目录 CONTENTS

中国石油大学(华东)
CHINA UNIVERSITY OF PETROLEUM

计算机科学与技术学院
Big Data & Cyber Security

第七章　机器学习

智能科学系

目录 CONTENTS

7.1 机器学习简介

- Machine learning is the study of **algorithms** and mathematical **models** that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model of **sample data**, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. -----**Wikipedia**

- 机器学习专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。 -----**百度百科**

- 对于某给定的任务T，在合理的性能度量方案P的前提下，某计算机程序可以自主学习任务T的经验E；随着提供合适、优质、大量的经验E，该程序对于任务T的性能逐步提高。 -----**Machine learning**

7.1 机器学习简介

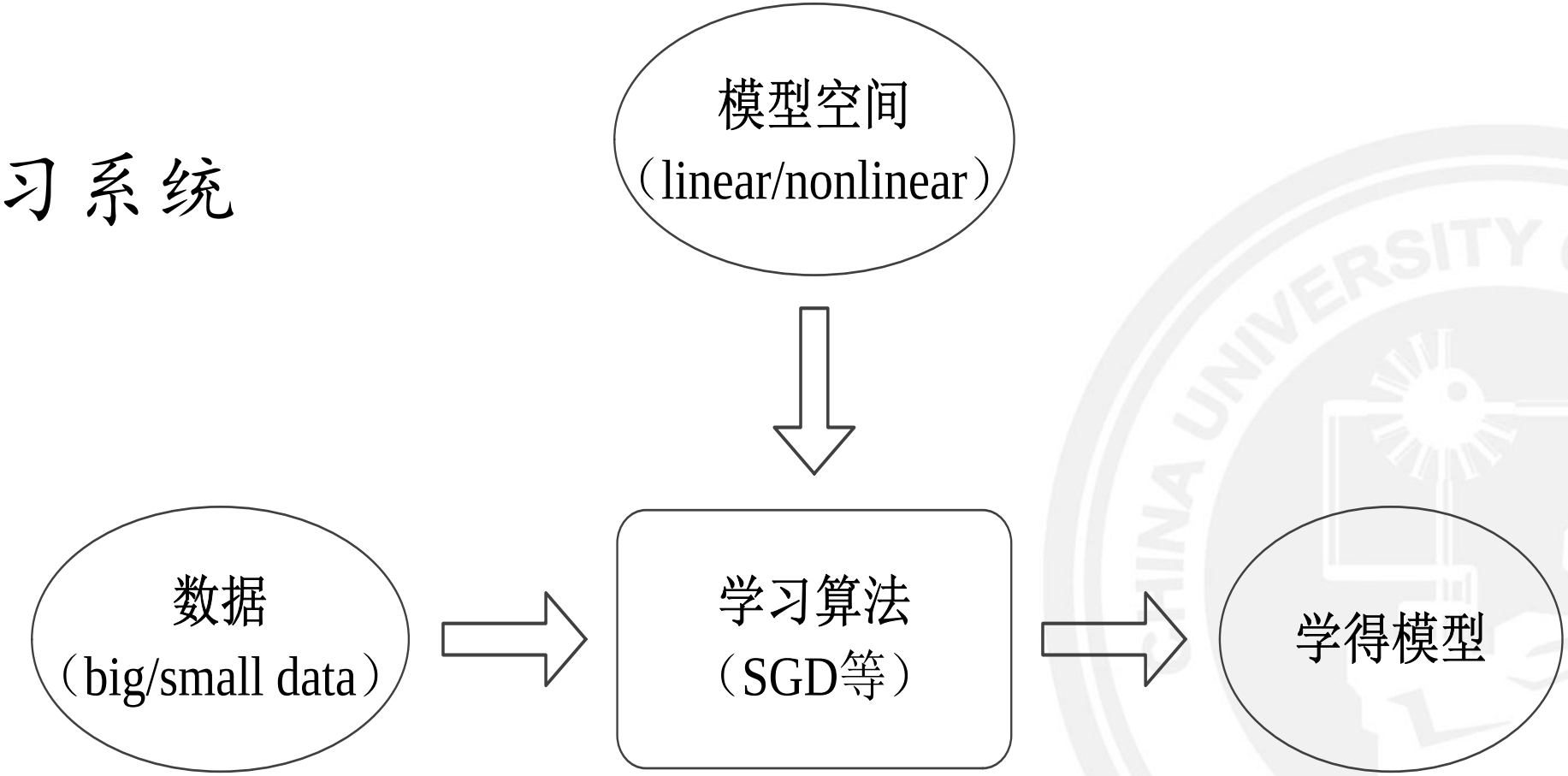
利用经验改善系统自身的性能

具体解释：

- ✓ 经验(数据和常识)，在此更多指的是数据（如互联网数据、科学实验数据等），即从数据中总结规律用于将来的预测
- ✓ 计算机系统对应于机器学习模型（如决策树、支持向量机等）
- ✓ 性能是模型对新数据的处理能力（如分类和预测性能等）
- ✓ 学习是一个蕴含特定目的的知识获取过程，其内部表现为新知识的不断建立和修正，而外部则表现为性能改善。

7.1 机器学习简介

学习系统



7.1 机器学习简介

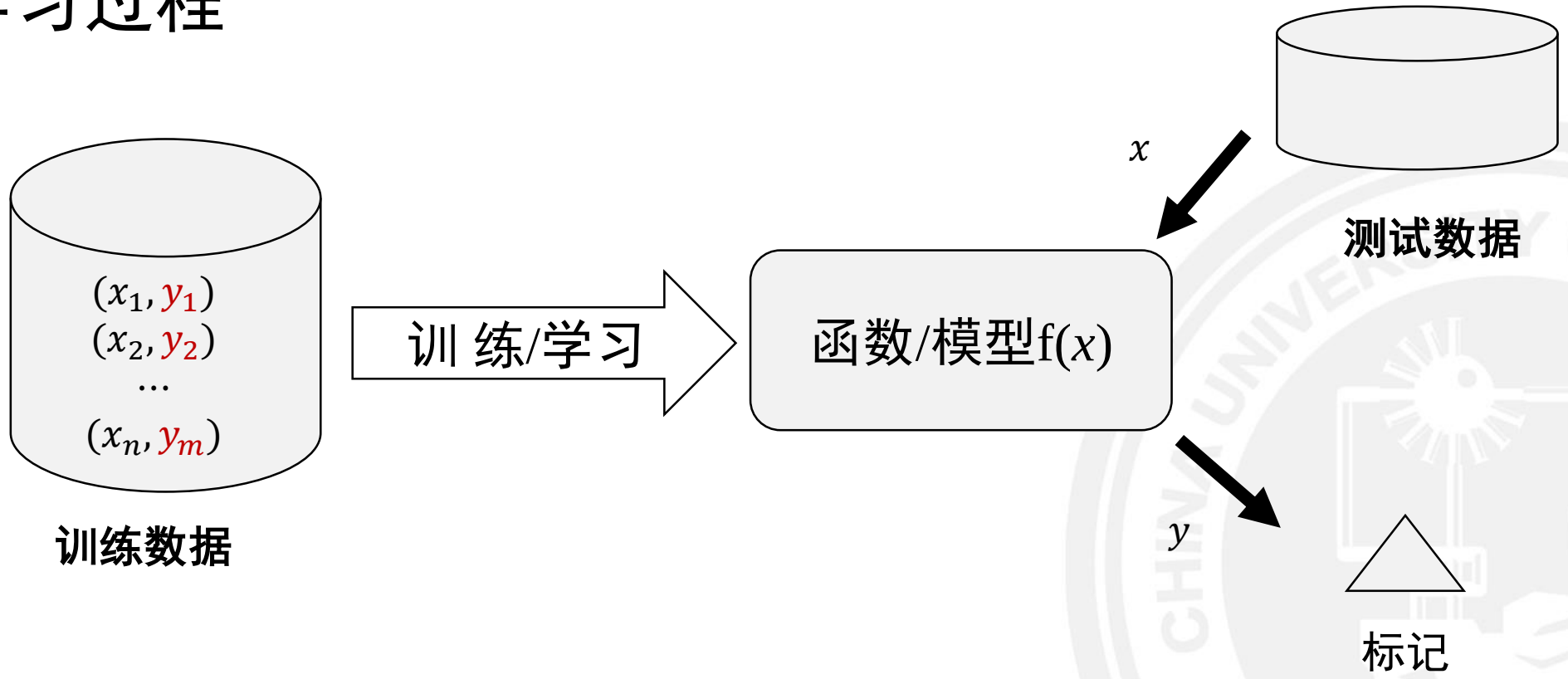
任何通过数据训练的学习算法都属于机器学习

经典ML算法

- 线性回归(Linear Regression)
- K-均值聚类方法(K-means)
- 主成分分析(Principal Component Analysis-PCA)
- 决策树(Decision Trees)和随机森林(Random Forest)
- 支持向量机(Support Vector Machines)
- 人工神经网络(Artificial Neural Networks)

7.2 监督学习

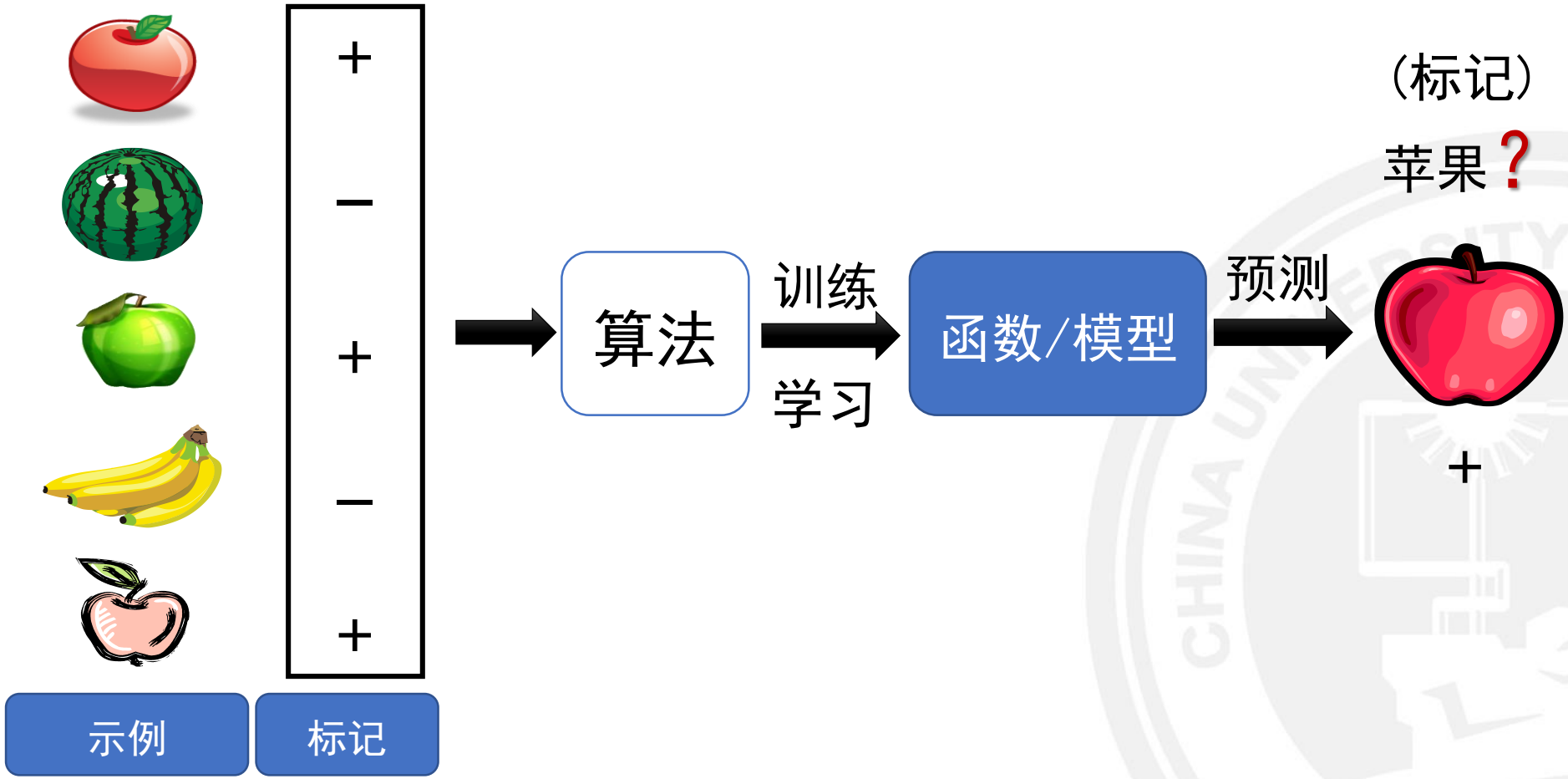
学习过程



事实上，f(x)是一个从x到y的映射

7.2 监督学习

学习过程



7.2 监督学习

- 已知：输入和输出

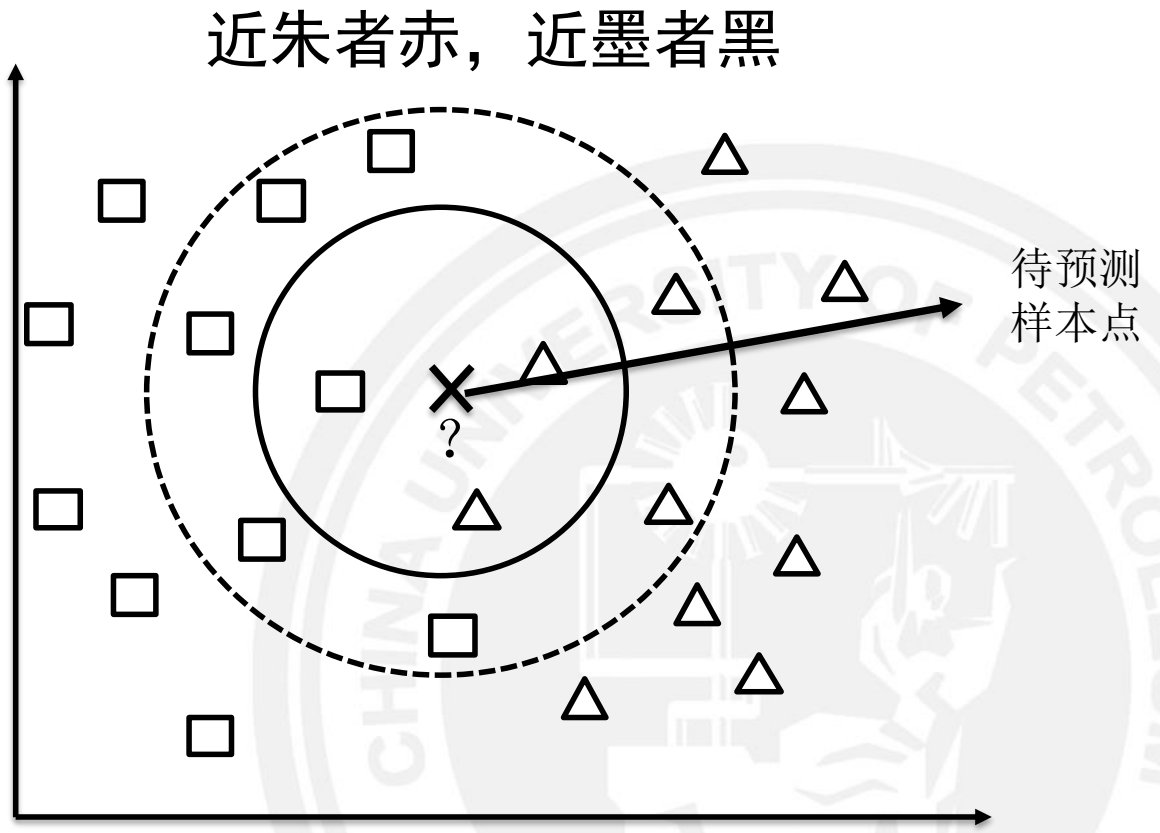
- 任务：（1）建立起一个将输入准确映射到输出的模型
（2）给模型输入新的值，能预测出对应的输出

7.2 监督学习

1. K-近邻

最简单的机器学习分类算法

- ✓计算待预测样本与训练样本集合所有样本距离
- ✓取前k个最相似训练样本，统计其类别
- ✓将占比最多的类别赋给待预测样本

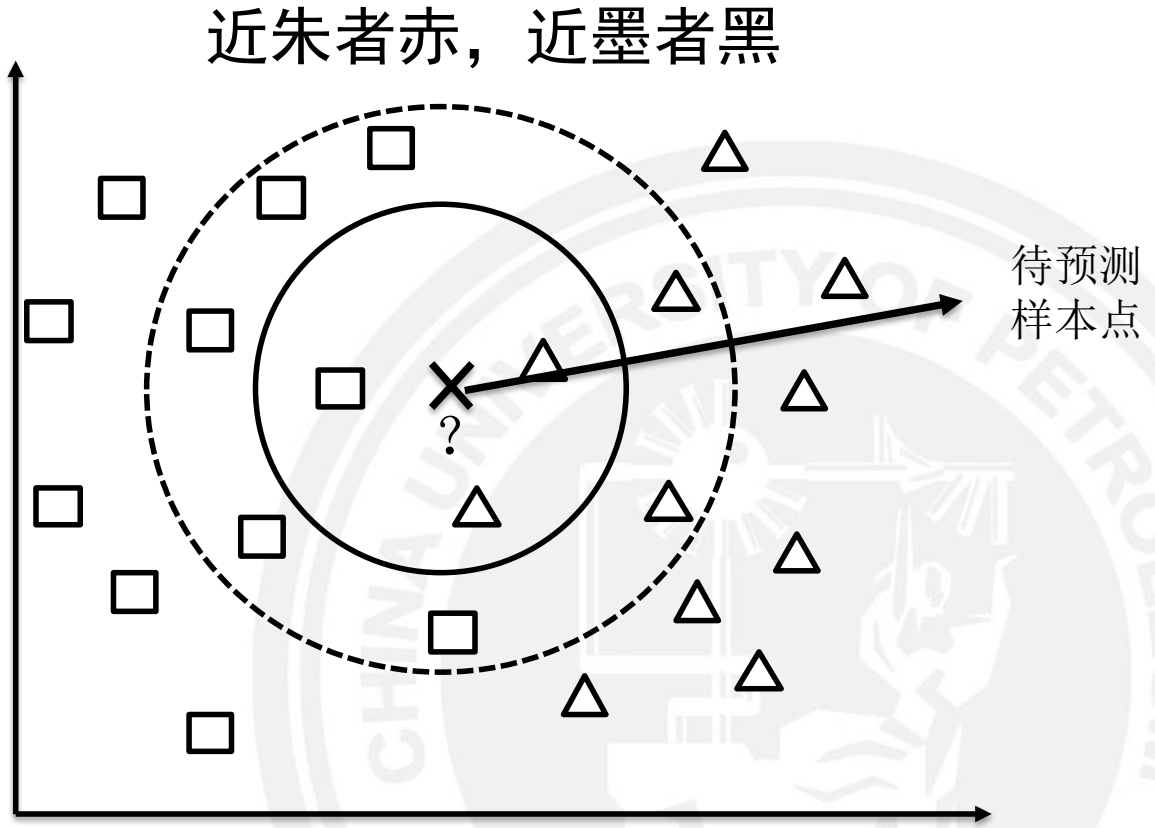


7.2 监督学习

1. K-近邻

最简单的机器学习分类算法

- ✓计算待预测样本与训练样本集合所有样本距离
- ✓取前k个最相似训练样本，统计其类别
- ✓将占比最多的类别赋给待预测样本



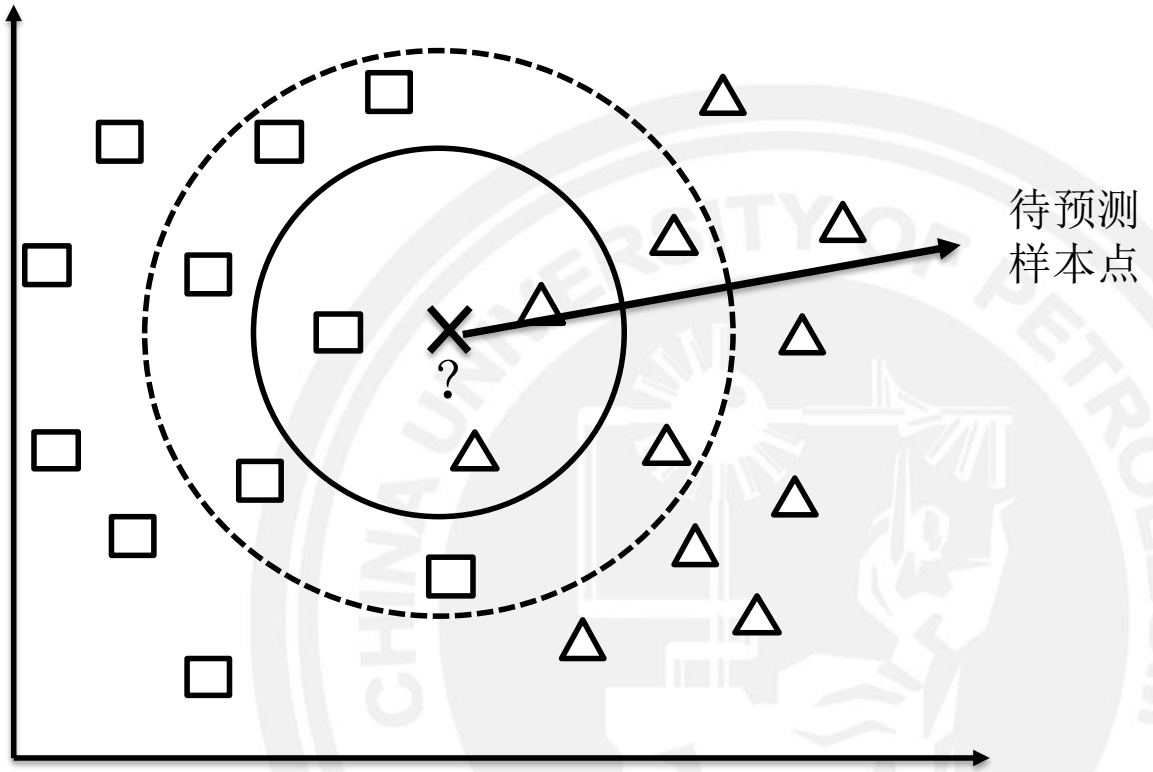
例：k=3时，三角形的类别
k=10时，正方形的类别

7.2 监督学习

1. K-近邻

算法优缺点：

- ✓优点是简单容易实现，支持多分类，不需要进行训练，直接用训练数据来实现分类。
- ✓对k敏感，计算量大



例：k=3时，三角形的类别
k=10时，正方形的类别

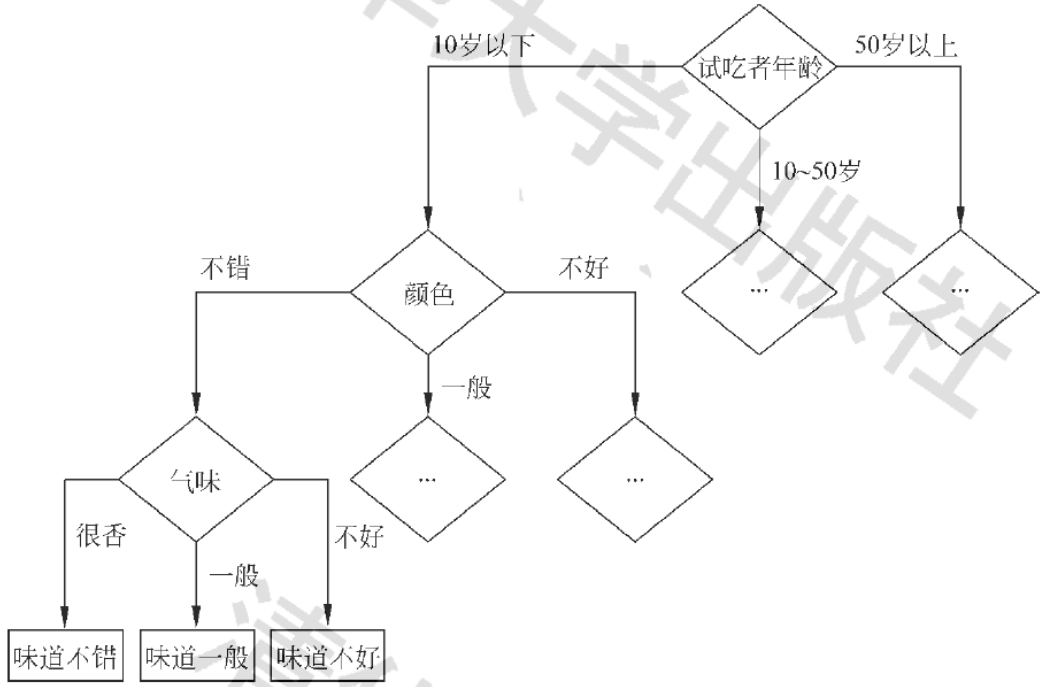
7.2 监督学习

2. 决策树

- 决策树是一种有监督的分类方法,它是用已有的数据构造出一棵树,用这棵树,对新的数据进行预测。

- 决策树基于“树”结构进行决策

  - 每个“内部结点”对应于某个属性上的“测试”(test)
  - 每个分支对应于该测试的一种可能结果(即该属性的某个取值)
  - 每个“叶结点”对应于一个“预测结果”

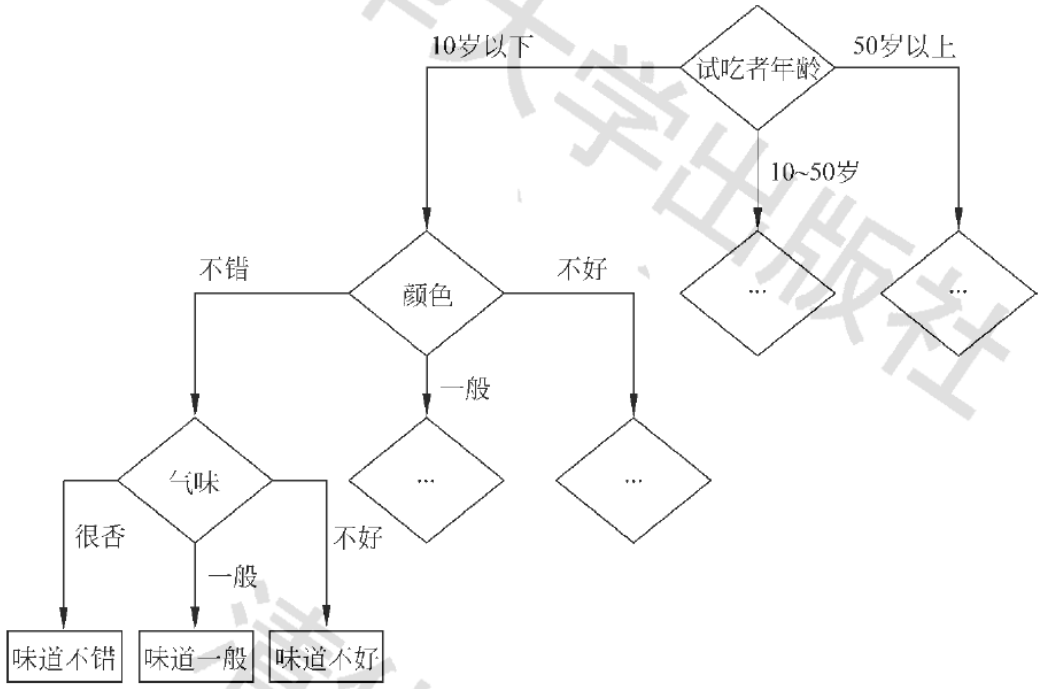


7.2 监督学习

2. 决策树

决策树模型

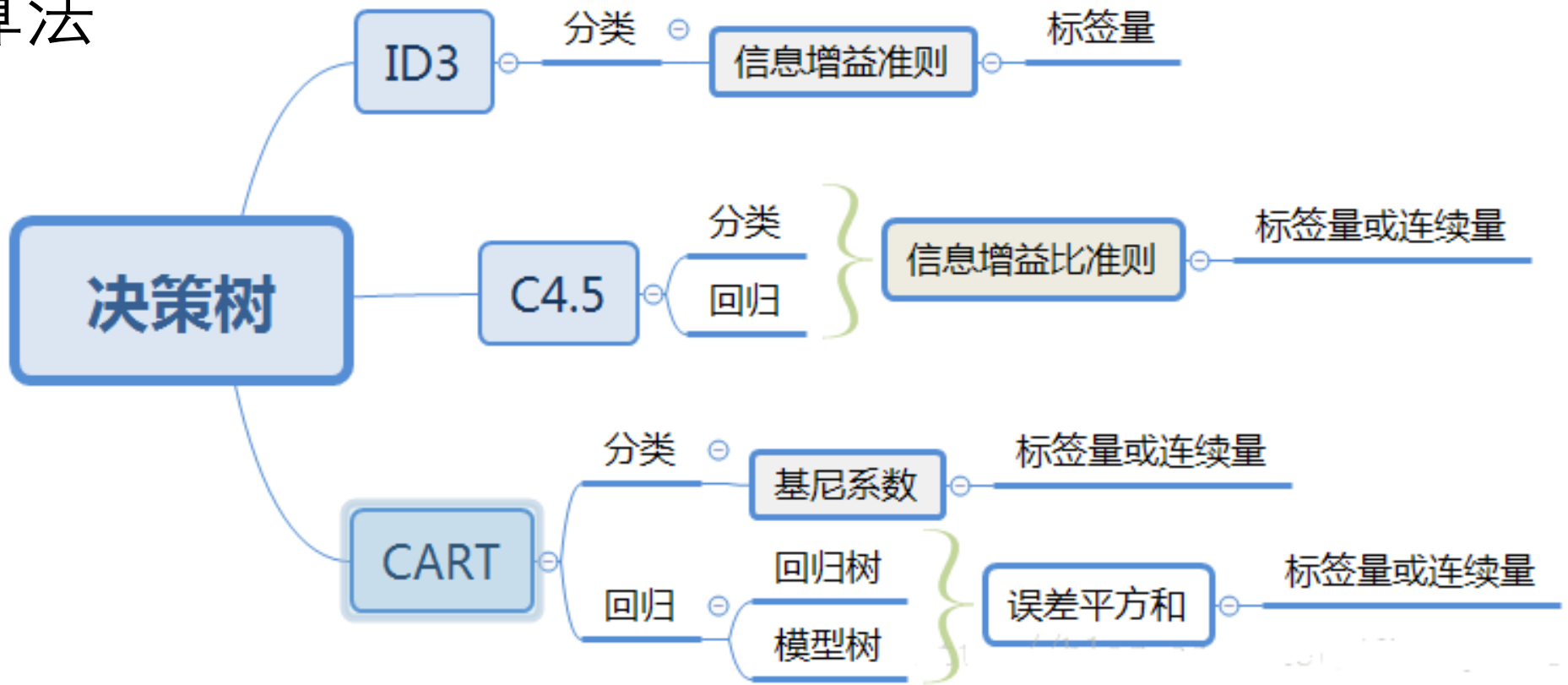
- 学习过程：通过对训练样本的分析来确定“划分属性”（即内部结点所对应的属性）
- 预测过程：将测试示例从根结点开始，沿着划分属性所构成的“判定测试序列”下行，直到叶结点。



7.2 监督学习

2. 决策树

- 决策树算法



7.2 监督学习

2.1 决策树的构造过程

一般包含三个部分

1、特征选择：特征选择是指从训练数据中众多的特征中选择一个特征作为当前节点的分裂标准，如何选择特征有着很多不同量化评估标准，从而衍生出不同的决策树算法，如CART, ID3, C4.5等。

2、决策树生成： 根据选择的特征评估标准，从上至下递归地生成子节点，直到数据集不可分则决策树停止生长。 对树结构来说，递归结构是最容易理解的方式。

3、剪枝：决策树容易过拟合，一般来需要剪枝，缩小树结构规模、缓解过拟合。剪枝技术有预剪枝和后剪枝两种。

7.2 监督学习

2.1 决策树的构造过程-基本流程

策略：“分而治之”(divide-and-conquer)

自根至叶的递归过程

在每个中间结点寻找一个“划分”(split or test)属性

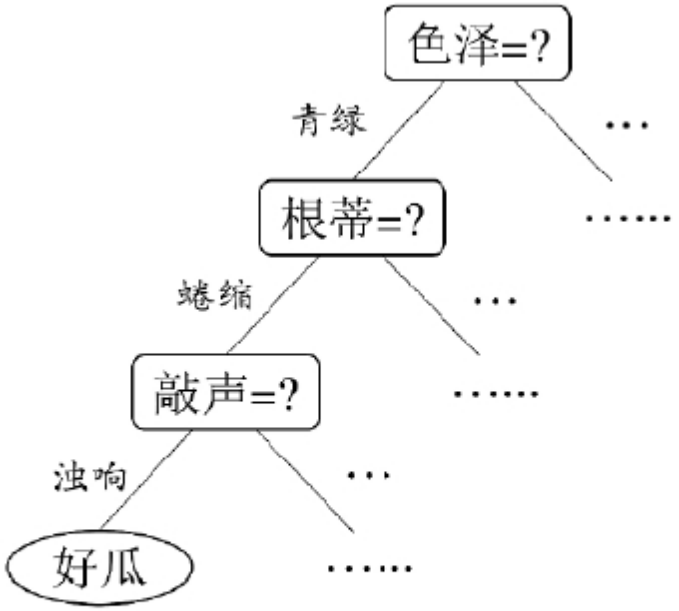


图 4.1 西瓜问题的一棵决策树

三种停止条件：

(1) 当前结点包含的样本全属于同一类别，无需划分;

(2) 当前属性集为空, 所有样本在所有属性上取值相同，无法划分;

(3) 当前结点包含的样本集合为空，不能划分。

7.2 监督学习

2.1 决策树的构造过程-信息增益（information gain)

信息熵 (entropy) 是度量样本集合“纯度”最常用的一种指标

假定当前样本集合 D 中第 k 类样本所占的比例为 p ，则 D 的信息熵定义为

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

计算信息熵时约定：若 $p=0$，则 $p\log_2 p = 0$.

$\mathrm{Ent}(D)$ 的最小值为 0，最大值为 $\log_2 |\mathcal{Y}|$.

Ent(D)的值越小，则D的纯度越高

信息增益直接以信息熵为基础，计算当前划分对信息熵所造成的变化

7.2 监督学习

2.1 决策树的构造过程-信息增益（information gain)

离散属性 a 的取值: $\{a^1, a^2, \ldots, a^V\}$

D^v: D 中在 a 上取值 = a^v 的样本集合

以属性 a 对 数据集 D 进行划分所获得的信息增益为:

$$\mathrm{Gain}(D,a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

划分前的信息熵

划分后的信息熵

第 v 个分支的权重，
样本越多越重要

7.2 监督学习

2.1 决策树的构造过程-信息增益

例子：共17个样本，分类为好瓜和坏瓜

$$|\mathcal{Y}| = 2$$

$$p_1 = \frac{8}{17} \qquad p_2 = \frac{9}{17}$$

因此划分前的信息熵为：

$$\mathrm{Ent}(D) = -\sum_{k=1}^{2} p_k \log_2 p_k =$$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

西瓜数据集

7.2 监督学习

2.1 决策树的构造过程-信息增益

例子：共17个样本，分类为好瓜和坏瓜

$$|\mathcal{Y}| = 2$$

$$p_1 = \frac{8}{17} \qquad p_2 = \frac{9}{17}$$

因此划分前的信息熵为：

$$\mathrm{Ent}(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(\frac{8}{17}\log_2\frac{8}{17} + \frac{9}{17}\log_2\frac{9}{17}) = 0.998$$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

西瓜数据集

7.2 监督学习

2.1 决策树的构造过程-信息增益

- 以“色泽”为例选择属性进行划分。

其对应的3个数据子集分别为:

D1 （色泽=青绿）， D2 (色泽=乌黑), D3(色泽=浅白)

$$\mathrm{Ent}(D^1) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

$$\mathrm{Ent}(D^2) =$$

$$\mathrm{Ent}(D^3) =$$

- 可以计算得到属性“色泽”的信息增益

$$\mathrm{Gain}(D, \text{色泽}) = \mathrm{Ent}(D) - \sum_{v=1}^{3} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

7.2 监督学习

2.1 决策树的构造过程-信息增益

- 以“色泽”为例选择属性进行划分。

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

其对应的3个数据子集分别为:

D1 （色泽=青绿）， D2 (色泽=乌黑), D3(色泽=浅白)

$$\mathrm{Ent}(D^1) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1.000$$

$$\mathrm{Ent}(D^2) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918$$

$$\mathrm{Ent}(D^3) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$$

- 可以计算得到属性“色泽”的信息增益

$$\mathrm{Gain}(D, \text{色泽}) = \mathrm{Ent}(D) - \sum_{v=1}^{3}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)$$

7.2 监督学习

2.1 决策树的构造过程-信息增益

- 以“色泽”为例选择属性进行划分。

其对应的3个数据子集分别为:

D1 （色泽=青绿）， D2 (色泽=乌黑), D3(色泽=浅白)

$$\mathrm{Ent}(D^1) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1.000$$

$$\mathrm{Ent}(D^2) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918$$

$$\mathrm{Ent}(D^3) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$$

- 可以计算得到属性“色泽”的信息增益

$$\begin{aligned}\mathrm{Gain}(D, \text{色泽}) &= \mathrm{Ent}(D) - \sum_{v=1}^{3}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v) \\ &= 0.998 - \left(\frac{6}{17}\times 1.000 + \frac{6}{17}\times 0.918 + \frac{5}{17}\times 0.722\right) \\ &= 0.109\end{aligned}$$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

7.2 监督学习

2.1 决策树的构造过程-信息增益

- 类似的，其他属性的信息增益为

$\text{Gain}(D, \text{根蒂}) =$　　$\text{Gain}(D, \text{敲声}) =$

$\text{Gain}(D, \text{纹理}) =$　　$\text{Gain}(D, \text{脐部}) =$

$\text{Gain}(D, \text{触感}) =$

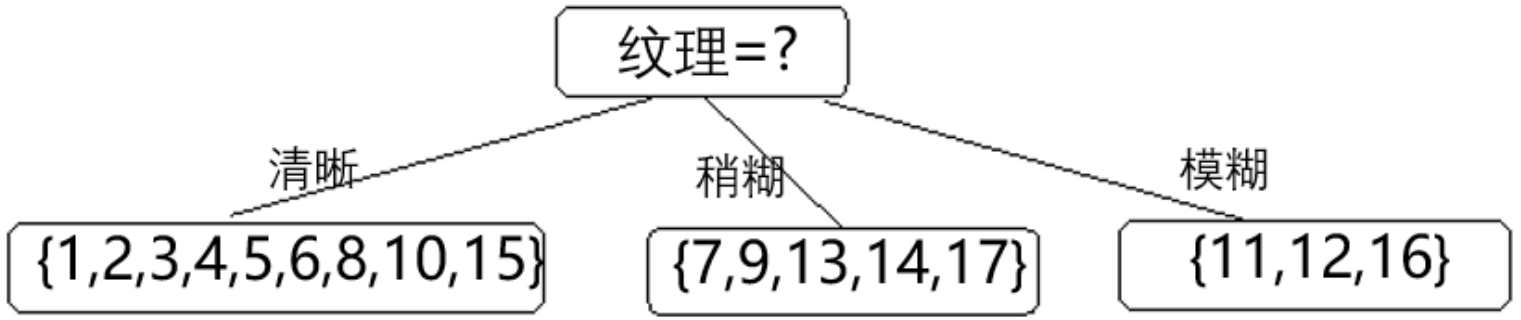
7.2 监督学习

2.1 决策树的构造过程-信息增益

- 类似的，其他属性的信息增益为

$$\text{Gain}(D, \text{根蒂}) = 0.143 \qquad \text{Gain}(D, \text{敲声}) = 0.141$$
$$\text{Gain}(D, \text{纹理}) = 0.381 \qquad \text{Gain}(D, \text{脐部}) = 0.289$$
$$\text{Gain}(D, \text{触感}) = 0.006$$

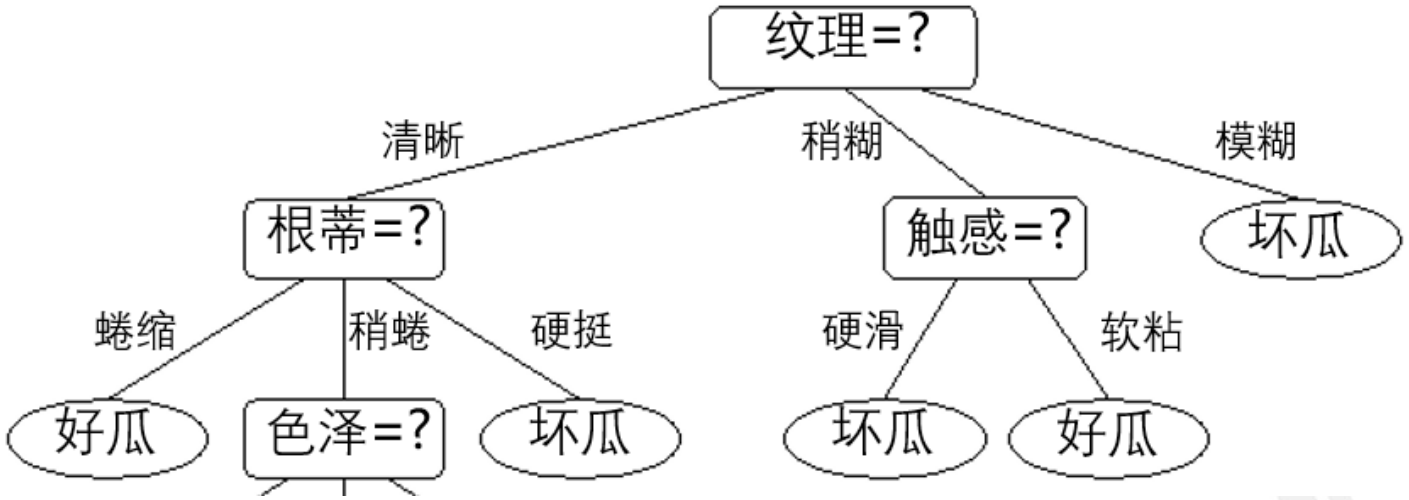
- 显然，属性“纹理”的信息增益最大，其被选为划分属性



7.2 监督学习

2.1 决策树的构造过程

- 对每个分支结点做进一步划分，最终得到决策树



7.2 监督学习

2.2 决策树的剪枝

研究表明: 划分选择的各种准则虽然对决策树的尺寸有较大影响，但对泛化性能的影响很有限

- 例如信息增益与基尼指数产生的结果，仅在约 2% 的情况下不同

  剪枝方法和程度对决策树泛化性能的影响更为显著

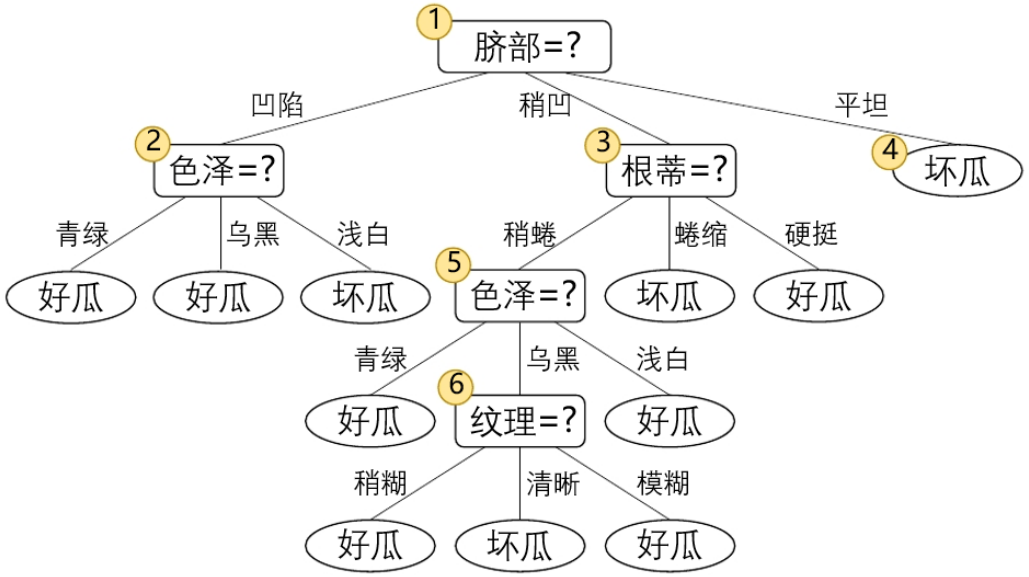
  剪枝 (pruning) 是决策树对付“过拟合”的主要手段!

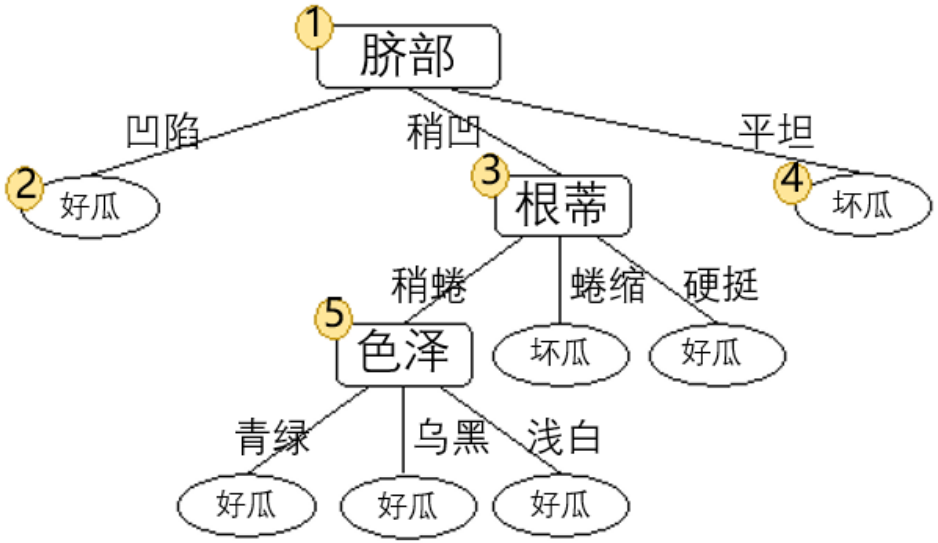
- 基本策略:
  - 预剪枝 (pre-pruning): 提前终止某些分支的生长
  - 后剪枝 (post-pruning): 生成一棵完全树，再“回头”剪枝

7.2 监督学习

2.2 决策树的剪枝：后剪枝

- 最终，后剪枝得到的决策树：



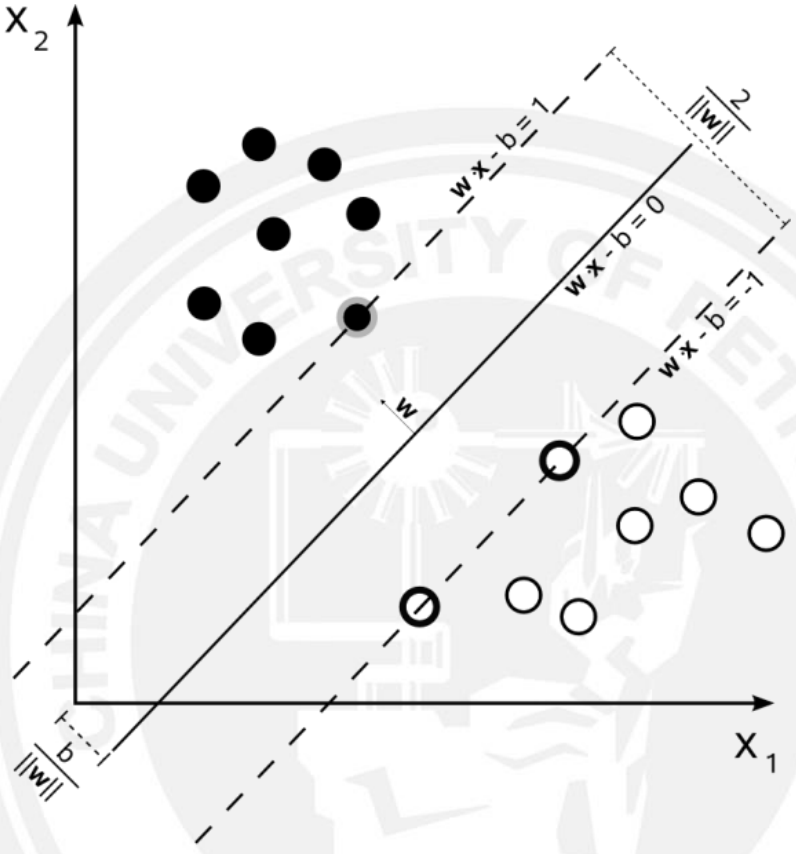


7.2 监督学习

3. 支持向量机

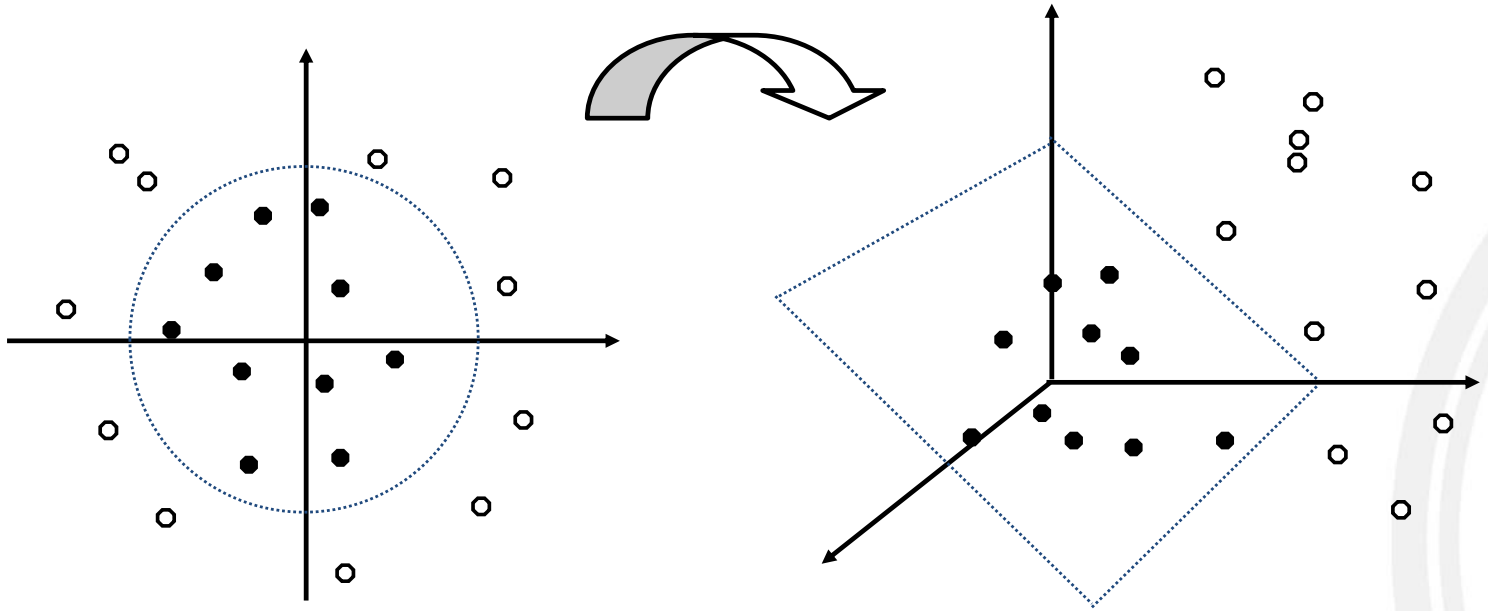
- 划分超平面
- 支持向量：离划分超平面最近的几个训练样本点。
- 间隔：两个类支持向量到超平面的距离之和。
- 支持向量机：

最大间隔准则(Maximize Margine)：支持向量机的目的就是找到具有最大间隔的划分超平面。



7.2 监督学习

3. 支持向量机-核方法



（1）线性不可分　　（2）线性可分

✓扩展低维特征到高维

✓ $\Phi(x)$：$x \rightarrow (1, x, x^2)$

$\Phi(x)^T\Phi(y)$

✓常用核函数

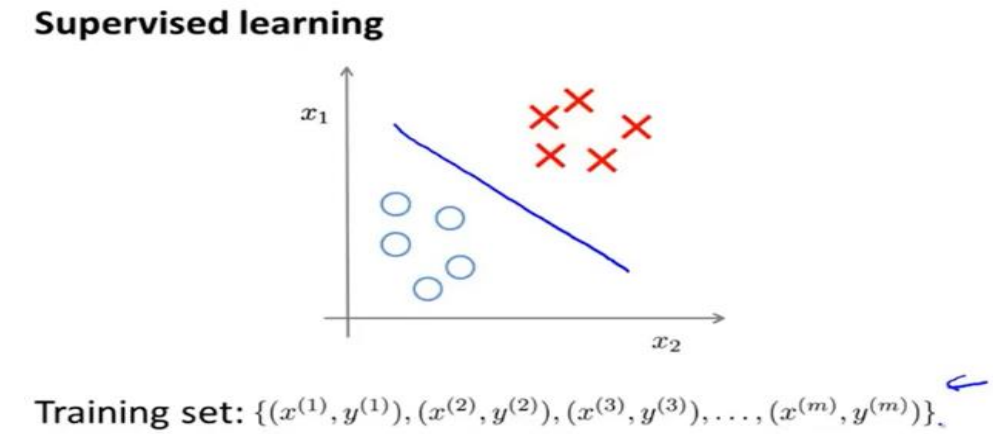
✓ 多项式核

✓ sigmoid核

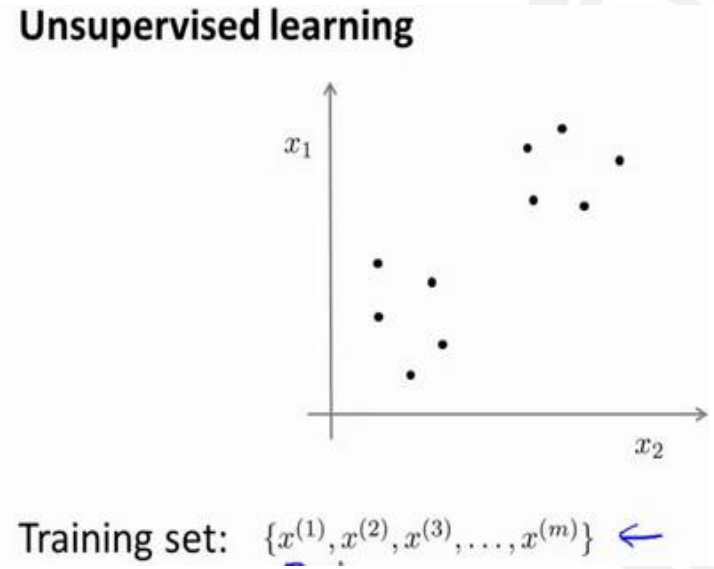
✓ 径向基核

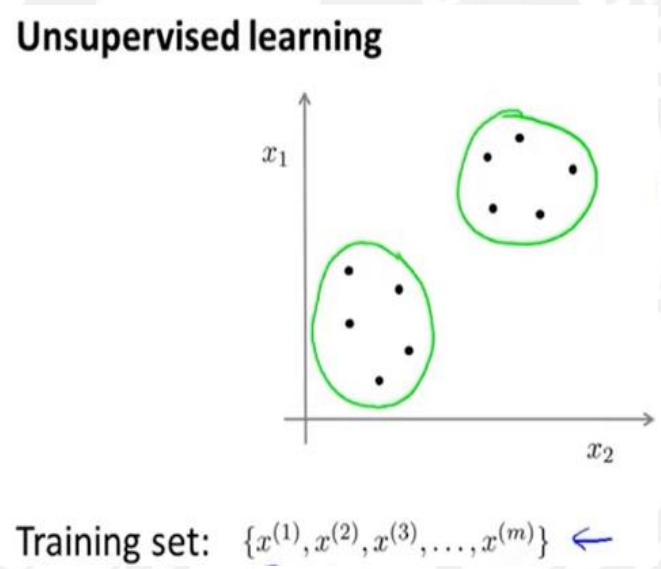
7.3 无监督学习

- 分类：事先确定有哪些类别，每一个数据有标签



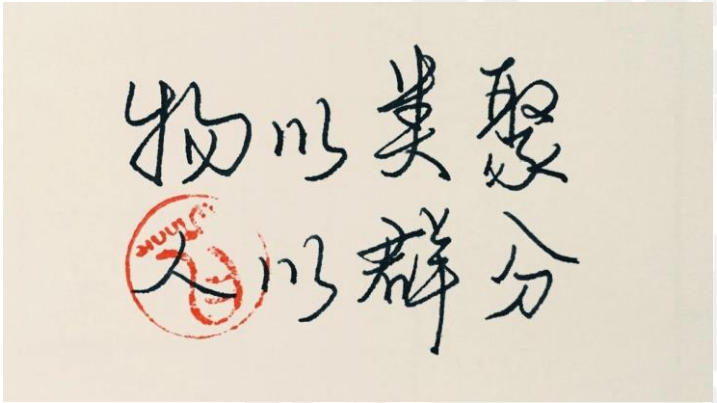
- 聚类：数据没有标签，根据一定的算法将相似数据聚到一类





7.3 无监督学习

- K-均值聚类



7.3 无监督学习

- 聚类
  - 聚类
    - ✓在同一个类中，数据对象是相似的
    - ✓不同类之间的对象是不相似的
  - 聚类算法
    - ✓根据给定的相似性评价标准，将一个数据集合分组成几个聚类
  - 一个好的聚类算法 - 聚类有效性
    - ✓聚类内部高相似性
    - ✓聚类之间低相似性

7.3 无监督学习

- 聚类
  - 算法流程：

1、初始化K个聚类中心

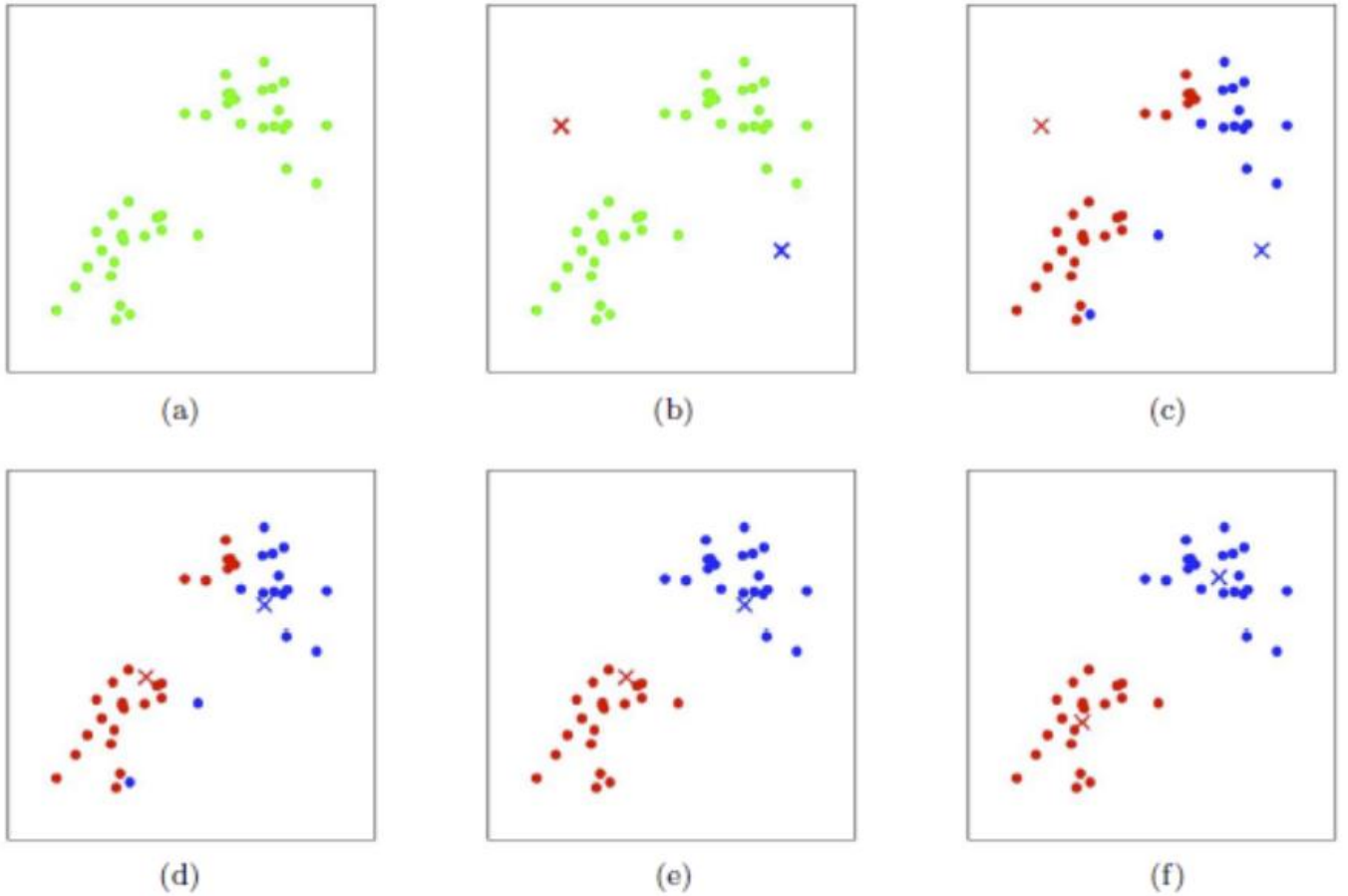
2、为每个个体分配聚类中心

$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

3、移动聚类中心

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

4、重复迭代, 直到聚类中心不变或者变化很小



7.3 无监督学习

- 聚类
  - 为了得到“簇内距离小，簇间距离大”的聚类效果，K-MEANS的做法是最终找到K个合适的聚类中心，来最小化平方误差 $J(c,\mu)=\sum_{i=1}^{m}||x^{(i)}-\mu_{c^{(i)}}||^2$

如何保证收敛

- J函数表示每个样本点到其质心的距离平方和。
- 假设当前J没有达到最小值，那么首先可以固定每个类的质心，调整每个样例所属的类别来让J函数减少
- 固定所属的类别，调整每个类的质心也可以使J减小。
- 这两个过程就是内循环中使J单调递减的过程。当J递减到最小时，聚类中心和所属类别同时收敛。

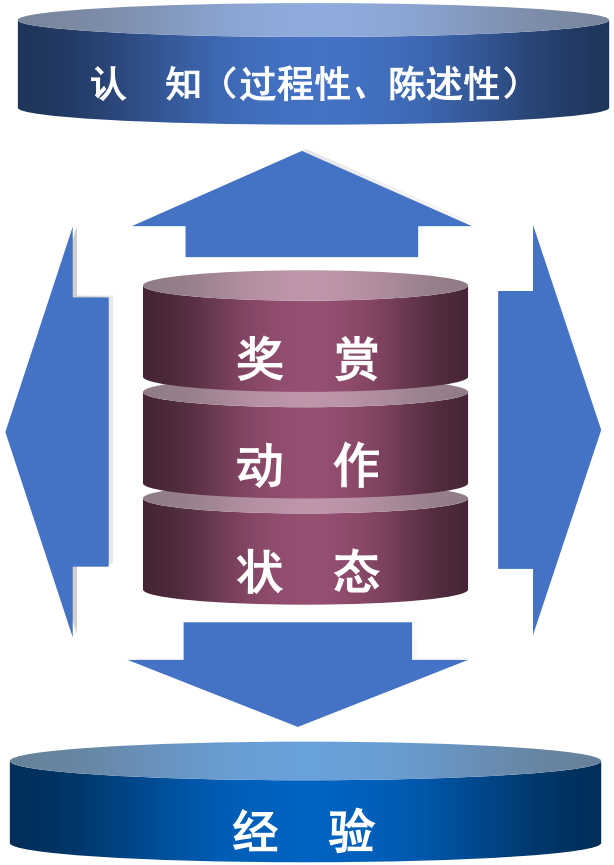
7.4 弱监督学习

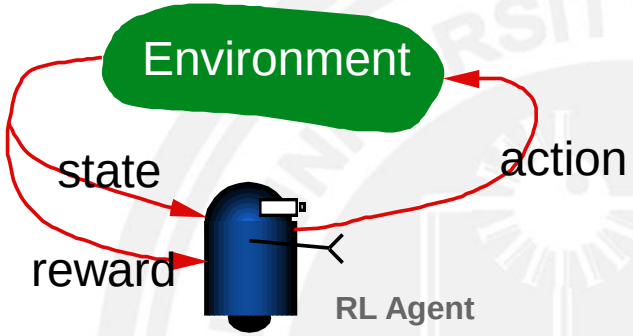
- 半监督学习(Semi-Supervised Learning)
  - 输入数据部分被标识，部分没有被标识，
  - 常用监督学习算法的延伸：图论推理算法，拉普拉斯支持向量机
- 迁移学习
  - 把为任务 A 开发的模型，重新使用在为任务 B 开发模型中
- 强化学习(reinforcement learning)
  - 输入数据直接反馈到模型，模型必须对此立刻作出调整

7.4 弱监督学习

- 强化学习

具体事例





抽象模型

强化学习的本质：奖惩和试错(Trial and error)

THANKS