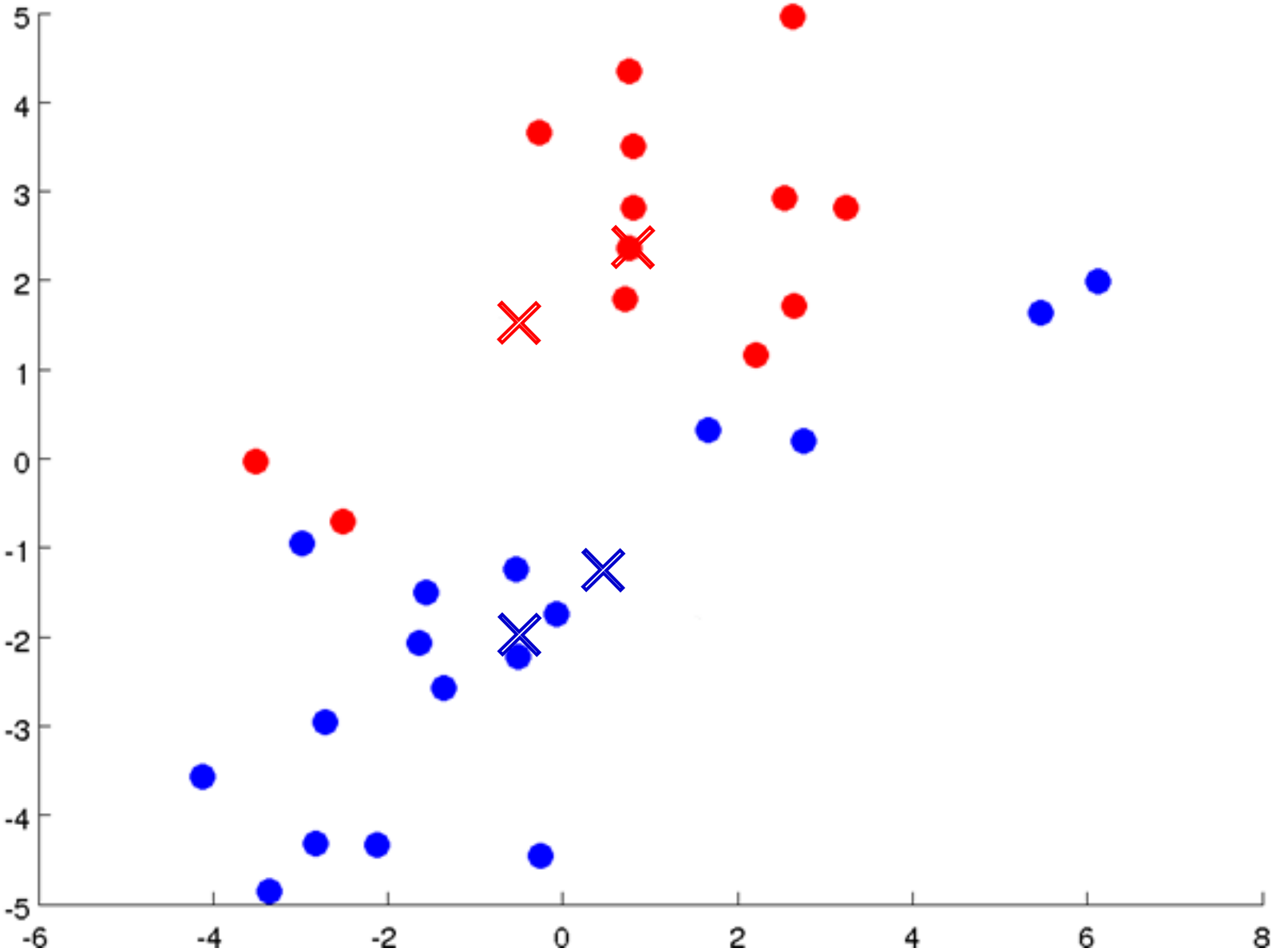
K=2



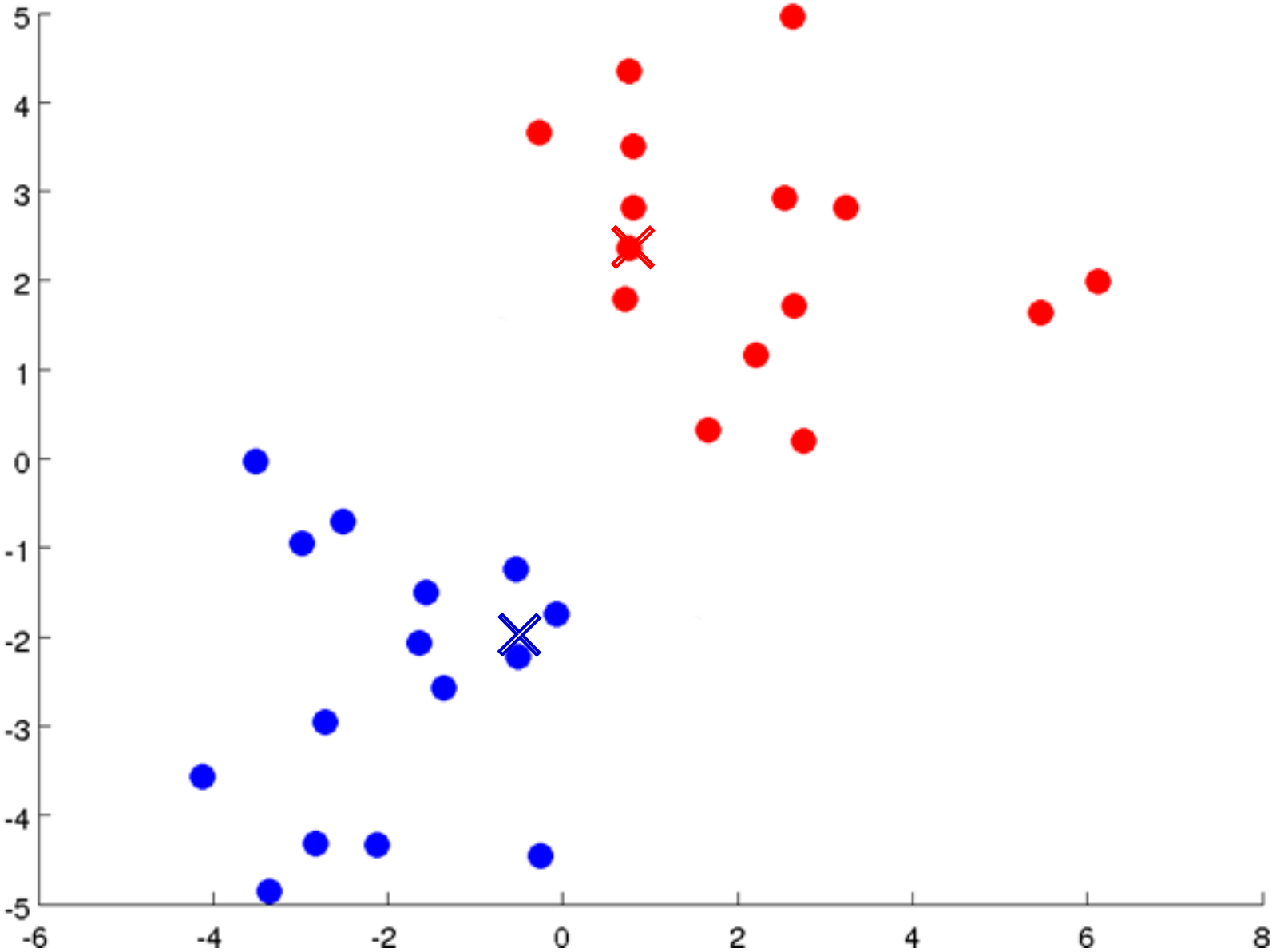
K=2



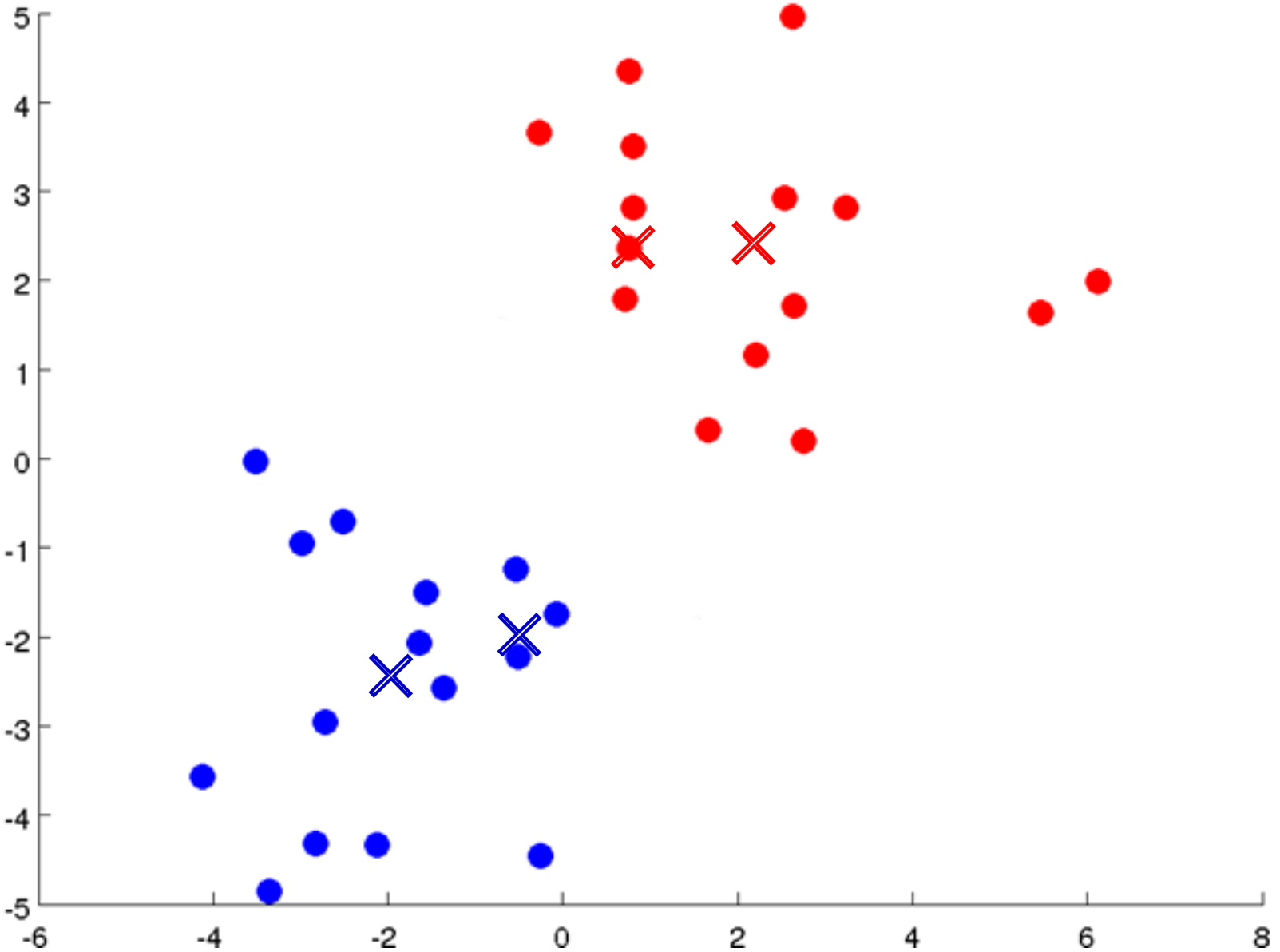
K=2



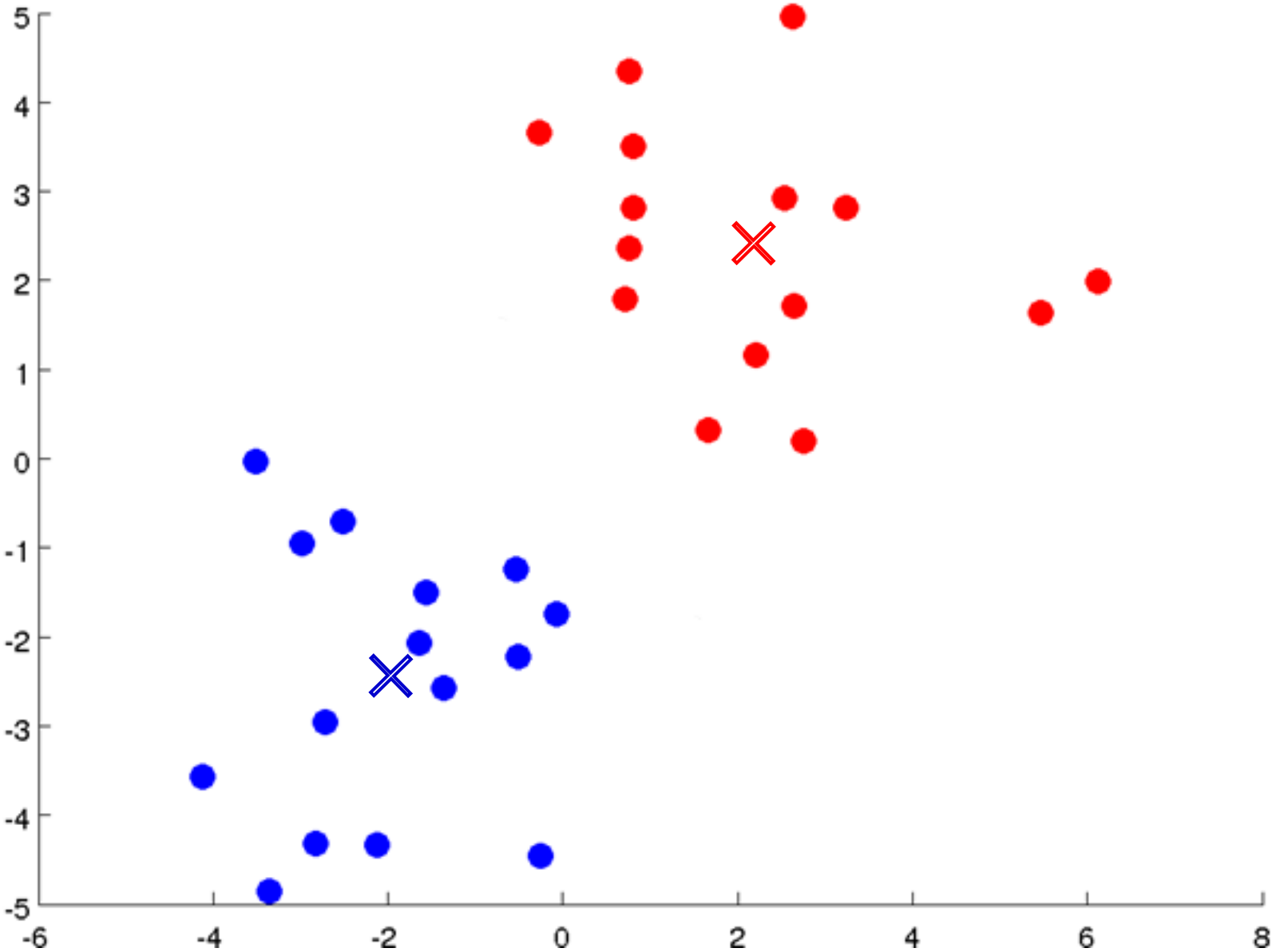
K=2



K=2



K=2



人脸数据库Olivetti的聚类分析

前100幅图像类别分配的图片表示。
具有同样颜色的人脸属于同一个类别，
而灰色图像表示没被分配到任何类别。
类别中心标有白色圆圈（9类）。

图像分割

根据颜色（RGB）的相似性分割图像.

K-Means 算法的主要问题

① **K-Means算法对参数的选择比较敏感**，不同的初始位置或者簇个数K的选择往往会导致完全不同的结果，当选取的K个初始簇中心点不合适时，不仅会增加聚类的迭代次数与时间复杂度，甚至有可能造成错误的聚类结果；

② 由于损失函数是非凸函数，则**不能保证计算出来的E的最小值是全局最小值**。在实际应用中，K-Means达到的局部最优已经可以满足需求。如果局部最优无法满足实际需要，可以重新选择不同的初始值，再次执行K-Means算法，直到达到满意的效果为止。

4.3.3.2 主成分分析

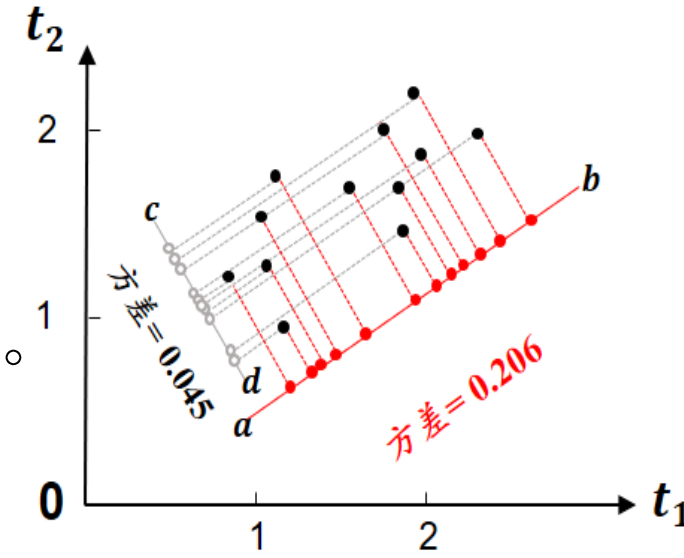
◆ 主成分分析（PCA）是一种常用于特征提取的线性降维方法。

◆ **PCA的主要原理**

通过某种线性投影，将高维的数据映射到低维的空间中，并期望在所投影的维度上的数据方差最大，方差的计算公式D：

$$D(X)=\frac{1}{m}\sum_{i=1}^{m}(x_i-\overline{x}_i)^2$$

其中，$X=(x_1, x_2, \cdots, x_m)$，为$m$个样本在某个维度上特征分量的集合，$\overline{x}_i$为$(x_1, x_2, \cdots, x_m)$的平均值。



◆ 方差大表示样本点在此超平面上的投影尽可能地被分开了，以此达到使用较少的数据维度来保留较多的原始样本点的特性的效果。

◆ 通过PCA还可以将一组可能存在相关性的特征分量（属性）转换为一组线性不相关的特征分量，转换后的这组特征分量叫**主成分**。

PCA算法的优点

(1) 仅用方差衡量信息量，不受数据集以外的因素影响。

(2) 各主成分之间正交，可消除原始特征分量之间相互影响的因素。

(3) 计算方法简单，主要运算是矩阵的特征值分解，易于实现。

4.4 弱监督学习

4.4.1 不完全监督学习

4.4.3.1 主动学习

4.4.3.2 半监督学习

4.4.3.3 迁移学习

4.4.3.4 强化学习

4.4.2 不确切监督学习

4.4.3 不准确监督学习

4.4 弱监督学习

◆数据集中**小部分有标注，大部分没标注**，即使有标注，也可能**质量不高、有噪声**。

◆**弱监督学习**是指已知数据及其**弱标签**（即**标签质量不高**，包含不完整、不确切、不准确的标签），训练一个智能算法，将输入数据映射到一组更强的标签的过程。

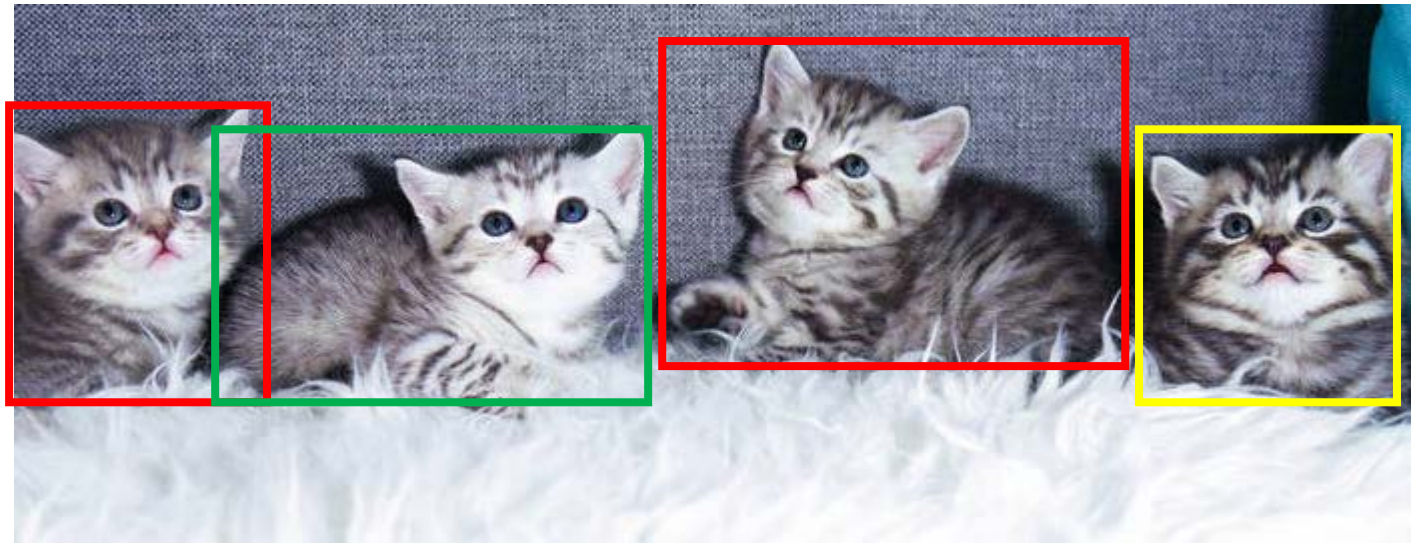
◆**标签的强弱**是指标签蕴合的信息量的多少，例如，相对于图像分割的标签而言，图像分类的标签就是弱标签。如果我们知道一幅图，告诉你图上有一只“狗”，然后需要你将“狗”在图像中的位置标出来，并且将“狗”从背景中分离出来，那么这就是已知弱标签（“狗”）、要去学习强标签（狗的位置、狗的轮廓）的弱监督学习问题（**不确切监督**）。

◆**与监督学习相比**，弱监督学习中的数据标签是不完全的，即训练集中只有一部分数据有标签，其余（甚至）大部分数据没有标签。

◆**与无监督学习相比**，弱监督学习还是有一定的监督信息，学习性能更好些。例如，在医学影像数据分析中，医生只能手工标注少量图像，很难获得完整的、全部的病灶标注。

弱标签的示例

- 对图像分割而言，图像分类的标签就是弱标签。
- 只知道类别“猫”，然后需要将“猫”在图像中的位置标出来，并且将“猫”从背景中分离出来，这就是已知**弱标签**（“猫”）、要去学习**强标签**（猫的位置、猫的轮廓）的**弱监督学习**问题。



监督学习、无监督学习、弱监督学习

- 监督学习：方法丰富、研究充分、性能好、**数据成本高**
- 无监督学习：方法简单、**数据成本低**、性能难以提升
- 弱监督学习：它介于监督学习和无监督学习之间，它利用带有弱标签的训练数据集进行监督学习，同时利用大量无标签数据进行无监督学习。

  弱监督学习是一种思路，是一类方法的总称。
- 根据数据**标注质量的不同以及监督信息量的多少**，**弱监督学习**可分为：
  - **不确切监督学习**
  - **不准确监督学习**
  - **不完全监督学习**

4.4.1 不完全监督学习（Incomplete Supervision Learning）

- 当训练集只有一个（通常很小的）子集有标签，而其它数据没有标签时，所进行的监督学习称为**不完全监督学习**。即：**少部分数据有人工标记，大部分数据无标记。**
- 例如，在图像分类任务中，很容易从互联网上获取大量图像，然而真实标签却需要人类标注者手工标出，考虑到标注的人工成本，往往只有一小部分图像能够被标注，导致训练集中大部分图像根本没有标签。
- 针对不完全监督学习，可以考虑采用**不同的技术**去改善和解决：

  - **主动学习**：假设未标注数据的真实标签可以**向人类专家查询**，让专家为估计模型最有价值的数据点打上标签。
  - **半监督学习**：通过学习有标记数据，逐渐扩展无标记数据，**无需人参与**。
  - **迁移学习**：样本迁移、特征迁移、模型迁移。
  - **强化学习**：通过不断试错，使下一次采取的动作能得到更多奖励，并且将奖励最大化。

4.4.1.1 主动学习

- **主动学习**（Active Learning）是一种机器学习框架，在这种框架中，
  - 学习算法可以从尚未标记的样本池中**主动选择**下一个**需要标记的可用样本子集**；
  - 然后交互式地、动态地向用户发出查询，**请求用户为其提供真实标签**，或从Oracle **数据库中查询**已人工标注好的标签，或由**人工标注员实时标记**；
  - 再反馈给学习算法，进行监督学习训练，逐步提升模型效果。
- 可见，在主动学习的训练过程中，**需要有人类的干预**。
- **主动学习的基本思想**：允许机器学习算法**自主选择**它想要学习的数据，它可以**在使用少量标签的同时达到更高的精度**。这种算法被称为主动学习器（Active Learner）。
- **目标**：使用尽可能少的查询来训练出性能良好的模型。

4.4.1.1 主动学习

- 在主动学习出现之前，系统会从未标注的样本中**随机选择待标注的样本**，进行人工标记。显然，这样选择出的样本**缺乏针对性**。
- 主动学习器**选择最重要的、信息量大的样本**，例如，易被错分的样本，或类边界附近的样本，让人进行标注。
- 主动学习过程通常包括**5个步骤**：

  （1）系统**主动选择**待标注样本

  （2）**人工标注**或数据库查询

  （3）模型训练

  （4）模型预测

  （5）模型更新。

上述步骤循环往复，直到训练错误率低于某个设定的阈值。

4.4.1.1 主动学习

- 主动学习的**关键**是如何挑选出合适的未标注样本子集用于人工标注。
- 主动学习在训练阶段，需要对数据进行**增量且动态的标注**，以使算法能够获得最大信息量的标签，从而迅速提升模型性能。
- 主动学习有广泛的应用场景，例如，

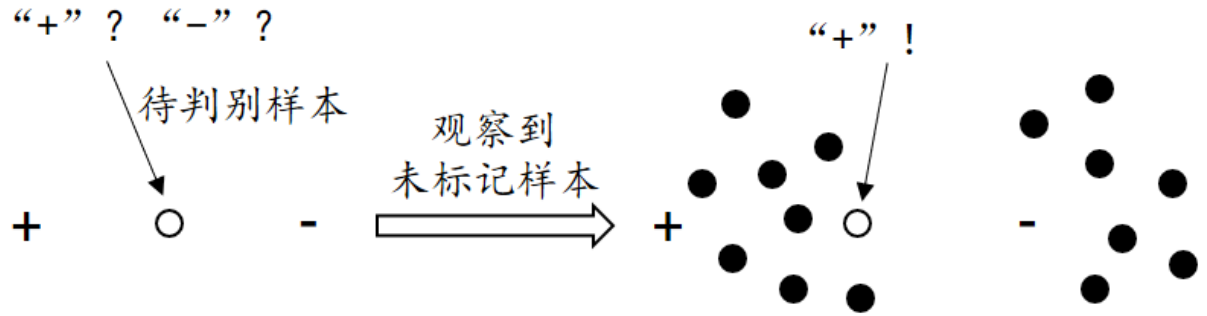
  ①个性化的邮件、短信分类：根据个人喜好来区分正常/垃圾短信和邮件。

  ②异常检测：如时间序列的异常检测（如频繁转账汇款，可能是诈骗）等。

  ③通用的图像识别系统，对某些极其相像的图像，还是需要人工识别，给予标注。

4.4.1.2 半监督学习（Semi-Supervised Learning）

- 半监督学习是监督学习与无监督学习相结合的一种学习方法，在其训练的过程中**无需人类干预**。
- 缺少已标注数据时，**主动学习**需要人类为未标注样本提供标签，而我们却**希望**：机器能通过对**有标签数据的学习，逐渐自动地将标签“传播”到无标签的数据上**。
- 半监督学习的**目的**：在不求助于人类专家的情况下，同时利用有标注数据和无标注数据来训练模型。
- 在半监督学习中，**有标注的数据往往是少数，未标注的数据是大多数**。少量的标注数据并不足以训练出好的模型，但可以**利用大量未标注数据来改善算法性能**。



半监督学习方法的实现思路

(1) 提取所有数据样本的特征信息，计算一个未标注样本 x_u 与所有标注样本之间的相似度。

(2) 认为：相似度越高的样本，它们的标签越倾向于一致。基于这个认知，找到与 x_u 相似度最高的已标注样本，将其标签赋予 x_u 作为标签。

(3) 重复前两步，逐步**将已标注样本的标签“传播”到未标注的样本上**，扩大标注数据集的规模；在标签传播过程中，**保持已标注样本的标签不变**，使其像一个源头将标签传向未标注样本。

(4) 上述迭代过程结束时，相似样本基本上都获得了相似的标签，从而完成了标签传播过程。

(5) 然后，利用规模较大的标注数据集训练学习模型，实现监督学习。

可见，半监督学习只需要低成本的人工标注，同时使用少量有标签数据和大量无标签数据，便能获得堪比甚至高于监督学习的性能。

4.4.1.3 迁移学习（Transfer Learning）

- 人类天生具有**举一反三**的能力。
  - 打乒乓球———打网球，下国际象棋———下中国象棋
  - C语言编程——— Python语言，跳拉丁舞后———跳探戈
- 计算机是否可以实现类似的功能？

  任务A 与任务B 具有某种相似性，利用A的学习经验，解决B，即**迁移学习**。
- **迁移学习**，是将在一个领域（称为**源领域**）的知识迁移到另外一个领域（称为**目标领域**），使得在目标领域能够取得更好的学习效果。
- 利用**源领域**的数据训练好学习模型，对其稍加调整，便可用于在**目标领域**完成任务。
  - 例1，将在 **ImageNet 图像集**上学习到的**分类模型**迁移到**医疗图像的分类**任务上；
  - 例2，将**中英互译**的翻译模型迁移到**日英互译**的任务上。
- 迁移学习的**关键要素**：寻找**源领域和目标领域**中**数据、任务**或**模型**之间的相似性或“不变性”。

(1) 样本迁移

样本迁移（Instance-Based Transfer Learning）

- 寻找任务A中与任务B相似的**标注数据**，能够直接作为任务B中的训练数据，用于训练任务B的模型。
- **基本思路**：在**源领域**中找与**目标领域**相似的数据样本，**加重该样本的权值**，**重复利用源领域中的样本标签数据**，用于训练目标领域中的模型。
- **主要目的**：找到**源领域**中有用的数据，用于在**目标领域**中训练学习模型。
- 例如：**任务1**：识别各种**交通工具**：汽车、飞机、轮船、自行车等，

  任务2：识别各种类型的**汽车**：轿车、卡车、面包车、跑车、越野车等
- 样本迁移方法**简单、易实现**,但**不太适用于源领域**与**目标领域数据分布相差较大**的情况,并且，对源领域中样本的选择、加重的权值、判断样本的相似性都**依赖于人的经验**，会使得模型的稳定性和可靠性降低。

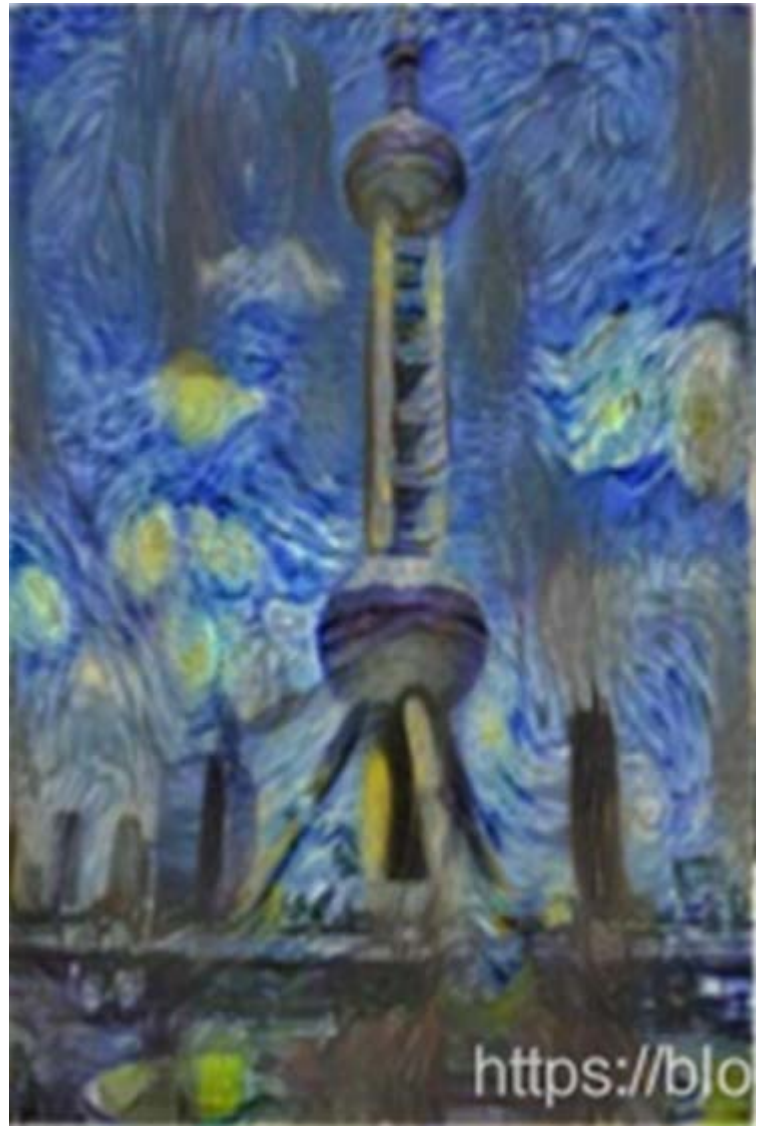
图像风格的样本迁移学习

“东方明珠”

+

梵高的《星月夜》

→



将梵高的画风迁移到“东方明珠”

图像风格的样本迁移学习

使用《星月夜》作为源数据

梵高作品《星月夜》

目标领域的原图

迁移后的画作

(2) 特征迁移

◆ **特征迁移**也称为**基于特征的迁移**（Feature-Based Transfer Learning）。

◆**基本思路**：当**源领域**和**目标领域**含有一些共同特征时，则可以通过特征变换将源领域和目标领域的特征映射到同一个特征空间中，使得在该空间中源领域数据与目标领域数据具有相同的数据分布，然后再利用传统的机器学习方法来求解。

◆特征迁移**旨在**通过引入**源领域**的**数据特征**来帮助完成**目标领域**的机器学习任务。

◆当目标领域缺少足够的标签时，可通过挖掘**源领域数据**与**目标领域数据的共同特征**，或者**借助中间数据进行“桥接”**，会有助于实现不同特征空间之间的知识迁移。

◆如：在**识别图像中花卉种类**（**目标领域**）时，若缺少**花卉种类的标签**，可借助**wiki**、**百度百科**等相关数据源（即**源领域**）获得带有标签的文本和图文并茂的中间数据。

◆特征迁移通常假设源领域和目标领域间有一些共同特征，在共同特征空间中迁移知识。

（3）模型迁移

- **模型迁移**也称为基于模型的迁移学习或基于参数的迁移学习（Parameter-Based Transfer Learning）。
- **基本思路**是：将**源领域**上训练好的**模型的一部分参数或者全部参数**应用到**目标领域**任务的模型上。
- 模型迁移可以利用模型之间的相似性来提高模型的能力。
- 例如，在需要使用模型完成**水果分类**任务时,可以将在**ImageNet上预训练的模型**用于**初始化水果分类模型**，再利用**目标领域**的几万个已标注样本进行**微调**，即可得到精度很高的模型。

4.4.1.3 迁移学习

- 与样本迁移和特征迁移一样，模型迁移也利用源领域的知识。然而，**三者在利用知识的层面上存在显著差别**：
  - **样本迁移**利用样本层面的知识，即**样本**；
  - **特征迁移**利用特征层面的知识，即**特征**；
  - **模型迁移**利用模型层面的知识，即**参数**。
- 模型迁移是目前**最主流的迁移学习方法**，可以很好地利用源领域中已有模型的参数，使得学习模型在面临新的任务时，只需**微调**，便可完成目标领域的任务。
- 目前，迁移学习已经在机器人控制、机器翻译、图像识别、人机交互等诸多领域获得了广泛的应用。

4.4.1.4 强化学习

- 强化学习（Reinforcement Learning），又称再励学习、评价学习或增强学习。
- 受到行为心理学的启发，强化学习的算法理论早在20世纪六七十年代就已形成，但直到最近才引起了学术界与工业界的广泛关注。
- **强化学习的思路**是：通过不断**试错**，使下一次采取的动作能够得到更多奖励，并且将奖励最大化。
- 强化学习用于描述和解决**智能体**（Agent）在与外部环境的交互过程中采取学习策略，以获得最大化的回报或实现特定目标的问题。其中，
  - **智能体**就是需要训练的学习模型，
  - **外部环境**则被表示为一个可以给出反馈信息的**模拟器**（Emulator），
  - **回报**（Reward）是通过人为设置的奖励函数计算得到的。

4.4.1.4 强化学习

◆ 智能体与环境通过在一系列“观察（Observation）—动作（Action）—回报（Reward）”的交互中获得知识，改进行动方案，以适应环境。

◆ 标准的强化学习过程如下：

(1) 初始化，令当前时刻t =1。

(2) 智能体获取当前时刻环境的状态信息，记为s_t。

(3) 智能体对环境采取试探性动作a_t。

(4)根据s_t和a_t，环境采用奖励函数计算出一个评价值r_t，反馈给智能体。

①若a_t正确，则智能体获得奖励，以后采取动作a_t的趋势将加强。

②若a_t错误，则智能体获得惩罚，以后采取动作a_t的趋势将减弱。

(5) 环境更新状态为s_{t+1}。

(6) 令t=t+1，若$t < T$（事先规定的时刻），返回第(2)步，否则算法结束。

4.4.1.4 强化学习

◆强化学习就是在上述反复交互的过程中不断修改从状态到动作的映射策略，以期获得最大化的累积回报，达到优化模型的目的。

◆强化学习不要求预先给定任何数据，而是通过接收环境对动作的反馈信息（奖励或惩罚）来获得学习数据，并用于更新模型参数，并非直接告诉智能体如何采取正确的动作。

◆**强化学习的关注点**在于对未知领域的探索和对已有知识的利用之间的平衡。

4.4.1.4 强化学习

◆ **强化学习与监督学习的不同之处在于**:

①监督学习输入的训练数据是包含样本及其标签的强监督信息，而强化学习从外部环境接收的反馈信息是一种弱监督信息;

②监督学习是采用正确答案来训练模型，给予的指导是即时的；而强化学习是采用“试错”的方式来训练模型，外部环境给予它的指导有时是延迟的。

◆ **强化学习与无监督学习的不同之处在于**:

- 无监督学习输入的是没有任何监督信息的无标注数据，
- 强化学习从环境获得的评价信息（不是正确答案）是一种弱监督信息，尽管监督信息很弱，但总比没有要好。

◆ 主流的强化学习算法包括Q学习（Q-Learning）、Deep Q-Network（深度强化学习的一种方法）等。

4.4.1.4 强化学习

◆ **强化学习最成功的应用案例无疑是在博弈领域**。

◆ 2016-2017年，谷歌DeepMind团队先后研发了围棋系统**AlphaGo**、升级版的围棋系统**AlphaGo Zero**和棋类游戏的拓展版**AlphaZero**。至此，半个多世纪以来，在游戏领域独占鳌头的博弈搜索方法被强化学习取代。

◆ 目前，**强化学习的应用场景越来越广泛**。例如，

- 工业领域的无人机、机器人作业、机械臂抓取物体等；
- 金融贸易领域的未来销售额预测、股价预测等；
- 自然语言处理领域的文本摘要、自动问答、机器翻译等；
- 新闻推荐、时尚推荐等推荐系统。

4.4.2 不确切监督学习

- **不确切监督**（Inexact Supervision Learning）是指在训练样本**只有粗粒度标签**的情况下进行的监督学习。
- 何为**粗粒度标签**？例如，
  - 有一张肺部 X 光图像，只知道该图像是一位肺炎患者的肺部影像，但并不清楚其中**哪个部位**的影像说明了其主人患有肺炎。
  - 假设一幅图像中有两只猫，现在只标注了整张图像的类别为“猫”，而**没有**标注两只猫在图像中各自的**边界框**。
- 若想在肺部影像上标出病灶的部位，或者想在猫的图像上进行目标定位，却只有粗粒度的标签——图像类别名称，而没有进一步的定位信息，即**只有图像级的标签，而没有对象级的标签**。

4.4.2 不确切监督学习

- 针对不确切监督学习，可以考虑采用**多实例学习**（Multi-Instance Learning，也译为多示例学习）技术去改善和解决。
- 在多实例学习中，定义“包”（bag）为多个实例（instance）的集合，即**每个“包”中包含多个实例**。
- 每个“包”都具有标签，但“包”中的实例却都没有标签。
- 当“包”中至少有一个正实例时，将该包定义为“正包”；
- 反之，当且仅当“包”中所有实例均为负实例时，该“包”定义为“负包”。
- 在多实例学习中，训练集由若干个具有标签的“包”组成，每个“包”都包含若干个未标注的实例。
- 多实例学习的**目标**是通过对具有标签的多实例“包”的学习，归纳出单个实例的标签类别，建立多实例分类器，尽可能准确地预测未知新“包”的标签。

4.4.3 不准确监督学习 (Inaccurate Supervision Learning)

◆**不准确监督**：也译为**不精确监督学习**。训练样本的**标签不总是正确的**，有些标签是错误的，在此场景下所进行的监督学习称为**不准确监督学习**。

◆例如，图像中本来是“**美洲豹**”，但图像标签却被错误地标记成了“**花豹**”。

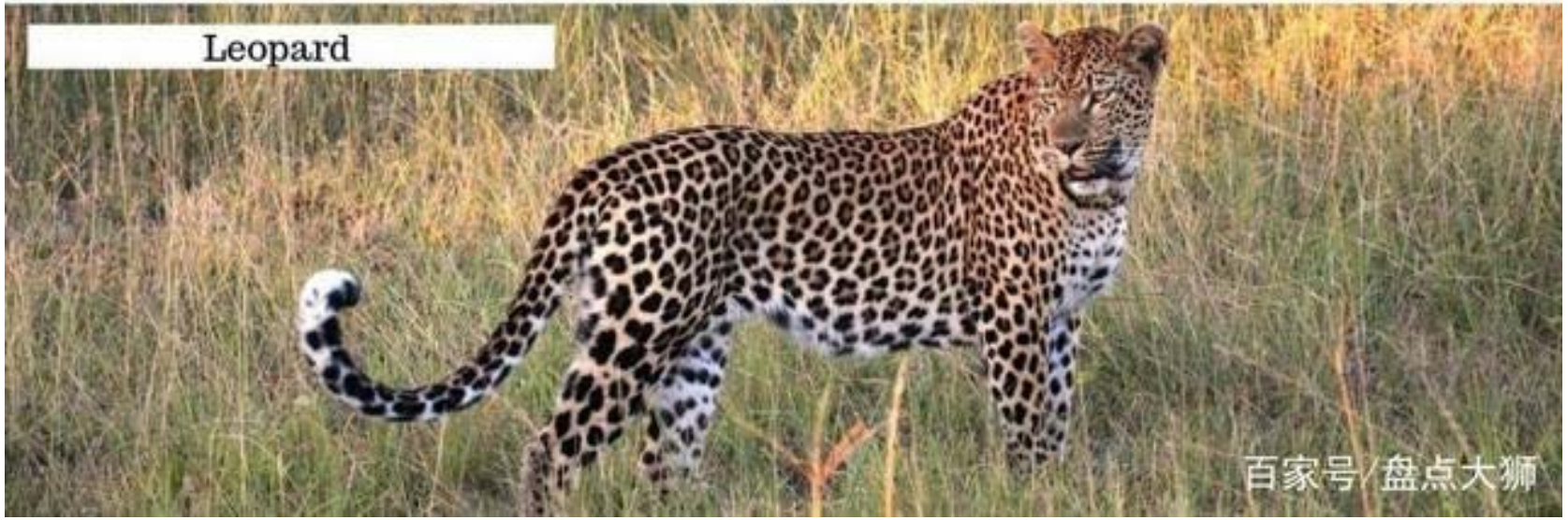
◆常见**原因**有三：

（1）因为图像标注者**粗心或疲倦**而导致的**失误**；

（2）标注者的**认知水平有限**，导致所标注的标签质量不高；

（3）图像属于**专业领域，标注难度大**，即使是人类专家也**难以统一认识**，例如医学影像的病理分析标注。

例如：“美洲豹” VS “花豹”



背上斑纹较小较密,环纹内部中空

美洲虎; 美洲豹（jaguar）

背上的斑纹呈现出更复杂的玫瑰花型;

花豹（leopard）

背上的斑纹类似于抽象的玫瑰花型。

4.4.3 不准确监督学习

- 一个典型的场景是在**标签有噪声**的情况下学习（Learning with Label Noise）。
- 已有许多相关理论研究，其中大多数都假设存在随机类型的噪声，即标签受到随机噪声的影响。
- 在实践中，一个**基本的想法**就是**识别出潜在的被标错的样本，然后试着进行更正**。
- **解决思路**，
  - 首先，用数据编辑方法**构建一个邻域图**，其中每个节点对应于一个训练样本，一条边连接两个具有不同标签的节点；
  - 然后，**判断一个节点是否可疑**，直觉上，若一个节点（即实例）与许多边相关联，则它是可疑的，可以删除或重新标注可疑节点。如此，则可以修正部分错误的标签。
  - 最后，再**用修正的数据训练学习模型**，会提升模型的性能。

无监督与自监督的区别

◆**自监督：在无标注数据上训练，其“标注”通常来自于数据本身**,其常规操作是通过各种各样的“auxiliary task”来提高学习表征(representation)的质量，从而提高下游任务的质量，如：**预测被盖住的文字或视频中的字幕。**

◆**自监督学习不再依赖标注，而是通过揭示数据各部分之间的关系，从数据中生成标签。**所以说，**自监督学习的核心：在于如何自动为数据产生标签。**

◆**无监督学习**：没有任何训练样本，即**没有训练过程**，而是直接对数据进行建模。**它不依赖于标签**，通过对数据内在特征的挖掘，找到样本间的关系，如聚类。

◆自监督学习和无监督学习不同，它主要是利用辅助任务（pretext）从大规模的无标签数据中挖掘自身的监督信息，通过这种构造的监督信息对网络进行训练，从而可以学习到对下游任务有价值的表征。即**自监督学习的监督信息不是人工标注的**，而是**算法在大规模无标记数据中自动构造监督信息**，来进行监督学习或训练。

◆ **以上各个概念的分类并不是严格互斥的。**

4.5 本章小结

1. 机器学习的三个视角：机器学习的任务、机器学习的范式、机器学习的模型。

2. **机器学习的任务**包括：分类、回归、排名、聚类、降维、密度估计等。

3. 各种**机器学习范式**及其经典算法。

（1）**监督学习**：用有标签的数据训练模型，经典算法包括KNN和SVM等，有训练过程。

（2）**无监督学习**：用无标签的数据建立模型，经典算法包括K-Means、PCA等，没有训练过程。

（3）**弱监督学习**：用带有弱标签的数据集训练模型，分为**不完全监督学习、不确切监督学习、不准确监督学习**。虽然将弱监督学习分为上述3种类别，但在实际操作中，它们经常同时发生。以上各个学习范式的分类并不是严格互斥的。

4.5 本章小结

不完全监督学习包括主动学习、半监督学习、迁移学习、强化学习。

① 主动学习：在训练模型的过程中，算法挑选出需要标注的未标注样本子集，让人来标注。主动学习过程需要人的干预。其目标是使用尽可能少的查询来训练出性能良好的模型。

② 半监督学习：同时用有标签和无标签的数据进行训练，已标注的样本向未标注的样本“传播”标签。半监督学习过程不需要人的干预。

③ 迁移学习有3种方式：样本迁移、特征迁移、模型迁移。

④ 强化学习：没有训练数据，根据外部环境反馈的奖惩信号调整智能体的动作。经典算法有Q-Learning。