Foundation of Artificial Intelligence

Prof. Fangshi Wang

Beijing Jiaotong University

Email: fshwang@bjtu.edu.cn

第4章 机器学习

4.1 机器学习的三个视角

4.1.1 机器学习的概念

4.1.2 机器学习的发展历史

4.1.3 机器学习的三个不同视角

4.1.4 机器学习的应用与术语

4.1.1 什么是机器学习

机器学习是人工智能的一个分支，是实现智能的关键。其目标是要构建可以从数据中学习、并对数据进行预测的系统。

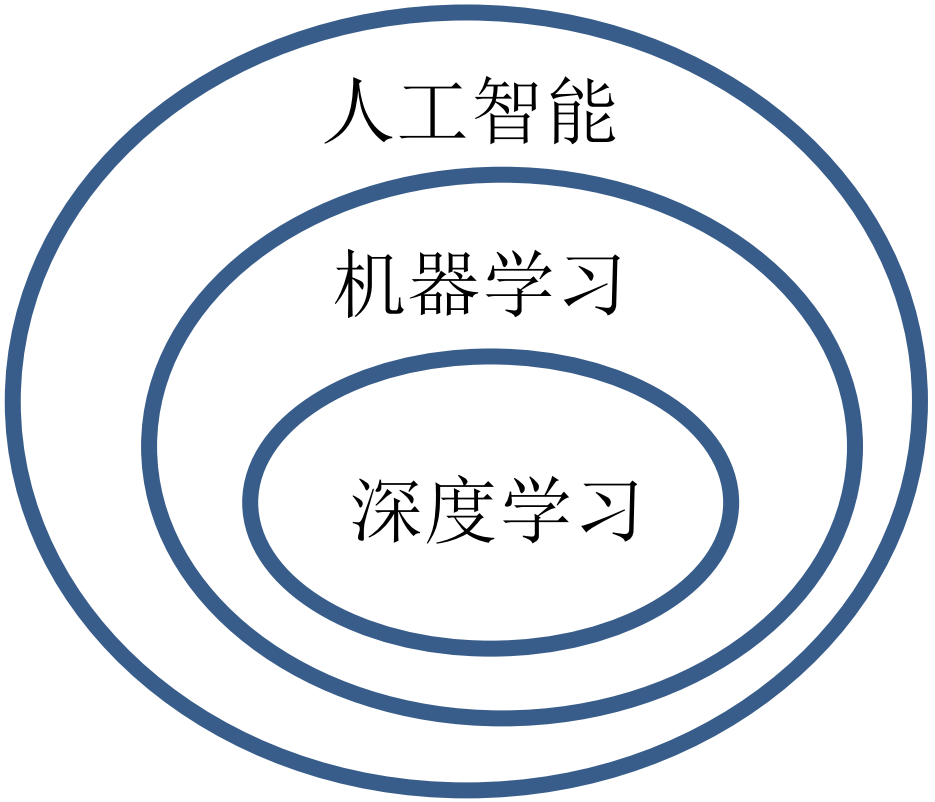
机器学习是对算法和数学模型的研究，而计算机系统则用它们来逐步提高其在特定任务上的性能。

Wikipedia

- 机器学习（Machine Learning）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 。
- 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。

百度百科

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**

Relations to Other Disciplines
与其他学科的关系

Statistical Learning 统计学习

- a machine learning framework drawing from statistics.
  取自于统计学的机器学习框架。

Pattern Recognition 模式识别

- the recognition of patterns in data. (≈ machine learning + data patterns)
  识别数据中的模式。(≈ 机器学习 + 数据模式)

Data Mining 数据挖掘

- the discovery of unknown properties in data. (≈ machine learning + database)
  发现数据中的未知特性。(≈ 机器学习 + 数据库)

Computer Vision 计算机视觉

- to extract information from images. (≈ machine learning + image processing)
  从图像中提取信息。(≈ 机器学习 + 图像处理)

数据挖掘、机器学习和统计学习的关系

- **统计学习**：是**机器学习**和**数据挖掘**这两门技术的基础，更偏重于理论上的完善；**统计学习**主要是通过机器学习为数据挖掘提供算法支撑。
- **机器学习**：是统计学习对实践技术的延伸，更偏重于解决**小数据量**的问题，为其提供算法技术的支撑；
- **数据挖掘**：更偏重于**大数据**的实际问题，更注重实际问题的解决，包括真实数据的数据清洗、建模、预测等操作。**机器学习**和**数据库**则是**数据挖掘**的两大支撑技术。
- **机器学习**可以分为以支持向量机为代表的**统计学习**和人工神经网络为代表的**联结主义学习**。
- 统计学习模型中的参数往往是可解释的，而人工神经网络就是一个黑箱。

机器学习与各领域之间的关系

- **模式识别**≈机器学习。PR源自工业界、偏应用，ML源自计算机学科、偏算法研究。
- **数据挖掘**=机器学习+数据库。目的是从大量数据中通过算法搜索隐藏于其中的信息。大部分数据挖掘算法，是机器学习算法在数据库中的优化。
- **统计学习**⊂机器学习。一个偏数学理论研究，一个偏算法研究。
- **计算机视觉**=机器学习+图像处理。CV就是用机器模拟生物视觉。图像处理负责给机器学习模型提供输入，机器学习负责学习并给出视觉结果。
- **语音识别**=机器学习+语音处理。其目的就是使机器能与人进行语音交流，把语音信号转变为相应的文本或命令。
- **自然语言处理**=机器学习+文本处理（如机器翻译）。

Review of Lecture 7

1. 爬山法、随机爬山法、随机重启爬山法共同的特点？完备吗？各自的特点？
2. 模拟退火算法完备吗？与爬山法的根本不同点？
3. 简述模拟退火算法的思路
4. 用遗传算法解决8皇后问题，适应度函数如何定义的？
5. 爬山法、模拟退火法、遗传算法的解释
6. 什么是机器学习？它与AI、深度学习的关系如何？
7. AI应用成功的三要素是什么？

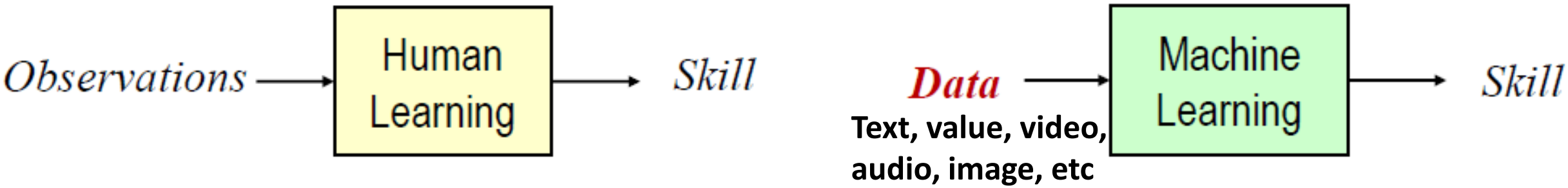
人工智能与机器学习

◆ **人类学习**

人类是从**观察**中积累**经验**来获取技能。

◆ **机器学习**

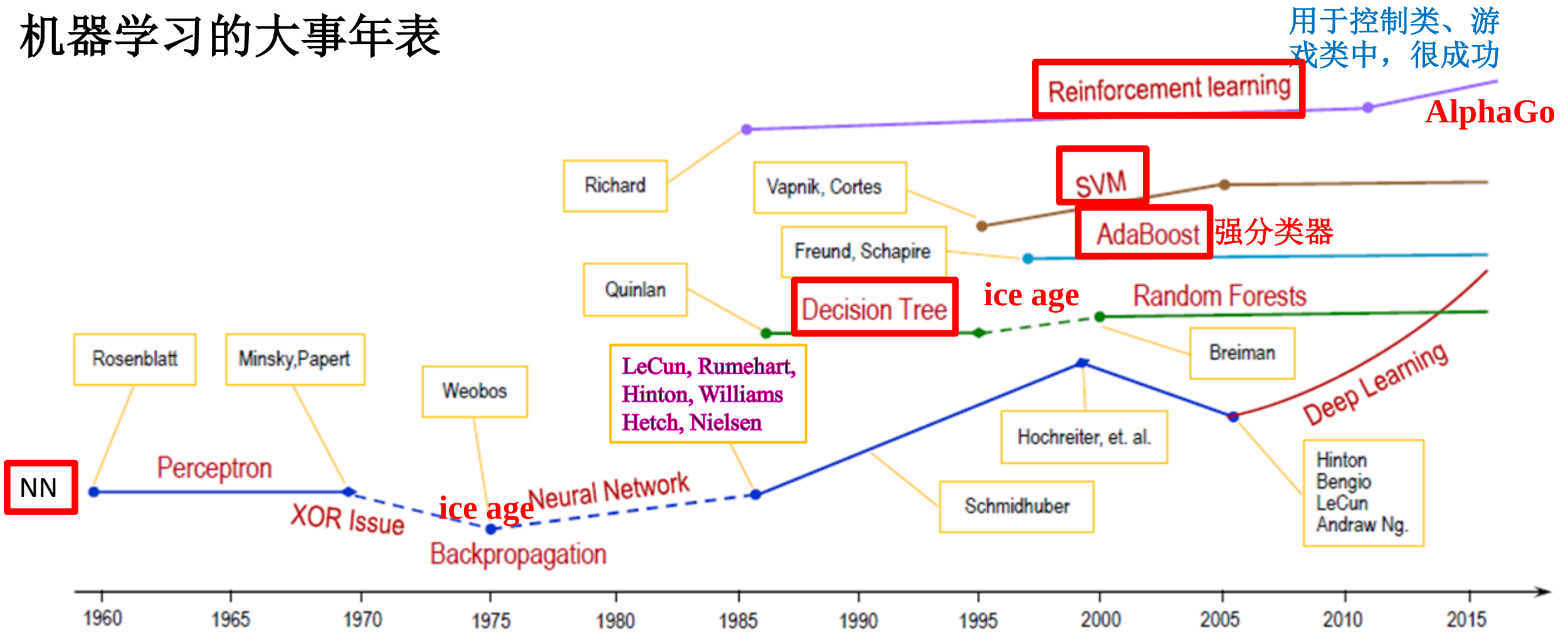
机器是从**数据**中积累或者计算的**经验中**获取技能。



机器模拟人类的学习行为.

4.1.2 机器学习的发展历史

机器学习的大事年表



Perceptron	**XOR Issue**	**Backpropagation**	**Neural Network**	**Reinforcement Learning**	**Decision Tree**	**SVM**	**AdaBoost**	**Random Forests**	**Deep Learning**
an algorithm of binary classifiers	perceptrons can't process XOR circuit	a common method of training neural networks	learning algorithms inspired by biological neural networks	concerned with how an agent take actions in an environment	a decision support tool that uses a tree-like graph	support vector machine	adaptive boosting algorithm	a multitude of decision trees	deep networks, hierarchical learning

机器学习的三个学派

Connectionism 联结主义

also called Bionicsism, Physiologism.
亦称仿生主义、生理学派

e.g., Perceptron, Artificial Neural Network, Deep learning.

Symbolicism 符号主义

also called Logicism, Psychologism, Computerism.
亦称逻辑主义、心理学派、计算机主义

e.g., Association Rules, Decision tree, Random Forests

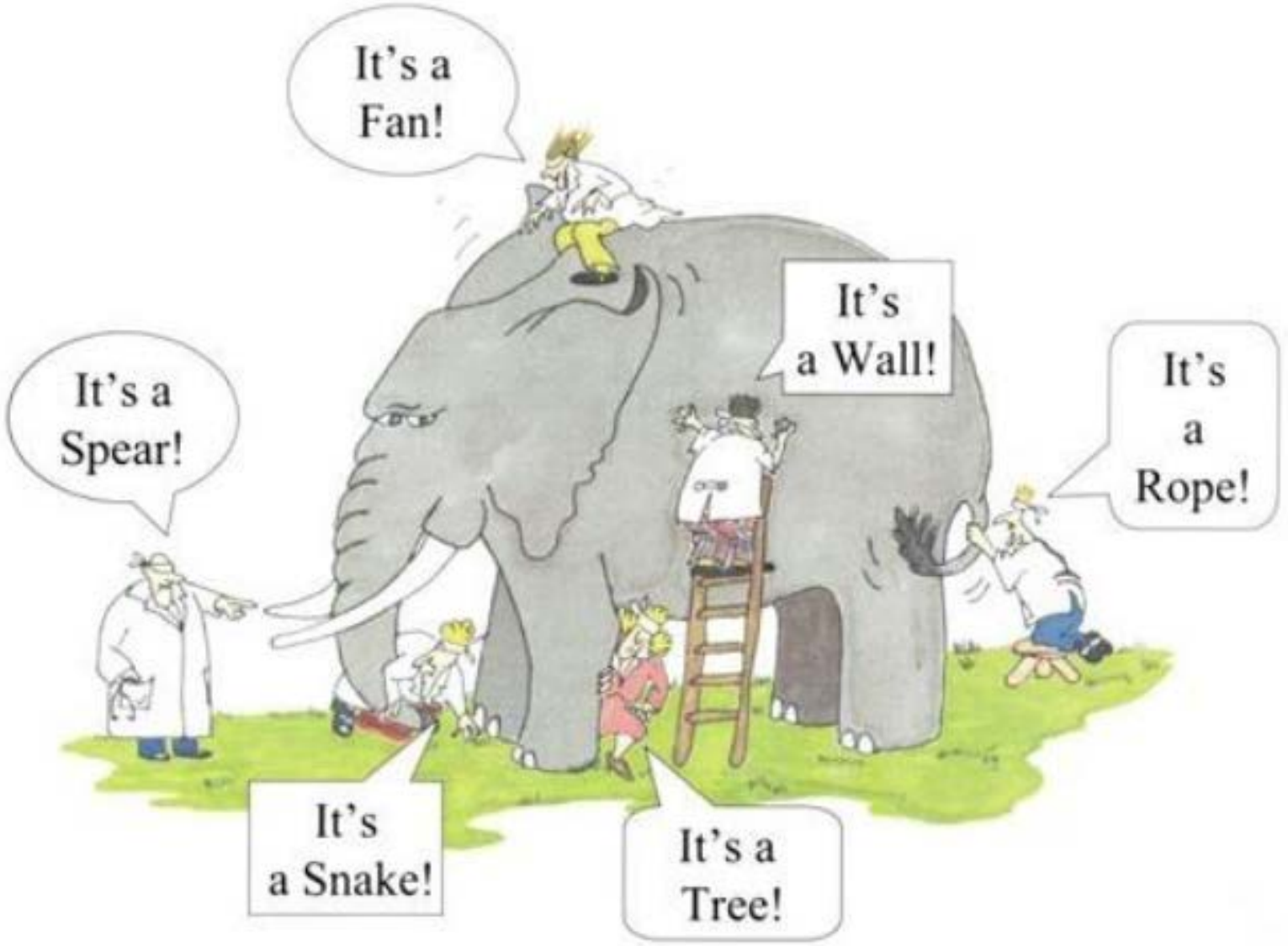
Behaviorism 行为主义

also called Actionism, Evolutionism, Cyberneticsism
亦称行动主义、进化主义、控制论学派

e.g., Reinforcement learning

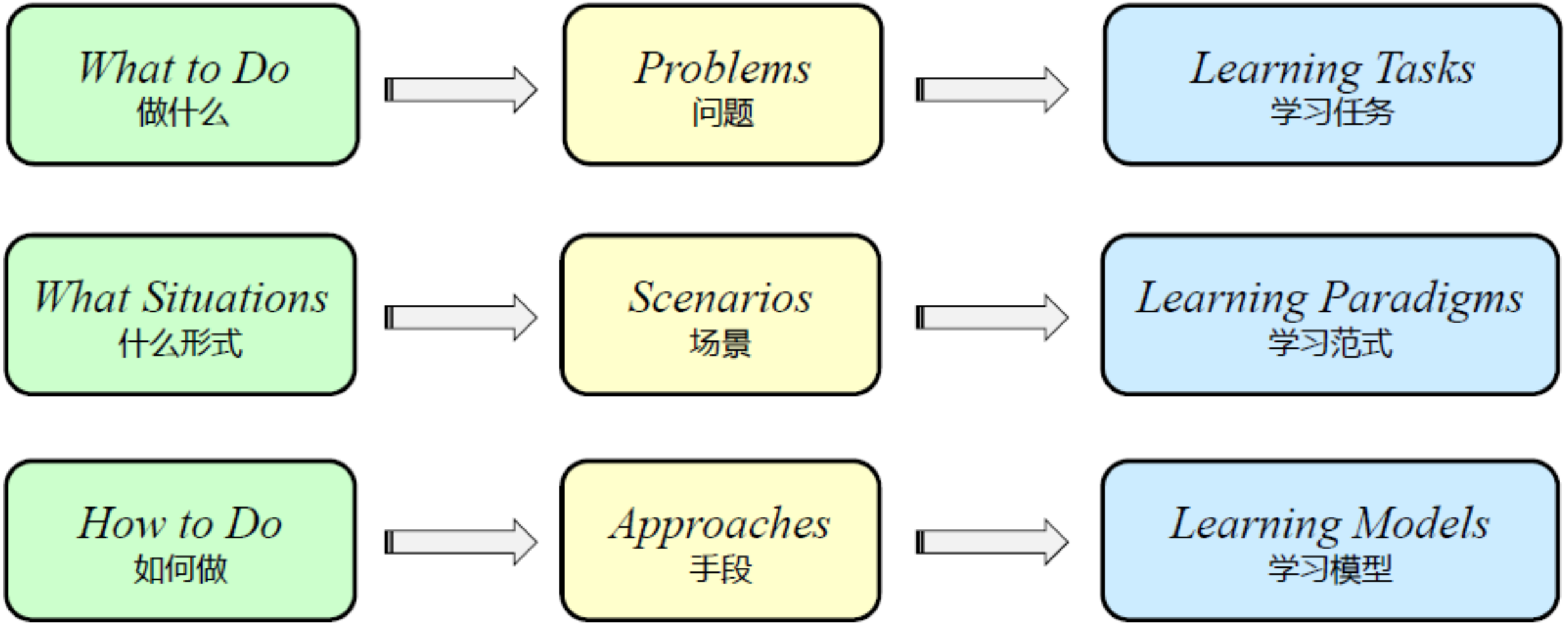
根据回报值采取相应行动

4.1.3 机器学习的三个不同视角



Maybe "Blind Men and an Elephant"

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



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	表示机器学习中问题发生的**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）**。

4.1.3 机器学习的三个不同视角

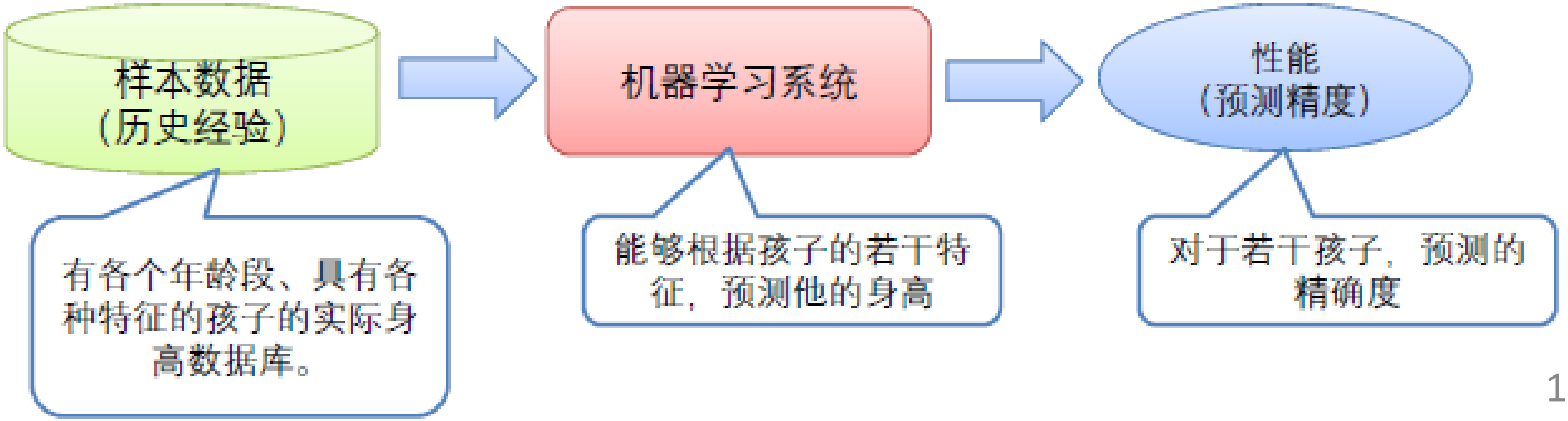
（1）什么是学习任务: 用于表示可以用机器学习解决的通用问题。

Tasks 任务	Brief Statements 简短描述	Typical algorithm 典型算法
Classification 分类	Inputs are divided into two or more known classes. 将输入划分成两个或多个类别。	SVM 支撑向量机
Regression 回归	Outputs are continuous values rather than discrete ones. 输出是连续值而不是离散的。	Bayesian linear regression 贝叶斯线性回归
Clustering 聚类	Inputs are divided into groups which are not known beforehand. 输入被划分为若干个事先未知的组。	*k*-means *k*-均值
Ranking 排名	Data transformation in which values are replaced by their rank. 用它们的排名来代替值的数据转换。	PageRank 网页排名
Density estimation 密度估计	Find the distribution of inputs in some space. 寻找某个空间中输入的分布。	Boosting Density Estimation 增强式密度估计
Dimensionality reduction 降维	Simplify inputs by mapping them into a lower dimensional space. 通过将输入映射到低维空间来将其简化。	Isomap 等距特征映射
Optimization 优化	Find the best solution from all feasible solutions 从所有可能的解中寻找最优解。	Q-learning Q-学习

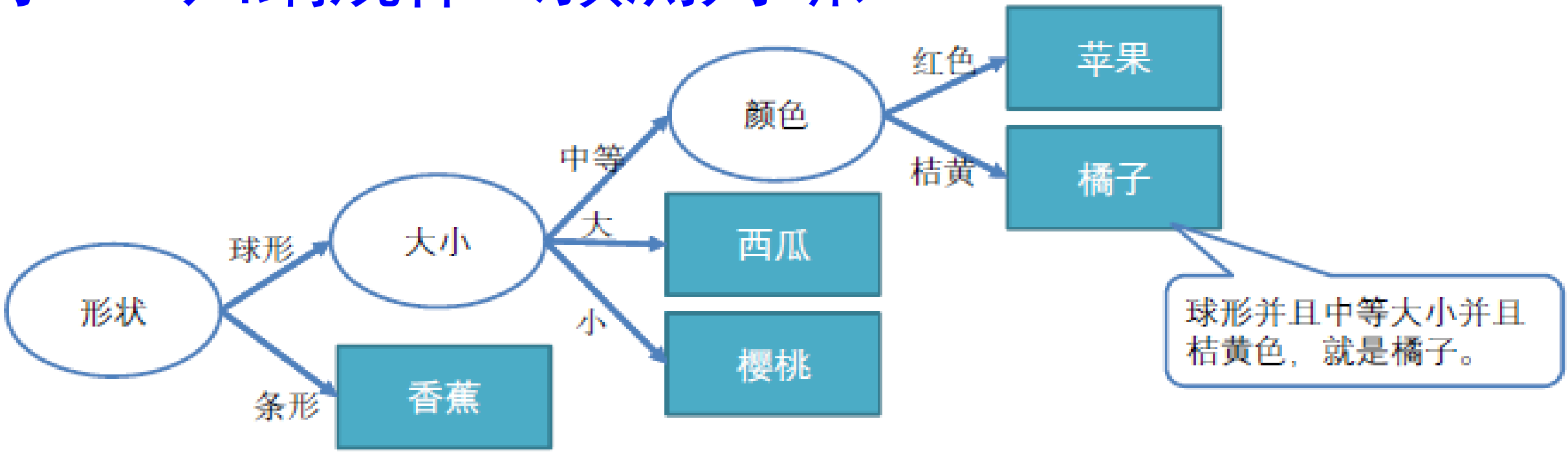
例4.1 图像分类的机器学习

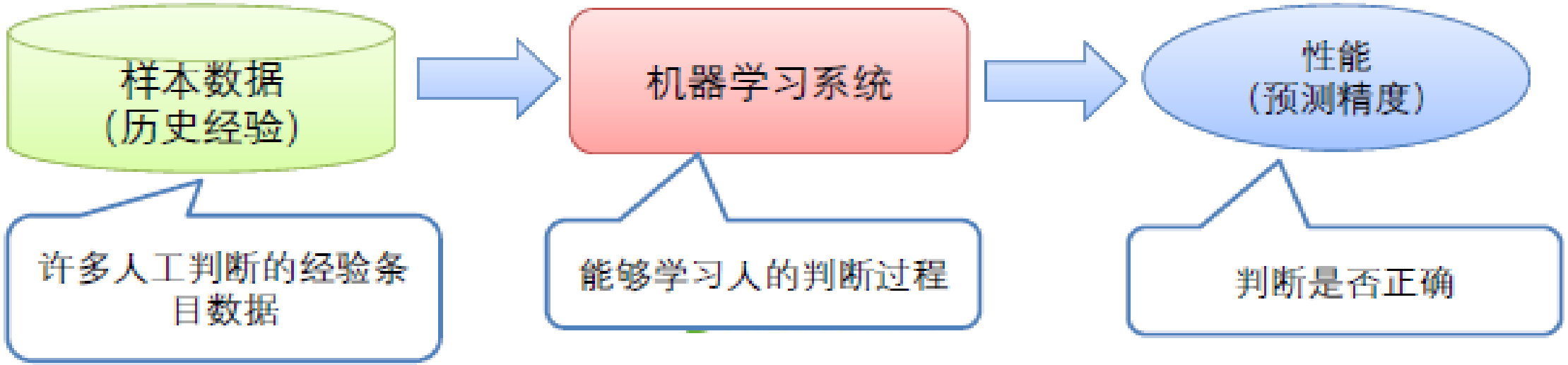
例4.2 身高预测（预测数值）

如何根据年龄、性别、体重等等特征，来预测小孩的身高？



例4.3 归纳规律（预测判断）

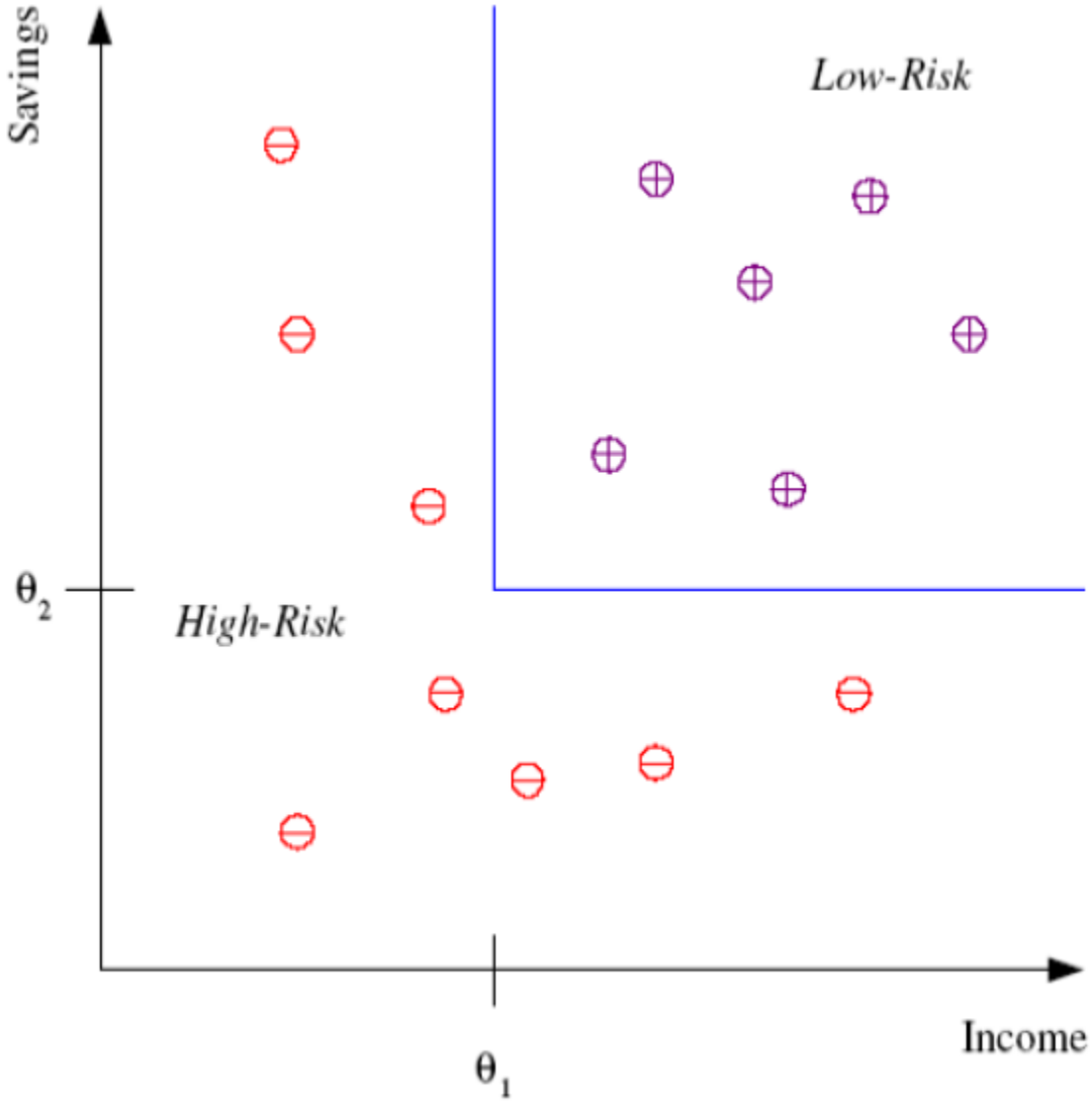




例4.4 信用评分

- **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*



(2) Learning Paradigms 学习范式

- 学习范式/类型：是用于表示机器学习中发生的典型场景。
- 根据机器学习的典型场景或样式，区分学习范式
  它怎样从数据（有标注/无标注）中学习，它如何同环境互动。

Paradigms 范式	Brief Statements 简短描述	Typical Algorithm 典型算法
Supervised 有监督 有老师教	The algorithm is trained by a set of labeled data, and makes predictions for all unseen points. 算法采用一组标注数据进行训练，再对所有的未知点做出预测。	Support vector machines 支撑向量机
Unsupervised 无监督 自学	The algorithm exclusively receives unlabeled data, and makes predictions for all unseen points. 算法仅接收未标注的数据，再对所有的未知点做出预测。	*k*-means *k*-均值
Reinforcement 强化	The algorithm interacts with environment, and receives an reward for each action 算法与外部环境交互，每个动作得到一个回报。 适用于博弈类游戏	Q-learning

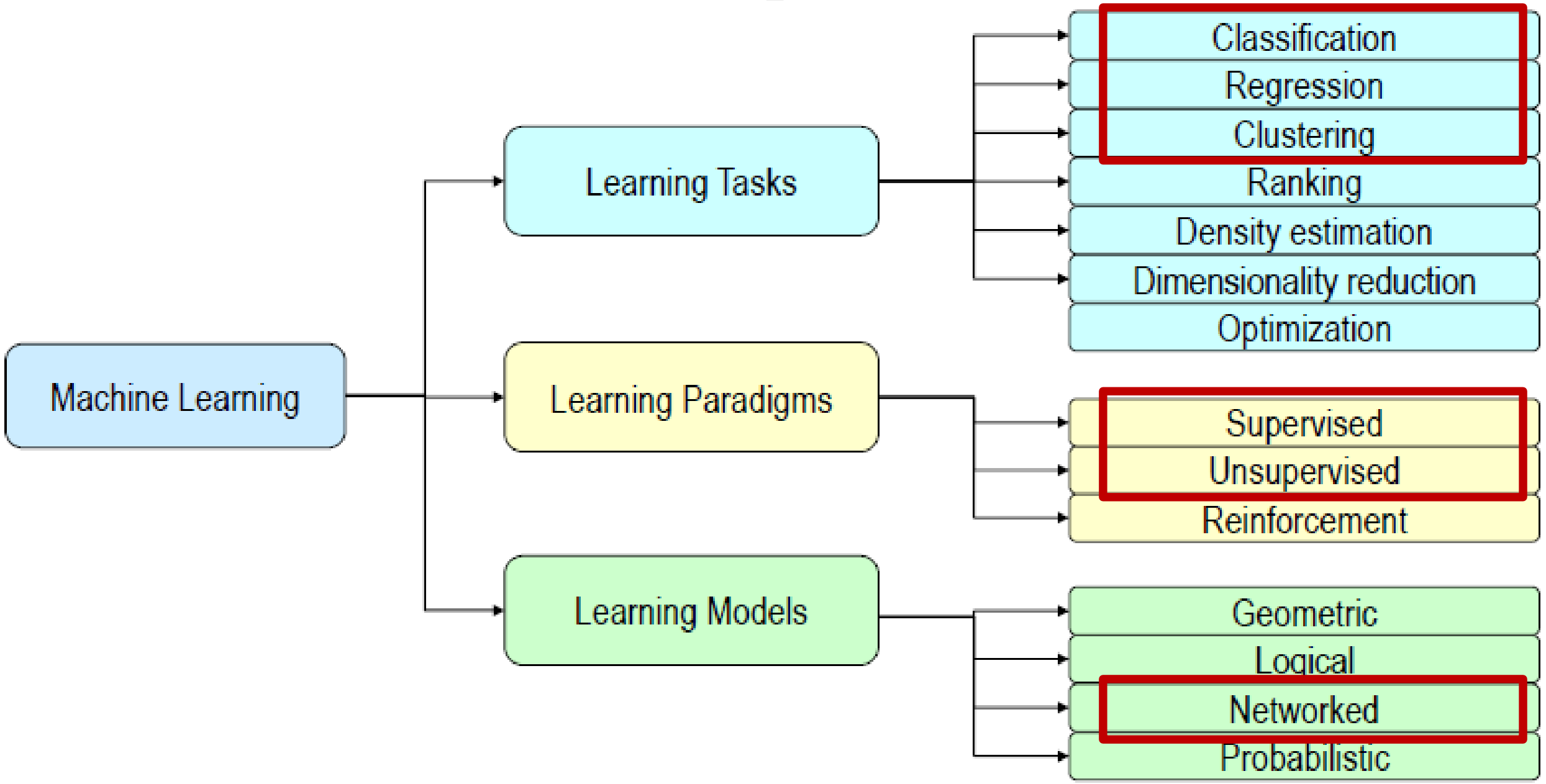
(3) Learning Models 学习模型

◆学习模型用于表示可以完成一个学习任务的方法。

◆**机器学习的效果**在很大程度上取决于解决该学习任务时所选用的方法.

Models 模型	Brief Statements 简短描述	Sub-models 子模型	Typical Algorithm 典型算法
Geometric 几何	Use geometric models such as line, plane, distance or manifold to construct learning algorithms. 采用线、面、距离或流行等几何图形模型来构建学习算法。	Line 线	Linear Regression 线性回归
		Plane 面	SVM 支撑向量机
		Distance 距离	*k*-NN *k*-近邻
		Manifold 流行	Isomap 等距映射
Logical 逻辑	Use logical models to construct learning algorithms. 采用逻辑模型来构建学习算法。	Logic 逻辑	Inductive Logic Program. 归纳逻辑编程
		Rule 规则	Association Rule 相关规则
Networked 网络	Use networked models to construct learning algorithms. 采用网络模式构建机器学习算法。	Shallow 浅层	Perceptron 感知机
		Deep 深层	CNN 卷积神经网络
Probabilistic 概率	Use probabilistic models to denote the conditional dependence between random variables. 采用概率模式来表示随机变量之间的条件相关性。	Bayes 贝叶斯	Bayesian Network 贝叶斯网络
		Generative 生成	Probabilistic Program. 概率规划
		Statistic 统计	Linear Regression 线性回归

The Three Perspectives 三个视角



4.1.4 Applications and Terminologies

(1) 机器学习的应用领域

- Machine Perception 机器感知
- Computer Vision 计算机视觉
- Video Analysis 视频分析
- Pattern Recognition 模式识别
  - Face/Speech/Fingerprint Recognition 人脸/语音/指纹识别
  - Optical Character Recognition (OCR) 光学字符识别 (OCR)
  - Handwriting Recognition 手写体识别
- Game Playing 玩游戏
- Natural Language Processing 自然语言处理
- Information Retrieval 信息检索

(1) 机器学习的应用领域

- Text or Document Classification (e.g. Spam Email Detection) — 文本与文档分类（例如垃圾邮件检测）
- Recommender Systems — 推荐系统
- Ad Placement — 广告配置
- Credit Scoring — 信用评分
- Fraud Detection — 欺诈检测
- Stock Trading — 股票交易
- Drug Design — 新药设计
- Medical Diagnosis — 医学诊断
- Robotics — 机器人学

(2) 机器学习中的一些术语

- **Samples 样本**

  用于学习或评估的数据项或实例。

- **Features 特征**

  属性集，通常表示为与样本相关的向量：

  - Handcrafted features: 手工式特征

    e.g., SIFT, HOG, SURF, LBP, GLOH, LESH, CENTRIST.

  - Learned features: 学习式特征

    例如：通过卷积神经网络获得的特征。

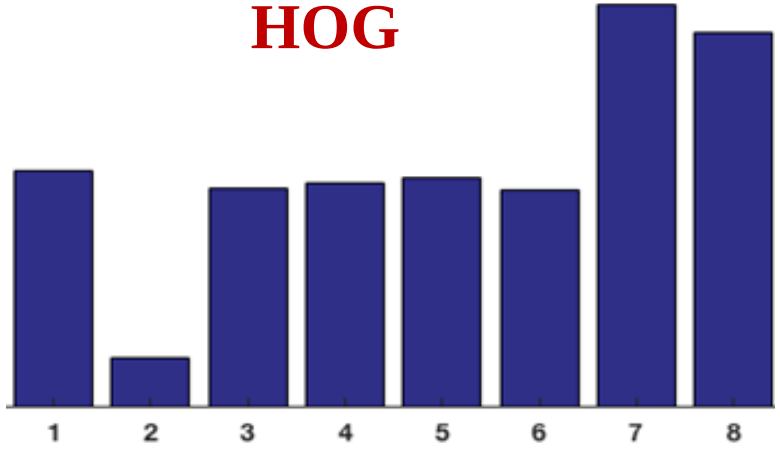
(2) 机器学习中的一些术语

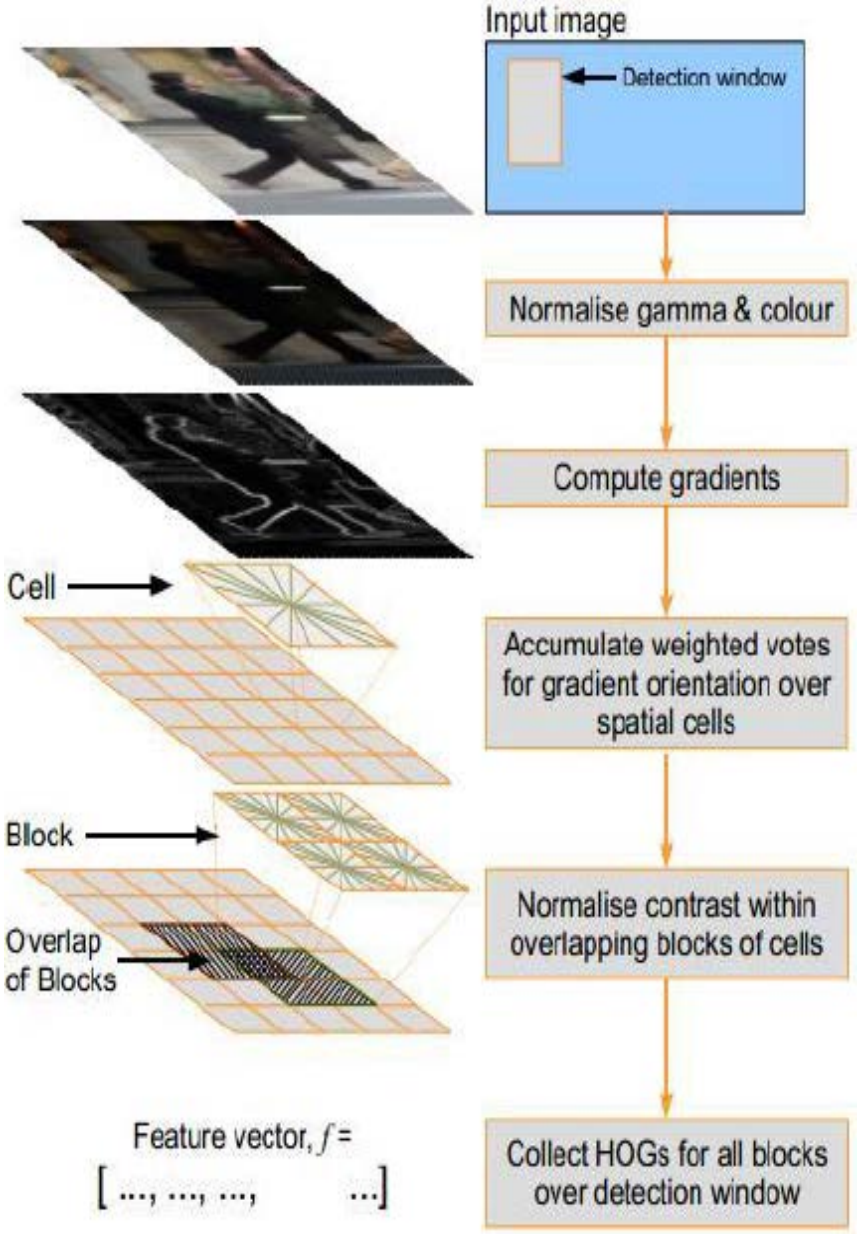
◆ Handcrafted Features /Designed Features

手工式特征 / 设计的特征

- 研究者设计的特征，称为手工式特征.
- e.g. HOG (Histogram of Oriented Gradients，定向梯度直方图)
- 按照强度梯度或边缘方向分布。
- 64×128 检测窗口

HOG



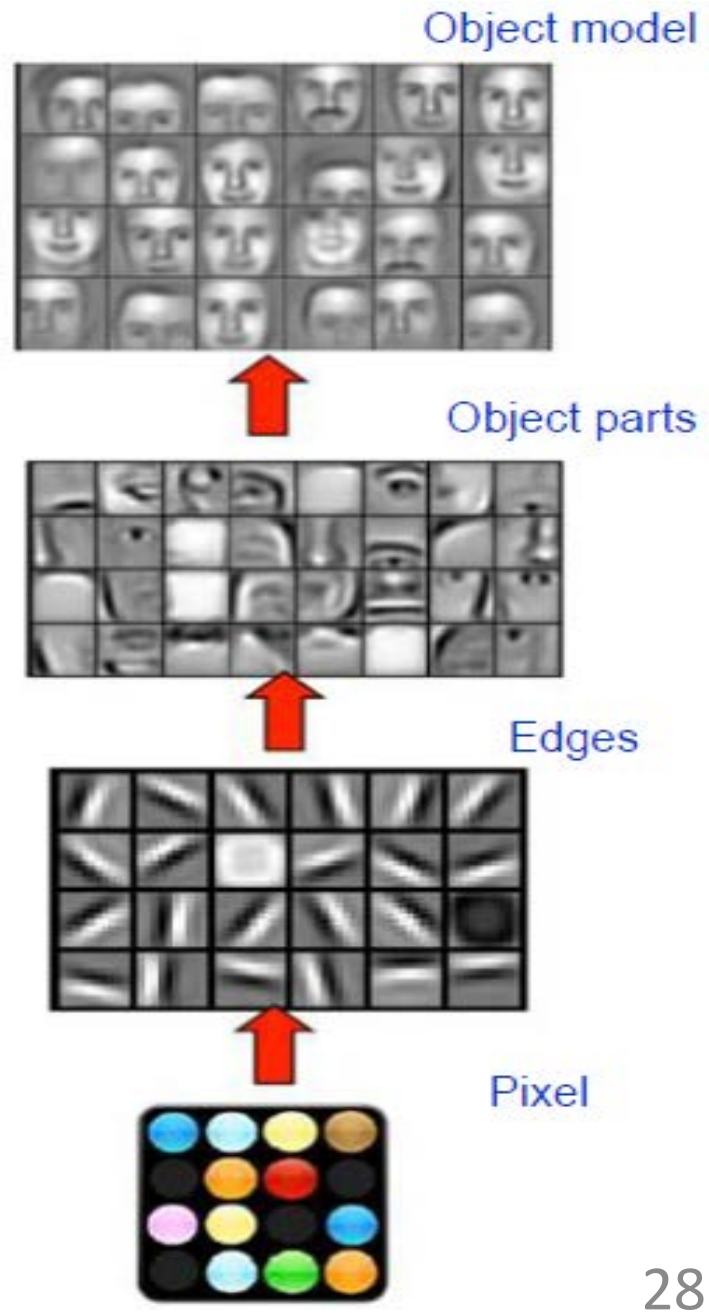


flow chart

(2) 机器学习中的一些术语

◆ Learned Features 学习式特征

- 人类可以有效地学会观察。因为大脑是深度的，具有许多处理层次。
- 具有这种深度架构的算法，能从原始数据中自动生成视觉特征。
- **特征学习**也被称为**表示学习**。
- 理解深度学习将使我们能够构建更智能的视觉认知机器。



(2) 机器学习中的一些术语

- **Labels 标记:** 在样本上指定的值或类别。
  - 分类问题中，标记就是样本被指定的特定类别。
  - 回归问题中，标记就是项被指定的实值。
- **Training sample 训练样本:** 用于训练学习算法的样本。
  - 对于垃圾邮件问题，训练样本由一组邮件样本以及相关标签组成。
- **Validation sample 验证样本**
  - **验证样本**是用于调整学习算法**超参数**的、已标注的数据。
  - 学习算法通常具有一个或多个**自由参数**，因而验证样本用于为这些模型的**超参数**（如网络层数、网络节点数、迭代次数、学习率 或 KNN中的k）选择适当的值。
- **Test sample 测试样本:**
  - 测试集既不参与参数的学习过程，也不参与参数的选择过程，仅用于模型评价.
  - 用于**评估学习算法性能**的样本。
  - 然后将这些预测与测试样本的标签进行比较，以衡量算法的性能。

(2) 机器学习中的一些术语

◆ Loss function 损失函数

- 用于度量预测标签和真实标签之间差异或损失。
- 将所有真实的标签集表示为Y， 将可能的预测集表示为Y’，则损失函数L为映射：

$$L: Y \times Y' \rightarrow \mathbb{R}+$$

◆ Hypothesis set 假设集（即函数集）

- **假设集**是将**特征**映射到**标签集**Y上的函数集。
- 例如，将电子邮件特征映射到Y上的函数集：

$$Y=\{spam, non-spam\}.$$

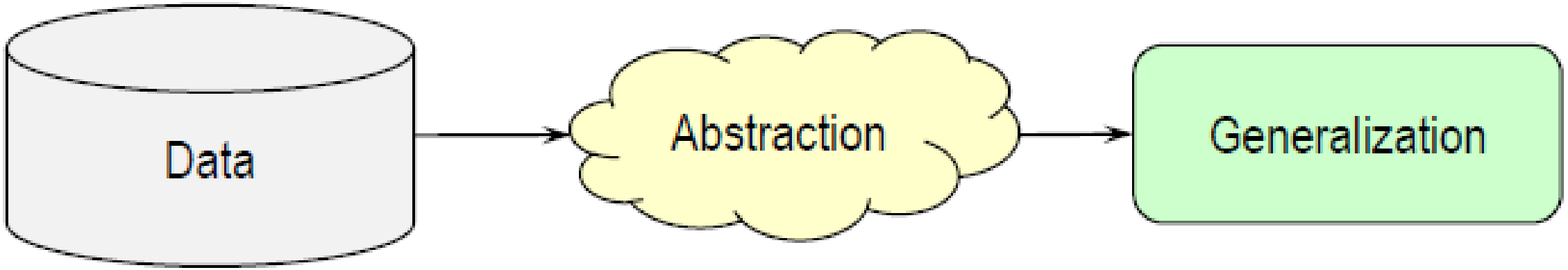
(2) 机器学习中的一些术语

- **Abstraction 抽象**

  其含义是将数据转化为更广泛的表示。

- **Generalization 泛化**

  它形容将抽象知识转化为可用于动作形式的过程。它也是学习算法具有学习数据集的经验后，可以**对未知样本正确地进行处理的能力**。



4.2 机器学习的任务

4.2.1 Classification 分类

4.2.2 Regression 回归

4.2.3 Clustering 聚类

4.2.4 Ranking 排名

4.2.5 Dimensionality Reduction 降维

4.2.1 Classification

- 分类
- 线性和非线性分类
- 维度与类别
- 应用与算法

(1) 分类 Classification

◆ **什么是分类**

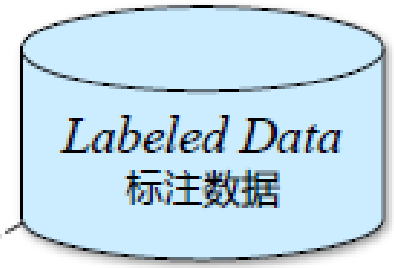
为每个数据项指定一个类别。

◆ **classifier 分类器**

用于分类的**算法**，称为**分类器**。

Classification: Training 分类：训练

Known Categories
已知类别



x_1	y_1
x_2	y_2

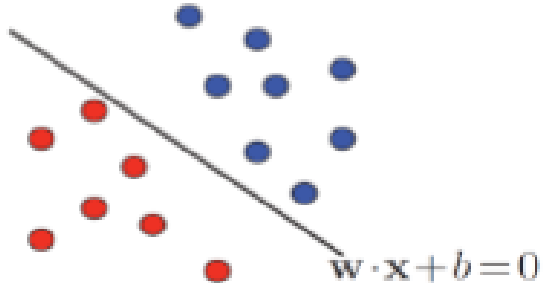
Labeling function
标注函数

Training
训练

(x, y)

Learning Algorithm
学习算法

$f(\mathcal{X})$

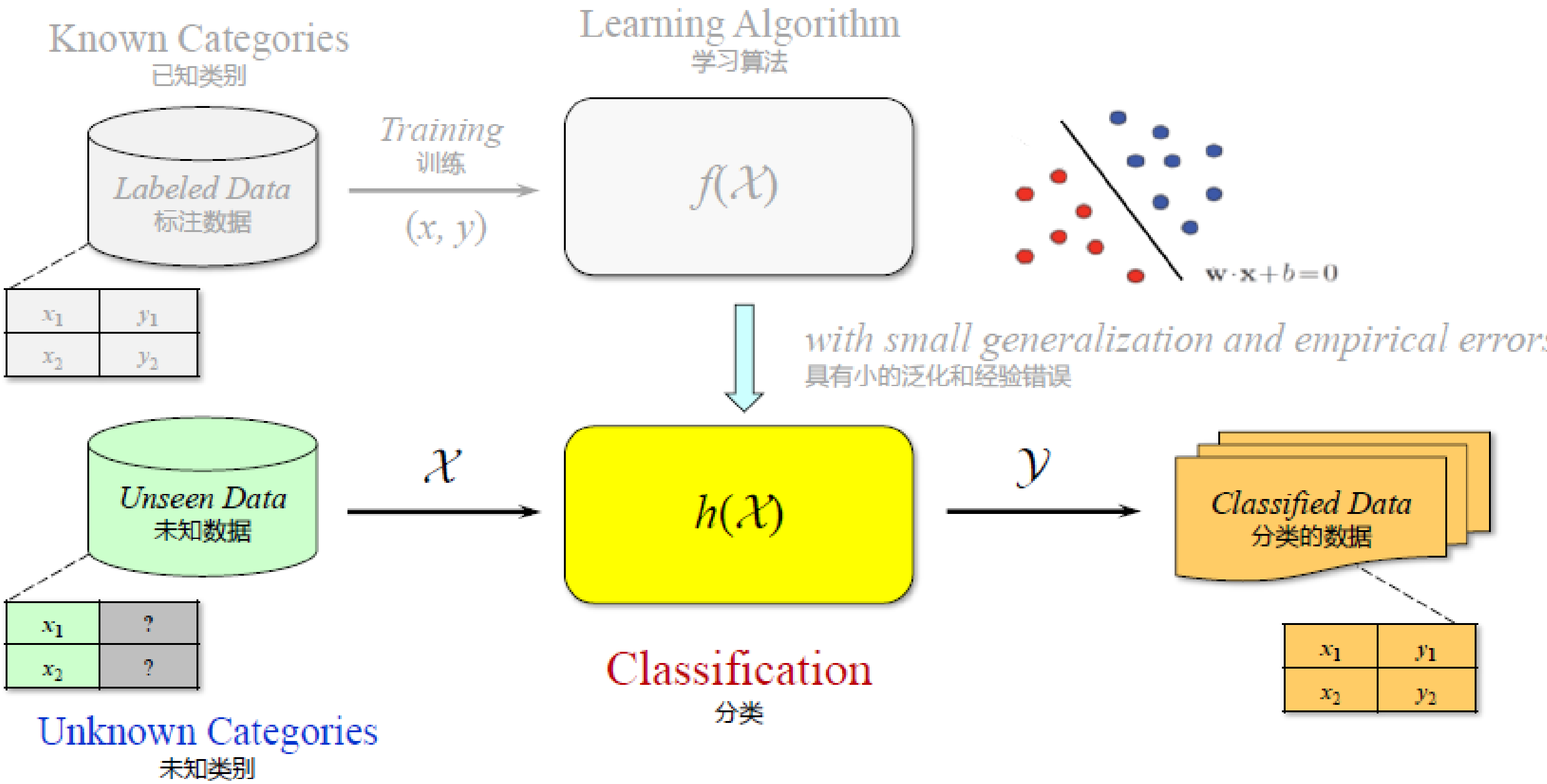


with small generalization and empirical errors
具有小的泛化和经验错误

$h(\mathcal{X})$

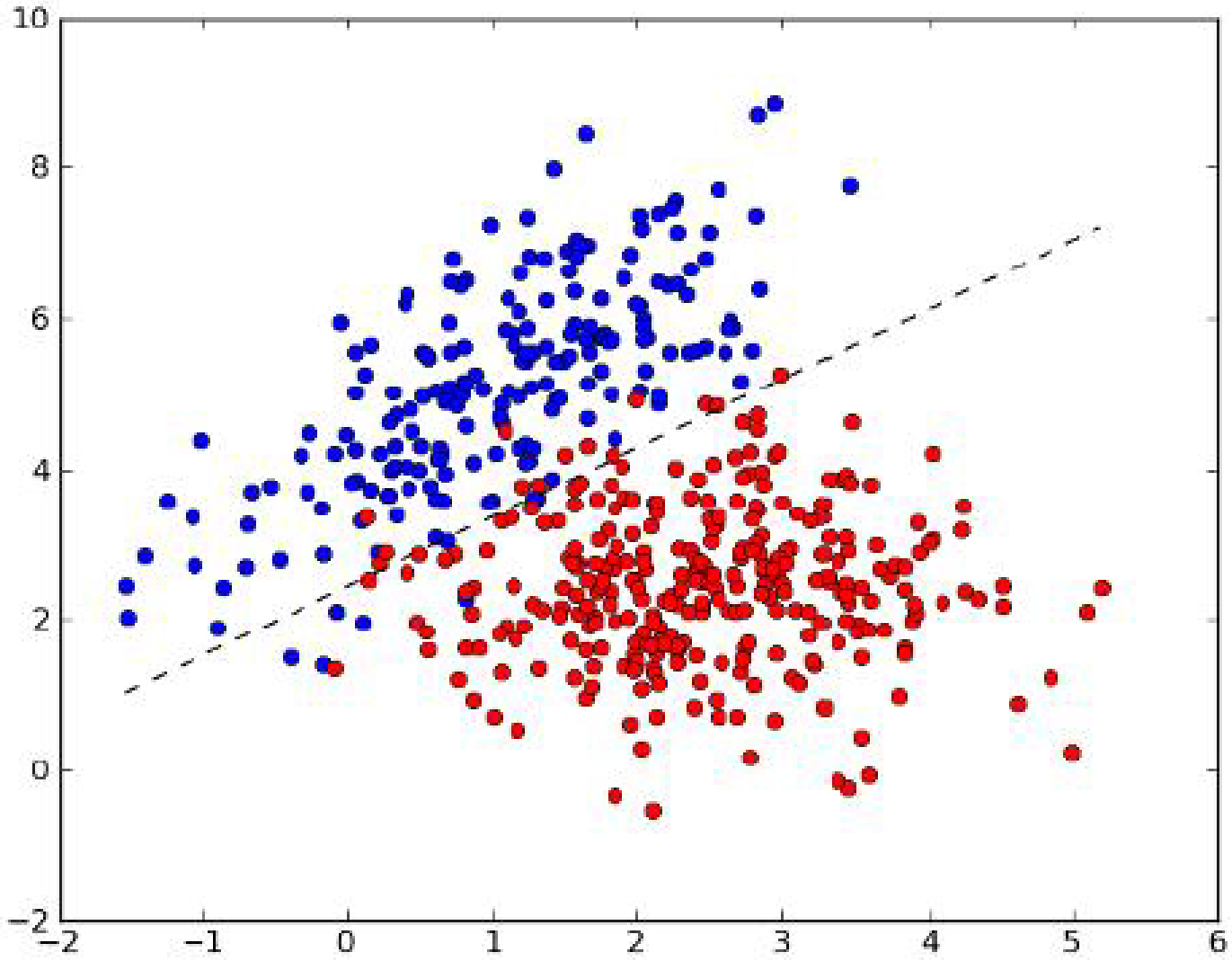
Hypothesis
(Classifier function)
假设（分类函数）

Classification: Training 分类：训练



（2）线性和非线性分类

- **Linear Classification**
  - **线性分类**是通过线性分类器来进行分类。
  - 一个线线性分类器就是一个线性判别函数。



Case Study: A Typical Linear Classifier
一个典型的线性分类器

$$H = \{\mathbf{x} \mapsto y(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \mid \mathbf{w} \in R^n, b \in R\}$$

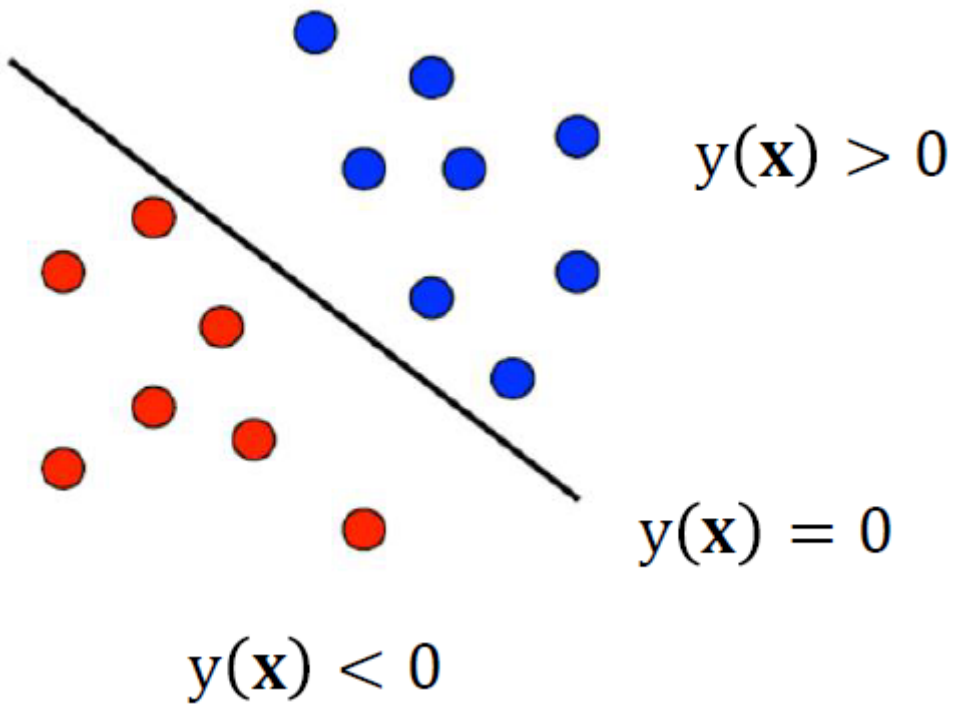
其中,**W**表示行向量, 称为权向量。

$$\mathbf{w} = (w1, \ldots, wn)$$

X表示列向量，称为输入向量。

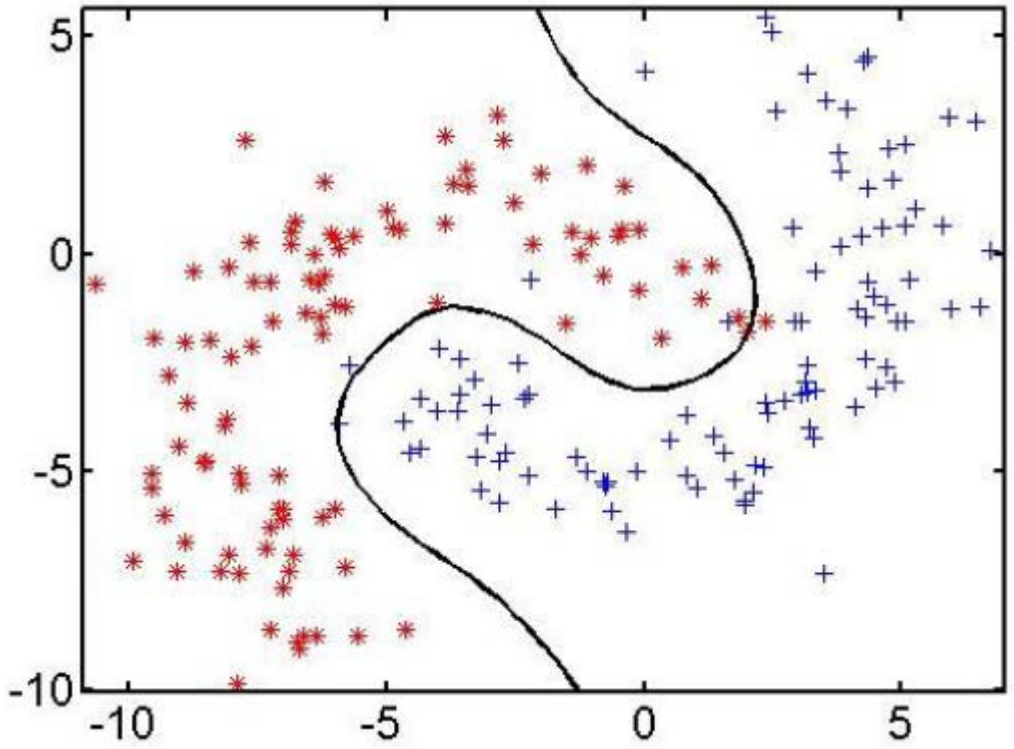
$$\mathbf{x} = (x1, \ldots, xn)^T$$

*b*表示偏差。



◆Nonlinear Classification 非线性分类

- 非线性分类是通过一个非线性分类器来进行分类。
- 一个非线性分类器就是一个非线性函数。

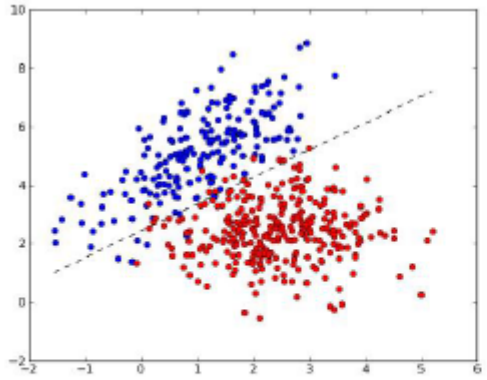


例如，在SVM中的非线性分类器是一个非线性核函数。

（3）Dimensions and Classes

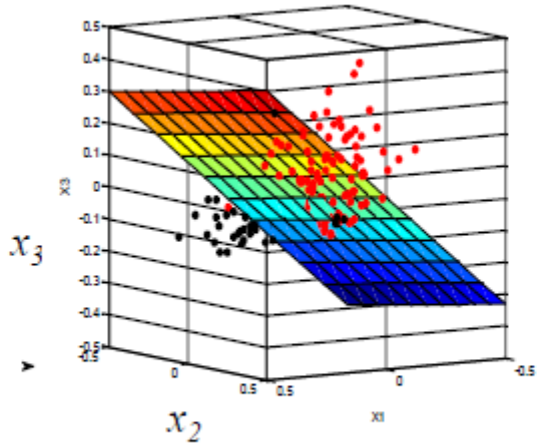
◆ Dimensions 维度

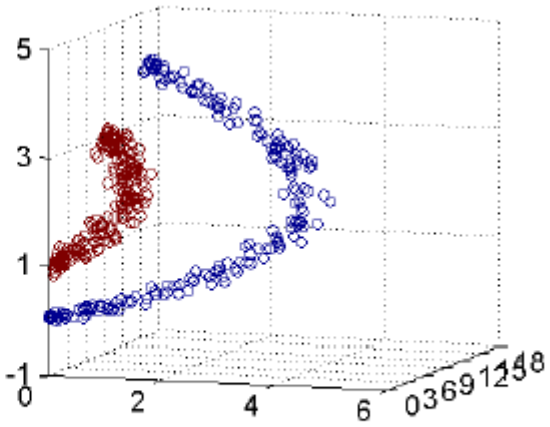
如果问题空间是*n*维的，则它的分类器是维度为 *n*-1的超平面。例如：



2-dimensions
2维

in 2-dimensions, the hyper-plane is a line
2维空间中，该超平面为一条线





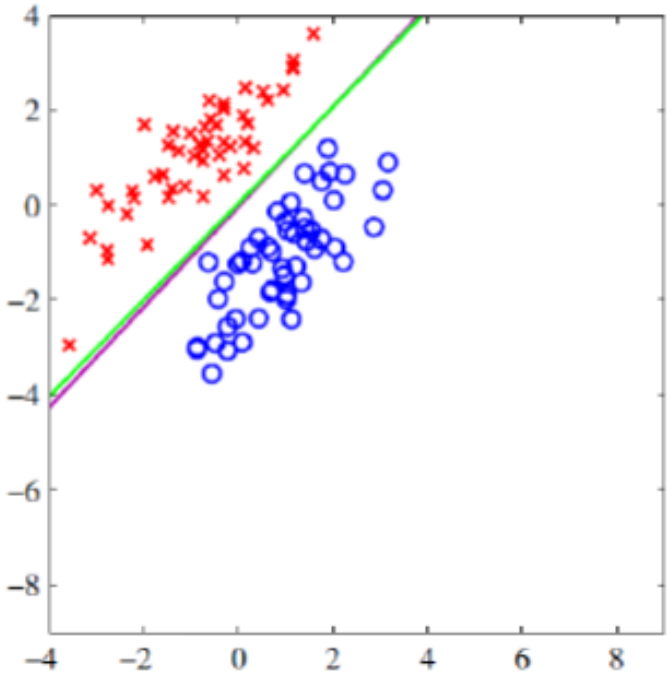
3-dimensions
3维

in 3-dimensions, the hyper-plane is a plane
3维空间中，该超平面为一个2维平面

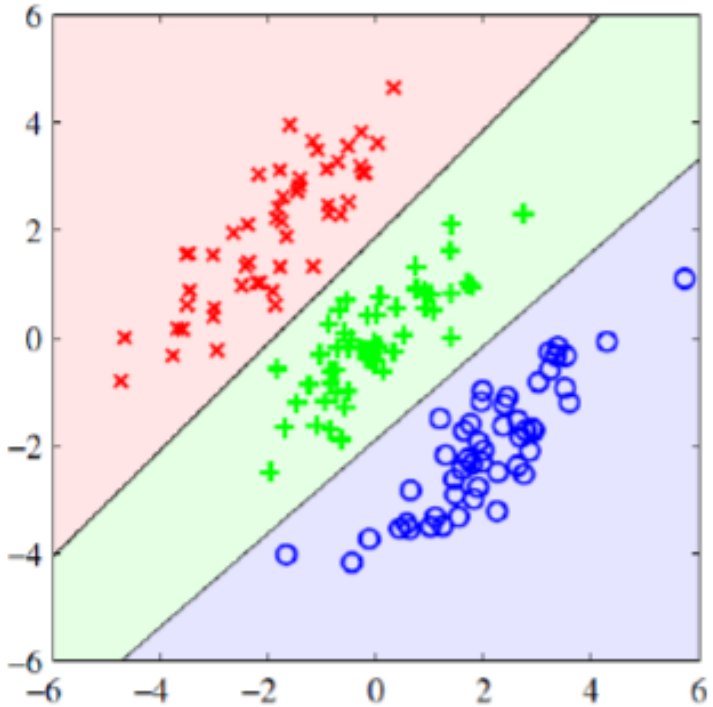
◆Classes 类别

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

- ◆ Two classes: 二元分类：$k = 2$
- ◆ Multiple classes: 多元分类：$k > 2$



Two classes
二元分类



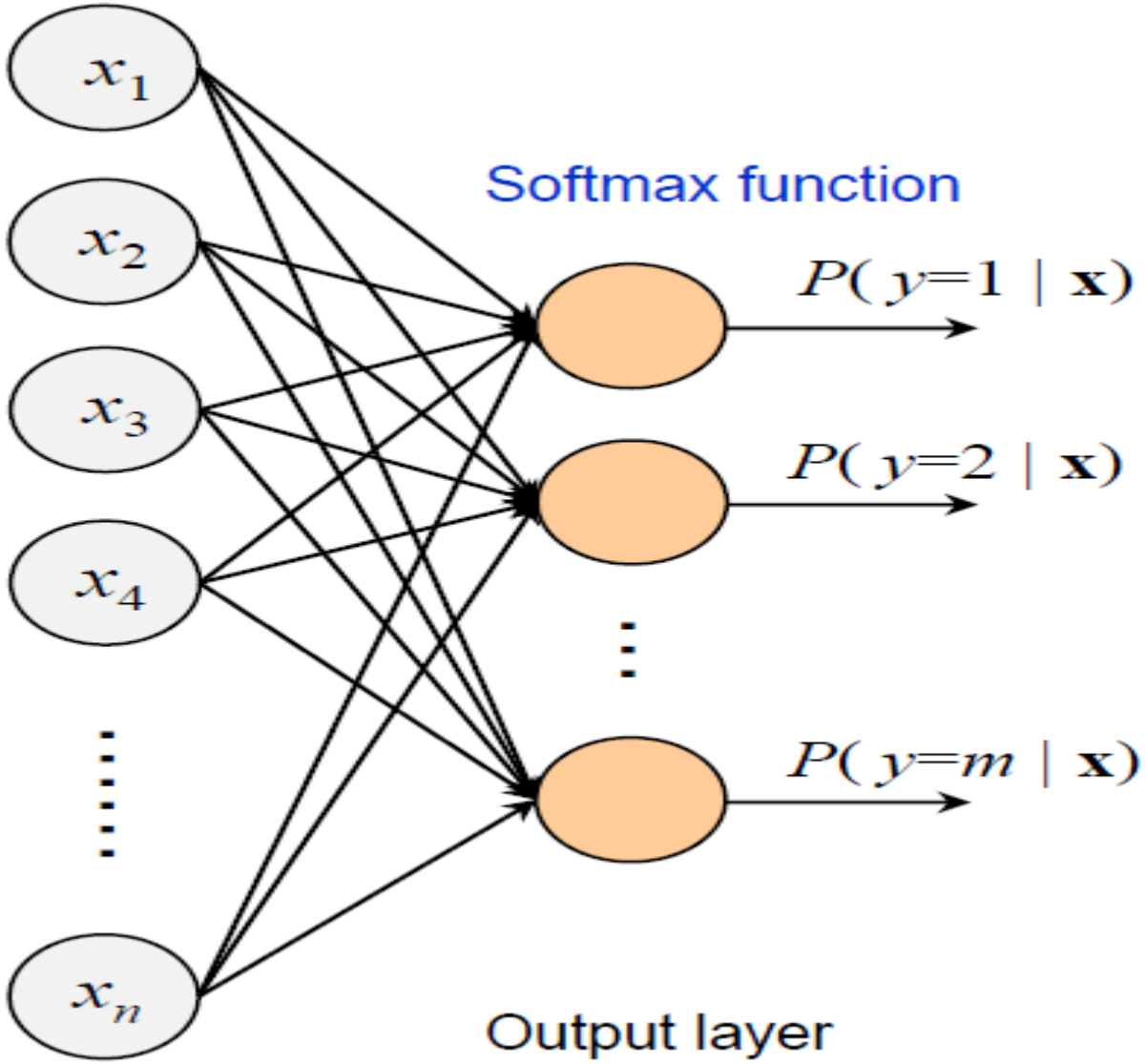
Three classes
三元分类

Case Study: Softmax Classifier （Softmax分类器）

- **Softmax分类器**是一个多元分类器，由 Softmax 函数来实现。
- Softmax 函数，记为$\sigma(\mathbf{x})$，它将一个任意实数值的K 维向量$\mathbf{x}$ 映射到一个实数值的K维向量 (范围0到1，和为1)。

$$\sigma(\mathbf{x})_j = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}} \qquad j = 1, \ldots, K$$

Case Study: Softmax Classifier （Softmax分类器）



Softmax 函数在ANN的最后一层用于多元分类。

(4) 分类的典型应用

- Computer vision
  - Face, handwriting recognition
  - Action recognition
  - Medical image analysis
  - Video tracking
- Pattern recognition
- Biometric identification
- Statistical natural language processing
- Document classification
- Internet search engines
- Credit scoring

计算机视觉

人脸、手写体识别

动作识别

医学图像分析

视频跟踪

模式识别

生物特征识别

统计自然语言处理

文档分类

互联网搜索引擎

信用评分

(4) 分类的典型算法

- **K-nearest neighbors** （KNN）K-近邻
- **Support vector machine** (SVM) 支撑向量机
- AdaBoost
- Decision tree 决策树
- Artificial neural networks 人工神经网络
- Bayesian networks 贝叶斯网络
- Hidden Markov models 隐马可夫模型
- Kernel method 核方法
- Linear discriminant analysis 线性判别分析
- Naive Bayes classifier 朴素贝叶斯分类器
- Softmax

4.2.2 回归

- 什么是回归?
- 线性与非线性回归
- 逻辑回归
- 应用与算法

4.2.2 Regression（回归)

◆ 回归

预测每个项的实数（连续）值。

◆ Regression vs. Classification 回归与分类

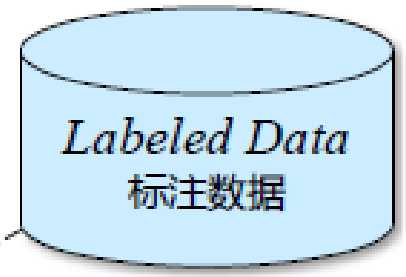
- 相似性：都需要训练过程
- 差异性：如下表所示

	Regression 回归	Classification 分类
Difference 差异性	Output is a real continuous value. 输出是**一个实数连续值。**	Output is a discrete categories. 输出是**一个离散的类别。**
Example 举例	➢ *Used-car price* 二手车价格 ➢ *Tomorrow's stock price* 明天的股票价格	➢ {*sunny, cloudy, rainy*} ➢ {0, 1, 2, ..., 9}

Regression: Training 回归：训练

Known Output Value
已知输出值

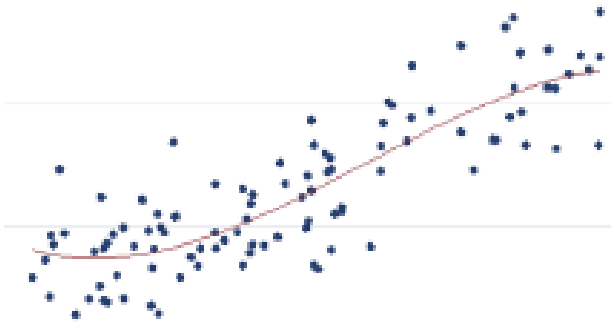
Learning Algorithm
学习算法



Training
训练

(x, y)

$f(X)$



x_1	y_1
x_2	y_2

Labeling function
标注函数

with small generalization and empirical errors
具有小的泛化和经验错误

$h(X)$

Hypothesis
(Regression function)
假设（回归函数）

Regression: Testing 回归：测试

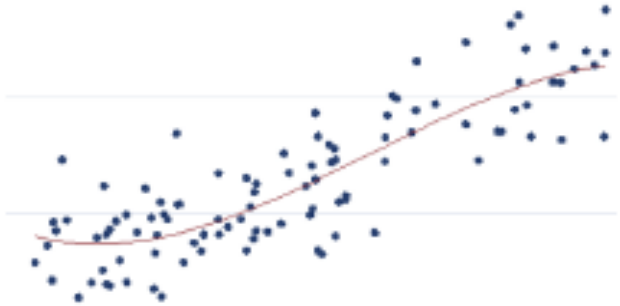
Known Output Value
已知输出值

Learning Algorithm
学习算法

Training
训练

(x, y)

$f(\mathcal{X})$



Labeled Data
标注数据

x_1	y_1
x_2	y_2

with small generalization and empirical errors
具有小的泛化和经验错误

Unseen Data
未知数据

x_1	?
x_2	?

$\mathcal{X}$

$h(\mathcal{X})$

$\mathcal{Y}$

Regressed Data
回归后的数据

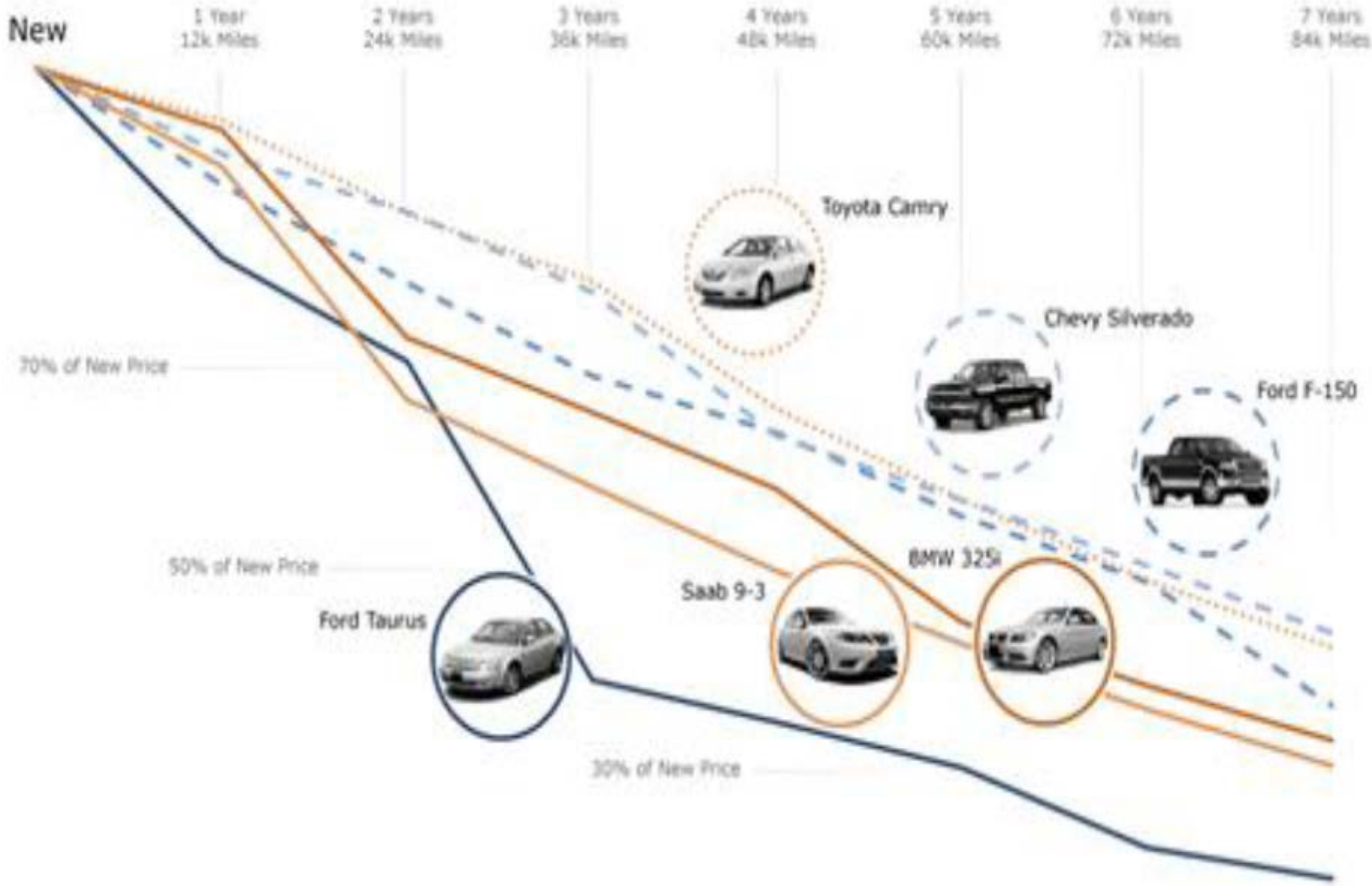
x_1	y_1
x_2	y_2

Regression
回归

Unknown Output Value
未知输出值

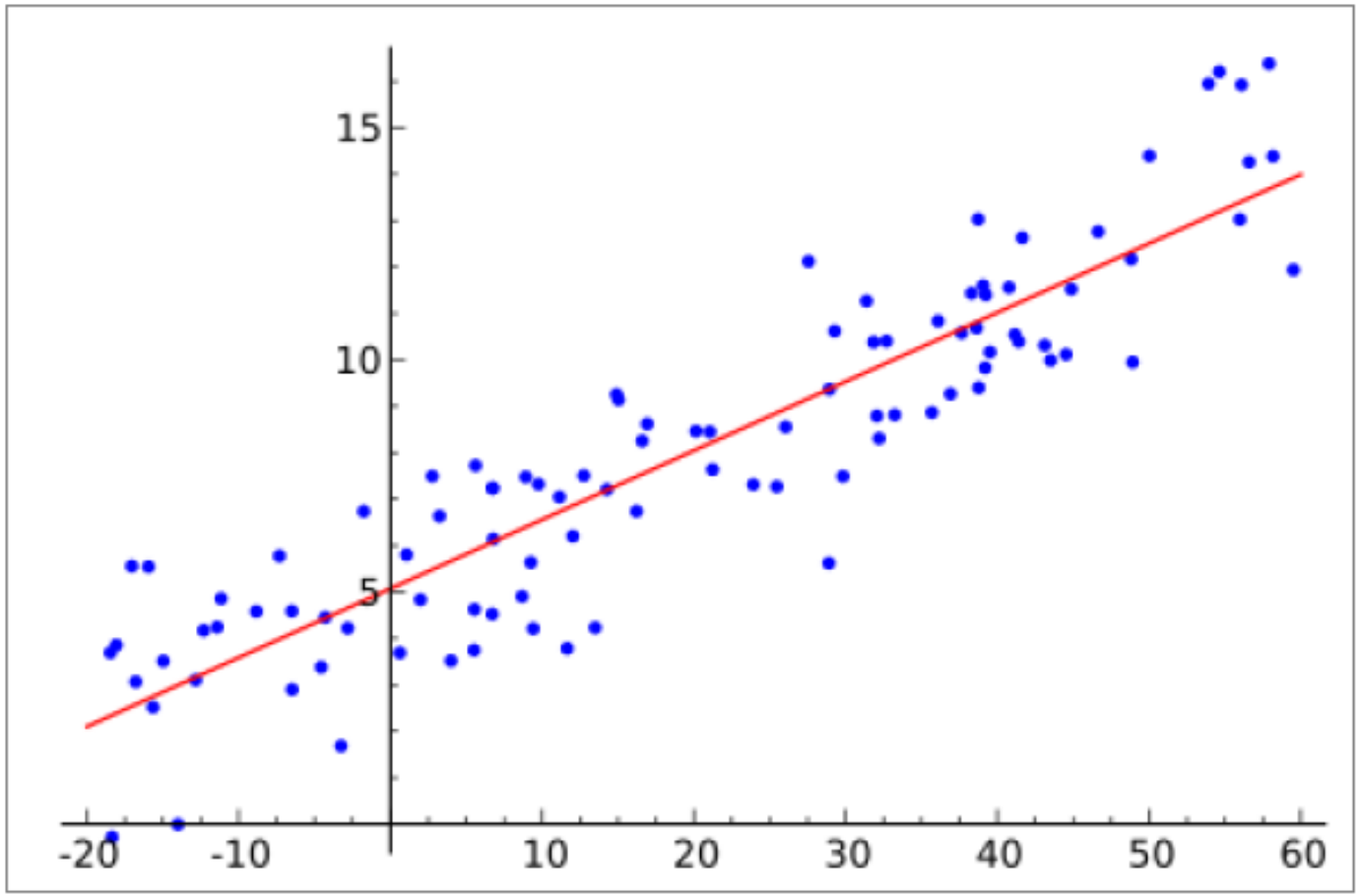
Example: Used Car Prices 二手车价格

- 构建一个预测二手车价格的系统。
- 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**。



Linear Regression 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；
  - 该函数取决于一个或多个独立变量。

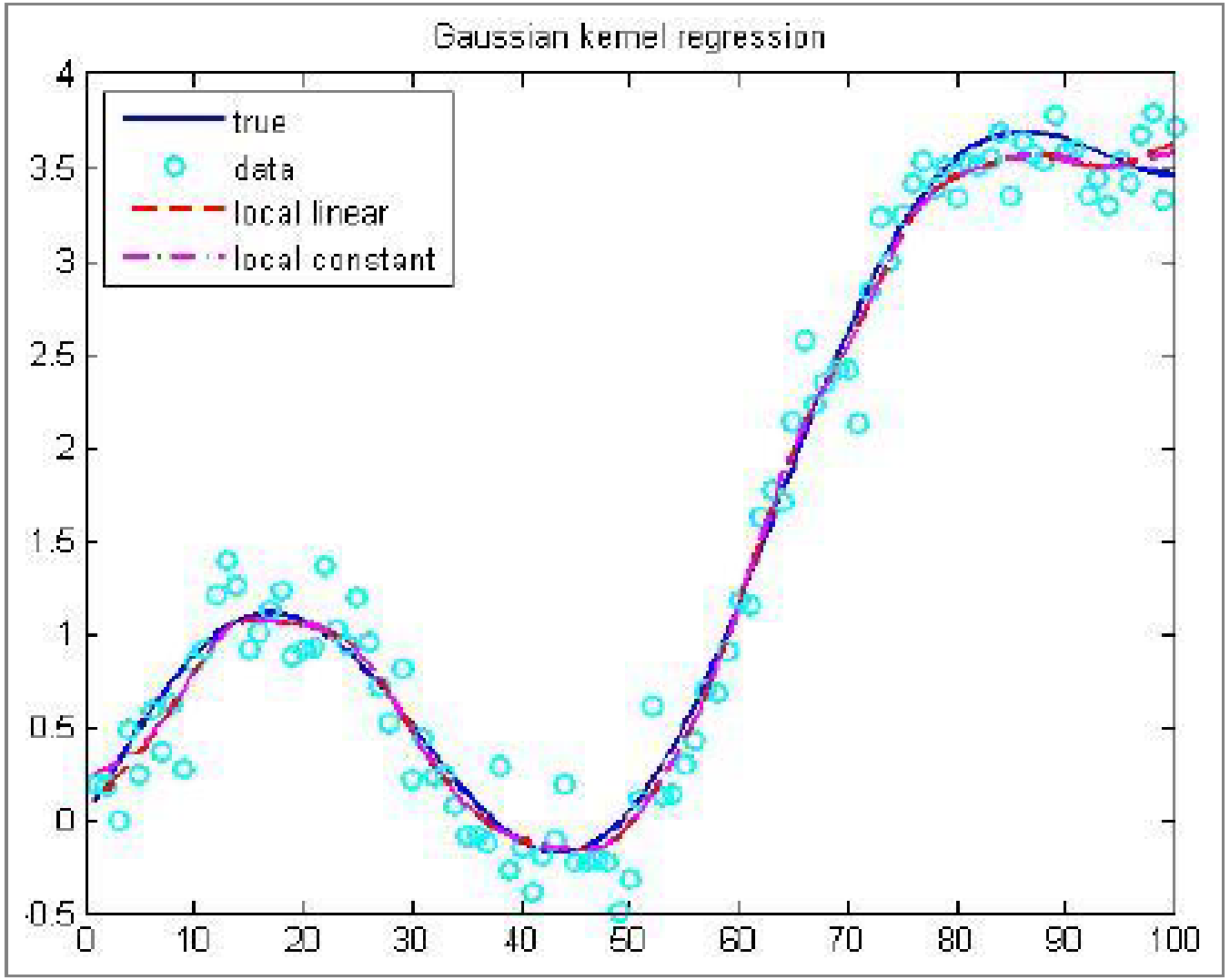


$$y(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b$$

模型表达：$y(x,w)=w_1x_1+\ldots+w_nx_n+b$

Nonlinear Regression 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。



$$y(\mathbf{x}) = \mathbf{w}_2 \cdot \mathbf{x}^2 + \mathbf{w}_1 \cdot \mathbf{x} + b$$

Logistic Regression 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。
- 例如，将用户的特征属性（性别，年龄，注册时间等）设置为**自变量**，将购买的概率（买/不买）设置为**因变量**。根据特征属性预测购买的概率。
- **线性回归**用于**预测**连续值，**逻辑回归**主要用于解决**分类**问题。
- 逻辑回归的**自变量**可以有一个，也可以有多个。一个自变量的叫做一元回归分析，超过一个自变量的叫做多元回归分析。

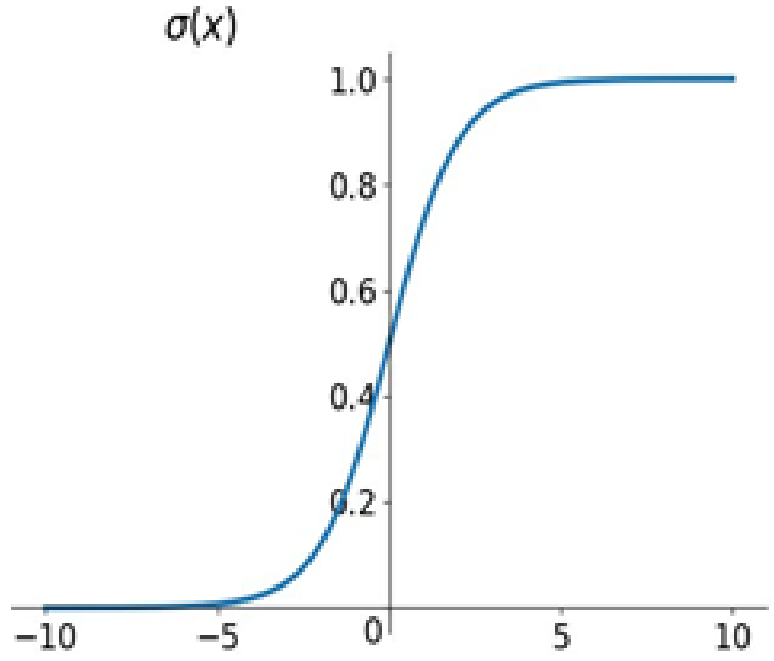
  logistic回归的**因变量**可以是**二分类**（binary classification），也可以是**多分类**。**二分类更为常用**，也更容易解释。
- **使用sigmoid函数，就是二分类**，若阈值取到0.5，也就是说大于0.5的是一类；小于0.5的是另一类。
- **使用softmax就是多分类**。

Logistic Regression 逻辑回归

二分类：使用一个Sigmoid 函数g，将回归函数值y(x)映射到 (0,1)间，适合解决**分类问题**。

Sigmoid function/ Logistic function :

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



输出: 实数值 $y \in (0,1)$--- 概率

含义: y表示未知样本 x 属于某一类别的概率.

Sigmoid函数特性

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

（1）定义域为R

（2）值域为（0,1）

（3）函数在定义域内为连续和光滑的函数

（4）处处可导，且导函数满足 g'(x) = g(x)(1-g(x))

（5）S形函数，中间梯度大，越趋近于两端梯度越小，容易梯度消失。

（6）导数值域为(0,0.25]

优点：平滑、易于求导。

缺点：激活函数计算量大，反向传播求误差梯度时，求导涉及除法；
反向传播时，很容易就会出现梯度消失的情况，从而无法完成深层网络的训练。

逻辑回归 vs. 线性回归

method	Independent variable 自变量（特征）	dependent variable 因变量（结果）	Function Type	usage
Linear Regression	**Continuous/ discrete values**	**Continuous real values**	**Linear**	**House/ used car price**
Logistic Regression	**Continuous/ discrete values**	**[0,1] Continuous real values**	**Non-Linear**	Tumor: Malignant / Benign

逻辑回归： 告诉病人肿瘤为恶性的概率为70%

回归的典型应用

回归被广泛地用于**预测**和**预报**。

- Trend estimation 趋势估计
- Epidemiology 传染病学
- Finance 金融：分析与量化投资的系统性风险。
- Economics 经济

  预测消费支出、固定资产投资支出、持有流动资产需求、等等。
- Environmental science 环境科学

Typical Algorithms of Regression 回归的典型算法

- Bayesian linear regression 贝叶斯线性回归
- Percentage regression 百分比回归
- Kernel ridge regression, 核岭回归
- Support-vector regression, 支撑向量回归
- Quantile regression, 分位数回归
- Regression Trees, 回归树
- Cascade Correlation, 级联相关
- Group Method Data Handling (GMDH), 分组方法数据处理
- Multivariate Adaptive Regression Splines (MARS), 多元自适应回归样条
- Multilinear Interpolation 多线性插值

4.2.3 Clustering 聚类

- 什么是聚类?
- 聚类的主要方法
- 应用与算法

4.2.3 聚类

◆ **聚类**是**将对象进行分组**的任务，使得同一组中的对象彼此之间比其他组中的对象更相似。即，聚类是**将相似的输入数据分在同一类别。**

Clustering vs. Classification 聚类与分类

◆ **相似性: 都是分组**

◆ **差异性:** 如下表所示

Clustering 聚类	Classification 分类
To identify similar groups for input objects 给输入对象**标识相似的组。**	To assign pre-defined classes for input items 给输入项**分派预定义的类。**
Without training data. **没有训练数据。**	With training data. **有训练数据。**
Clusters are discovered based on distances, density, etc. 基于距离、密度等发现类**别。**	Classifiers need to have a high accuracy for classification. 分类器需要具有较高的分类精度。

聚类算法的典型方法

1)　层次聚类

基于对象间距离的层次聚类。

2) **基于中心点聚类**

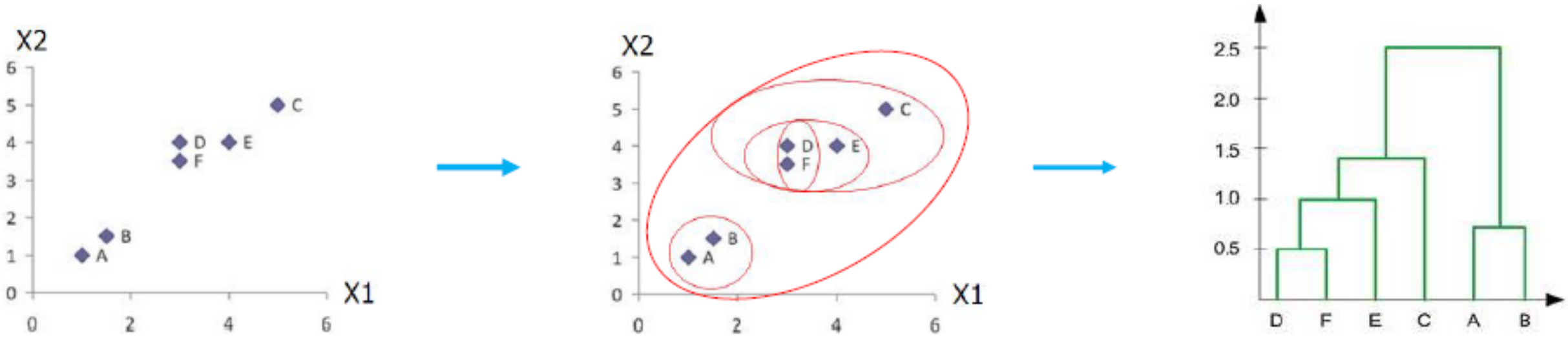
找到 k个类别中心，并将对象分配到最近的类别中心点，也称为**划分聚类**。

3) **基于密度聚类**

将稠密区域连接的对象组成一个类别。

1) 层次聚类

- 基于这样一个核心理念：对象与其附近的对象更相关，而不是较远的对象。
- 采用某种准则来创建数据对象集的层次分解。

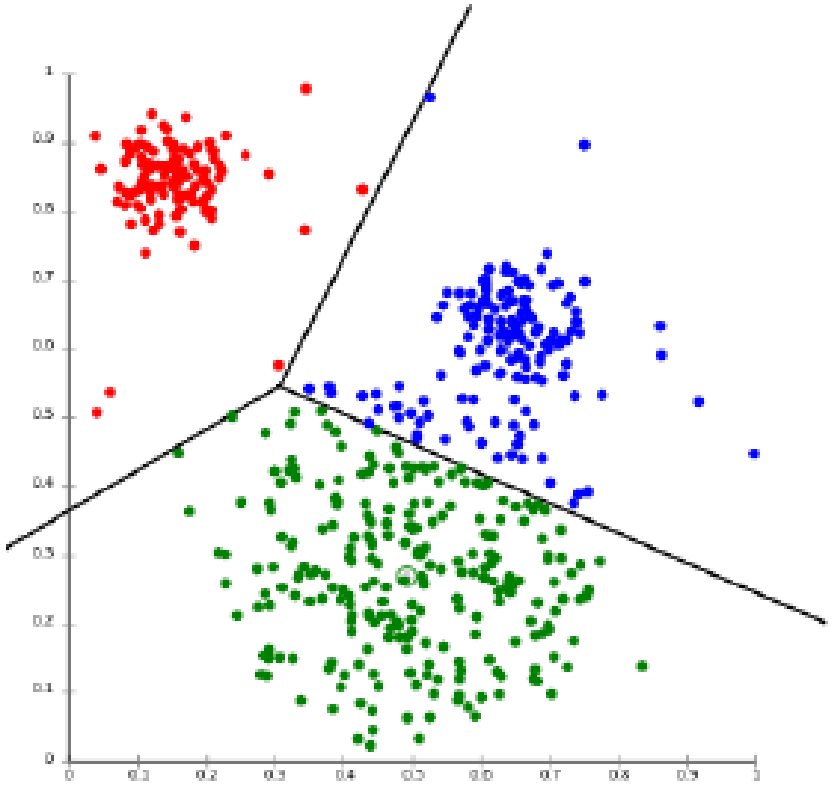


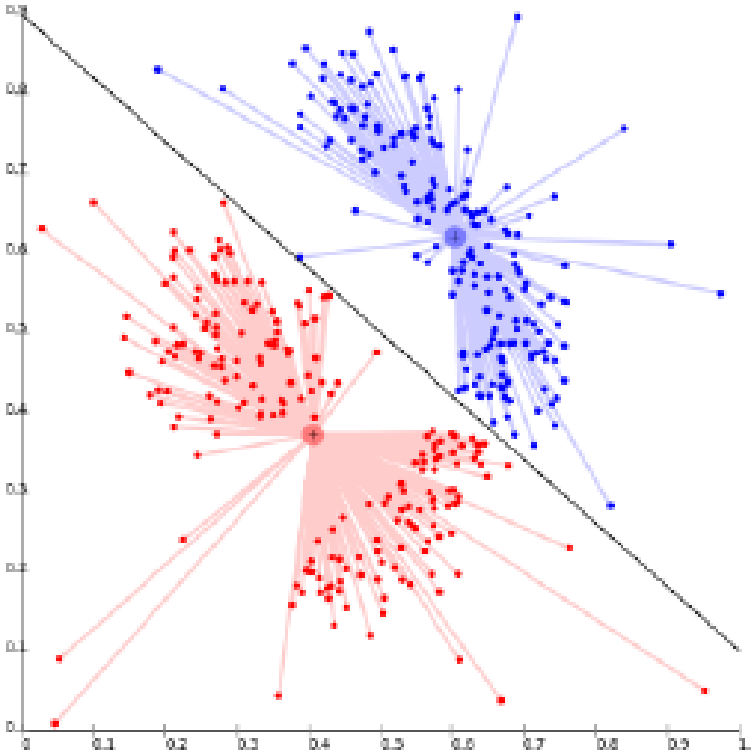
Typical algorithms: AGNES (Agglomerative NESting), DIANA (Divisive Analysis), ……

典型算法：AGNES (集聚嵌套), DIANA (分裂分析), ...

2) 基于中心点聚类/基于划分的聚类

构建各种不同的分区，再根据某种准则（例如最小平方距离代价之和）对其进行评价。



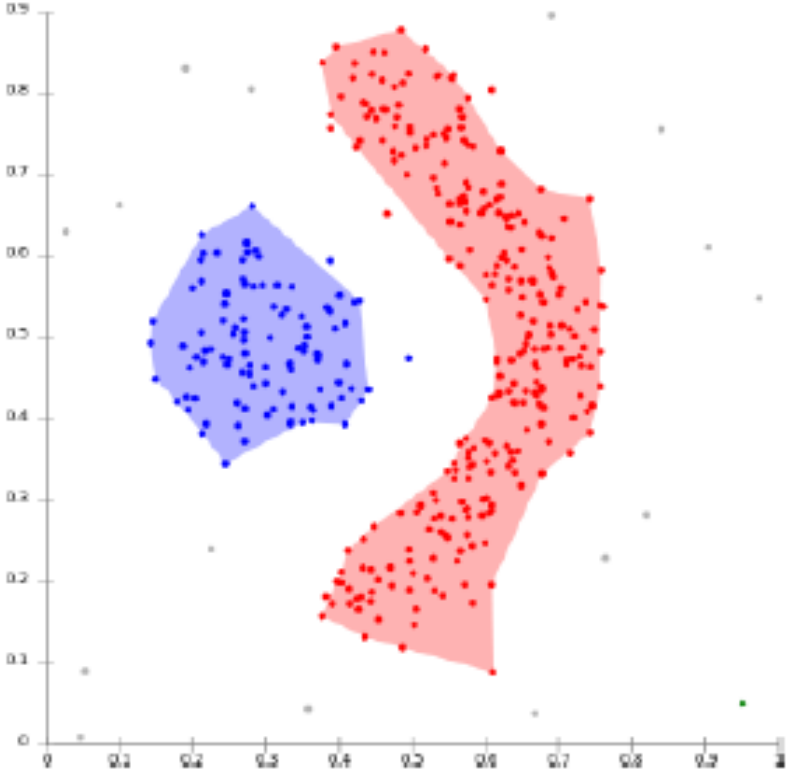


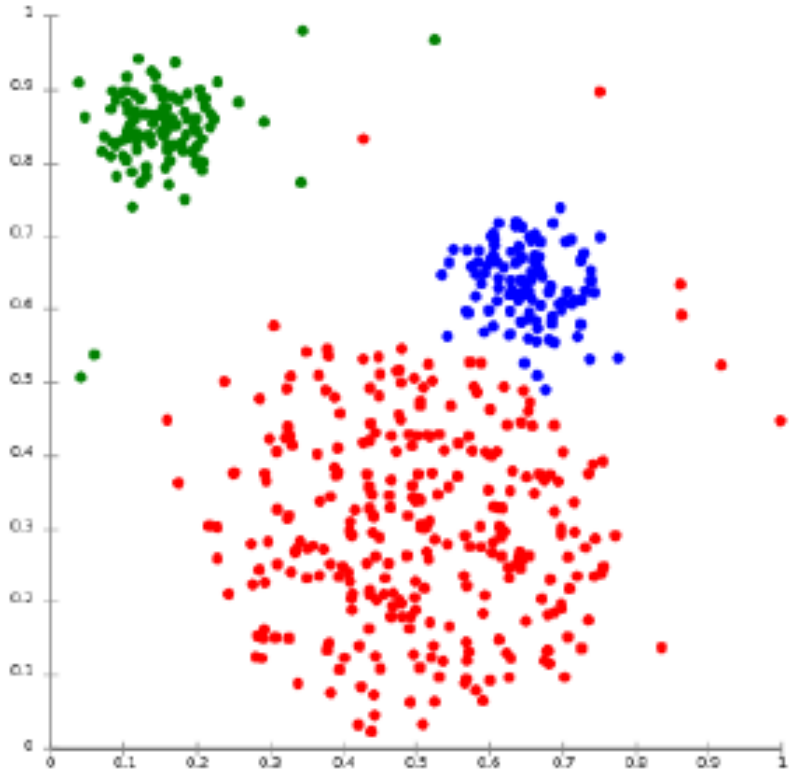
Typical algorithms: *k*-means, *k*-medoids, ……

典型算法：*k*-均值, *k*-中心点, ...

3) 基于密度聚类

类别被定义为比数据集其余部分密度更高的区域。





Typical algorithms: DBSCAN (Density-Based Spatial Clustering of Applications with Noise), ……

典型算法：DBSCAN (基于密度的噪声应用空间聚类), ...

Case Study: Clustering by density peaks

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）。

聚类的典型应用

- Medicine
  - Medical imaging
- Business and marketing
  - Grouping of customers
  - Grouping of shopping items
- World wide web
  - Social network analysis
  - Search result grouping
- Computer science
  - Image segmentation
  - Recommender systems

医学

医学影像

商务和营销

顾客分组

购物商品分组

万维网

社交网络分析

搜索结果分组

计算机科学

图像分割

推荐系统

典型的聚类算法

- ***k*-means（k均值）**
- *k*-modes（k众数）
- PAM
- CLARA
- FCM
- BIRCH
- CURE
- ROCK
- Chameleon
- Echidna
- DBSCAN
- DBCLASD
- OPTICS
- DENCLUE
- Wave-Cluster
- CLIQUE
- STING
- OptiGrid
- EM
- CLASSIT
- COBWEB
- SOMs

4.3 机器学习的范式

4.3.1 监督学习

4.3.2 无监督学习

4.3.3 强化学习

机器学习中的典型学习范式

Paradigms 范式	Brief Statements 简短描述	Typical Algorithm 典型算法
Supervised 有监督	The algorithm is trained by a set of labeled data, and makes predictions for all unseen points. 算法采用一组标注数据进行训练，再对所有的未知点做出预测。	Support vector machines 支撑向量机
Unsupervised 无监督	The algorithm exclusively receives unlabeled data, and makes predictions for all unseen points. 算法仅接收未标注的数据，再对所有的未知点做出预测。	*k*-means *k*-均值
Reinforcement 强化	The algorithm interacts with environment, and receives an reward for each action. 算法与外部环境交互，每个动作得到一个回报。	Q-learning

4.3.1 有监督学习

什么是有监督学习

- 智能体接收一组**标注**的对象作为训练数据，然后对所有的未知点进行推测。
- 这种方式试图**生成**从输入到输出的**函数**或映射，然后可以将其用于对预先未知的数据生成输出。

It is a way of "teaching" the learning algorithm, like that a "teacher" gives the classes (courses).

这是一种“教”学习算法的方式，就像“老师”讲授课程那样。

◆ 有监督学习中的训练数据：

- 每个训练数据具有一个**已知标注**作为输入数据，
- 标注是由输入对象和预期输出值组成的**对**，例如垃圾与非垃圾邮件、或猫与狗

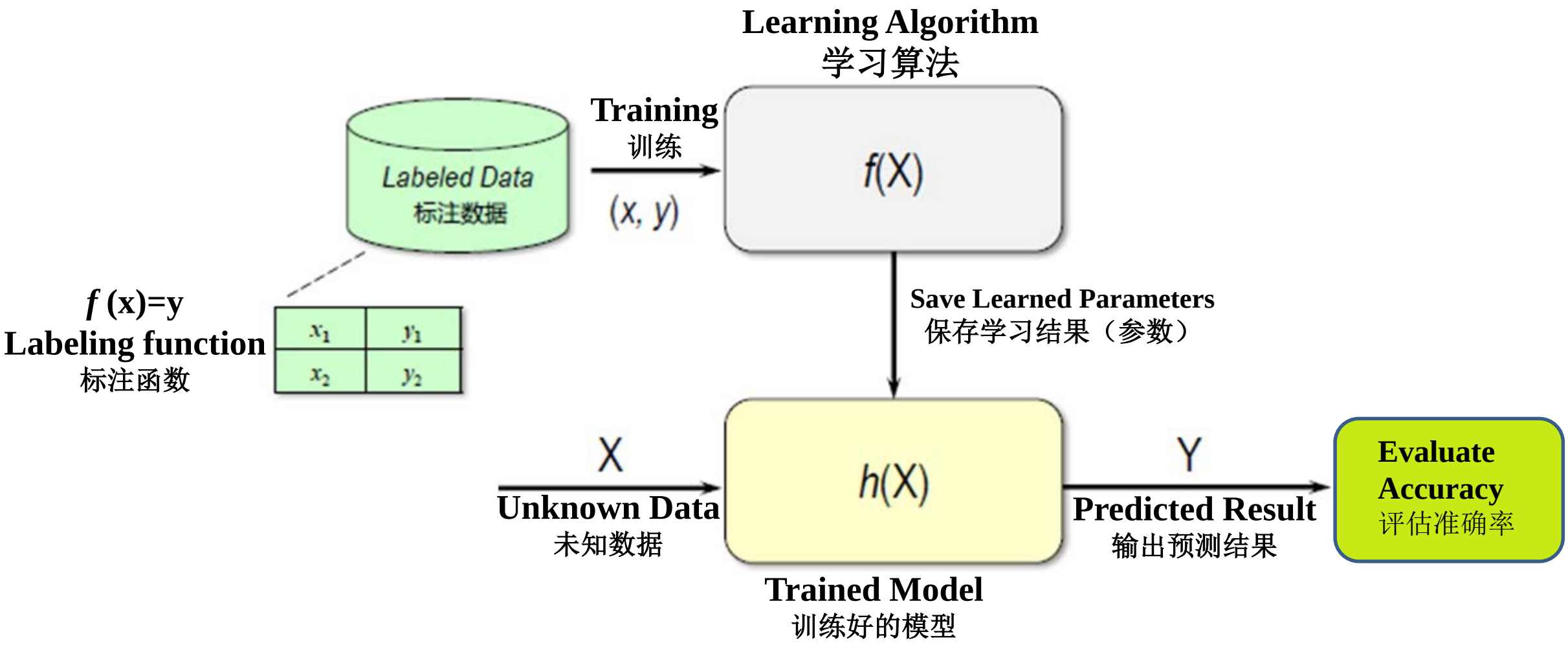
◆ 训练后的假设函数（模型）：

可用于映射新的未知数据。

desired output： **cat**

desired output： **dog**

有监督学习的步骤


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
x1 y1
x2 y2
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

有监督学习的步骤

1) 收集**训练集**和**测试集**

训练集和**测试集是两个不相交的数据集。**

例如，对于手写体识别，可以是手写字符、手写单词等。

2) 确定特征提取方法

通常，有两种从输入数据提取特征的方法：

- **手工特征提取**：通过某种特征描述子。
  - SIFT (Scale-invariant feature transform, 尺度不变特征变换)
  - HOG (Histogram of Oriented Gradient, 方向梯度直方图)
- **自动特征提取**：通过某种深度神经网络。

有监督学习的步骤

3) 选择完成该任务的学习算法，这取决于你的任务是什么。

例如，对于**分类**来说，你可以选择使用SVM、决策树、KNN、等等。

聚类：k-means

4) 采用该学习算法训练模型

在收集的训练数据集上运行该学习算法。

某些算法需要用户来确定某些**控制参数（超参）**。

这些**参数**可以通过在训练子集上优化性能来调整。

5) 评估模型的精确性

在**参数**调整和学习之后，应当在（独立于训练集的）**测试集**上对模型（函数）的性能进行度量。

与有监督学习相关的任务

◆ 分类

输出空间Y是一组**类别**。离散的类别，如：二分类（垃圾、非垃圾）、多分类（0---9）

◆回归

输出空间Y是一组**连续的实数值**。如房屋价格预测：建筑年代、面积大小、地理位置、学区房 School District ）

◆排名

输出空间Y是一组**相对的顺序**。例如：对搜索输出结果的排名

有监督学习的一些应用

Object recognition in computer vision	◆ 计算机视觉中的物体识别
Optical character recognition (OCR)	◆ 光学字符识别(OCR)
Handwriting recognition	◆ 手写体识别
Information retrieval	◆ 息检索
Learning to rank	◆ 学会排名
Spam detection	◆ 垃圾邮件检测
Speech recognition	◆ 语音识别
Bioinformatics	◆ 生物信息学
Cheminformatics	◆ 化学信息学

几个有监督学习的例子

- **垃圾邮件检测（二分类问题）**

  将电子邮件分为{Spam, Not Spam}

- **数字识别（多分类问题）**

  将手写体数字映射为{0, 1, 2, 3, 4, 5, 6, 7, 8, 9}

- **二手车/房屋价格预测（线性回归）**

  根据二手车市场收集到的历史数据，估算一台二手车的实际价格。

典型的分类与回归算法

Algorithm 算法	Task Types 任务类型	Predictive accuracy 预测精度	Training speed 训练速度
AdaBoost 自适应增强	Either 两者	Higher 高	Slow 慢
Artificial neural network 人工神经网络	Either 两者	Higher 高	Slow 慢
k-Nearest neighbor *k*近邻	Either 两者	Lower 低	Fast 快
Linear regression 线性回归	Regression 回归	Lower 低	Fast 快
Logistic regression 逻辑回归	Classification 分类	Lower 低	Fast 快
Naive Bayes 朴素贝叶斯	Classification 分类	Lower 低	Fast 快
Decision tree 决策树	Either 两者	Lower 低	Fast 快
Random Forests 随机森林	Either 两者	Higher 高	Slow 慢
Support vector machines 支撑向量机	Either 两者	Higher 高	Slow 慢

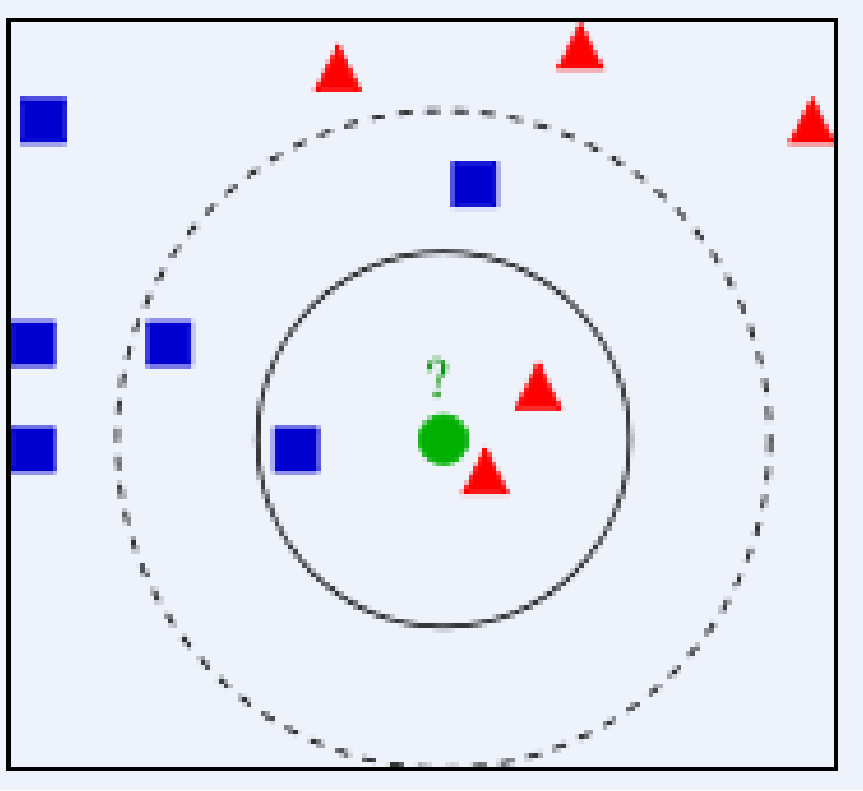
K-nearest Neighbor （KNN） Algorithm

- KNN算法是典型的监督学习算法，可以用于**分类**，还可以用于**回归**。
- KNN算法的思路：
  - **分类：** 给定待测样本A，在特征空间中找出与样本A最相似（即在特征空间中最邻近）的k个样本，然后统计这k个样本属于各类别的样本数，找到样本数最多的类别，则样本A也归于该类，属于分类问题。
  - **回归：**找出一个样本A在特征空间中的k个最相似样本，将这**k个样本**属性的**平均值**赋给样本A，就可以得到样本A的属性，属于回归问题。

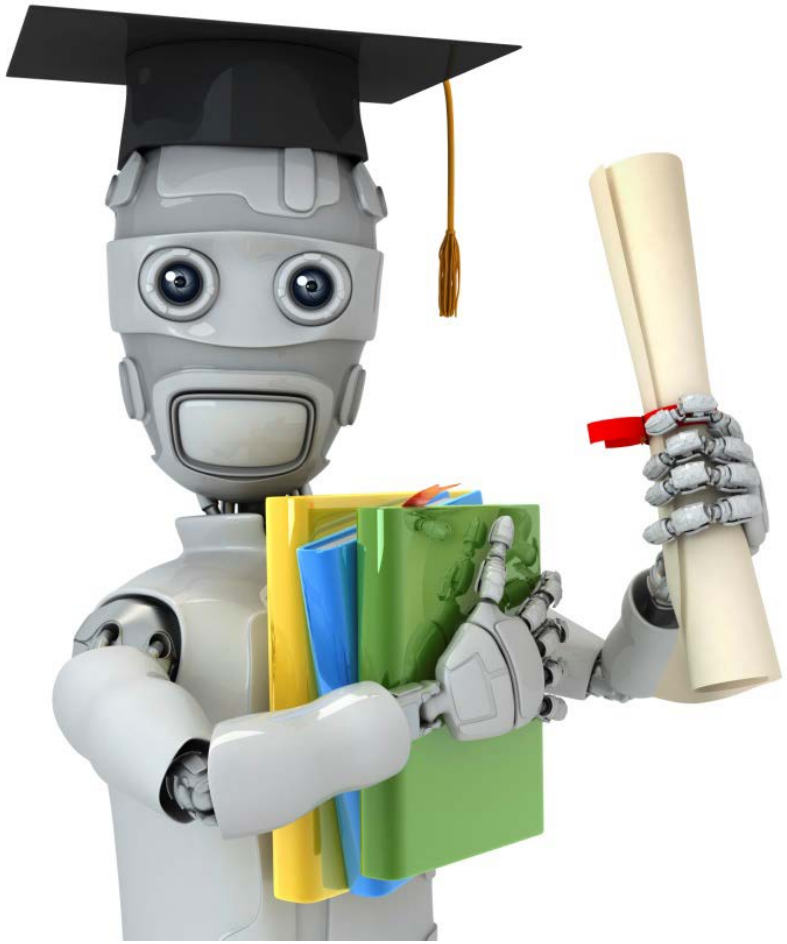
K-nearest Neighbor （KNN） Algorithm

KNN algorithm process:

1. 计算待测样本A到训练集中每个样本的距离
2. 将所有样本按照与A的距离的**增序**排列
3. 选取与A最近的k个训练样本，即k个最近的邻居
4. 统计这k个邻居的类别频率
5. 找到k个邻居中频率最高的类别，作为测试样本的类别。



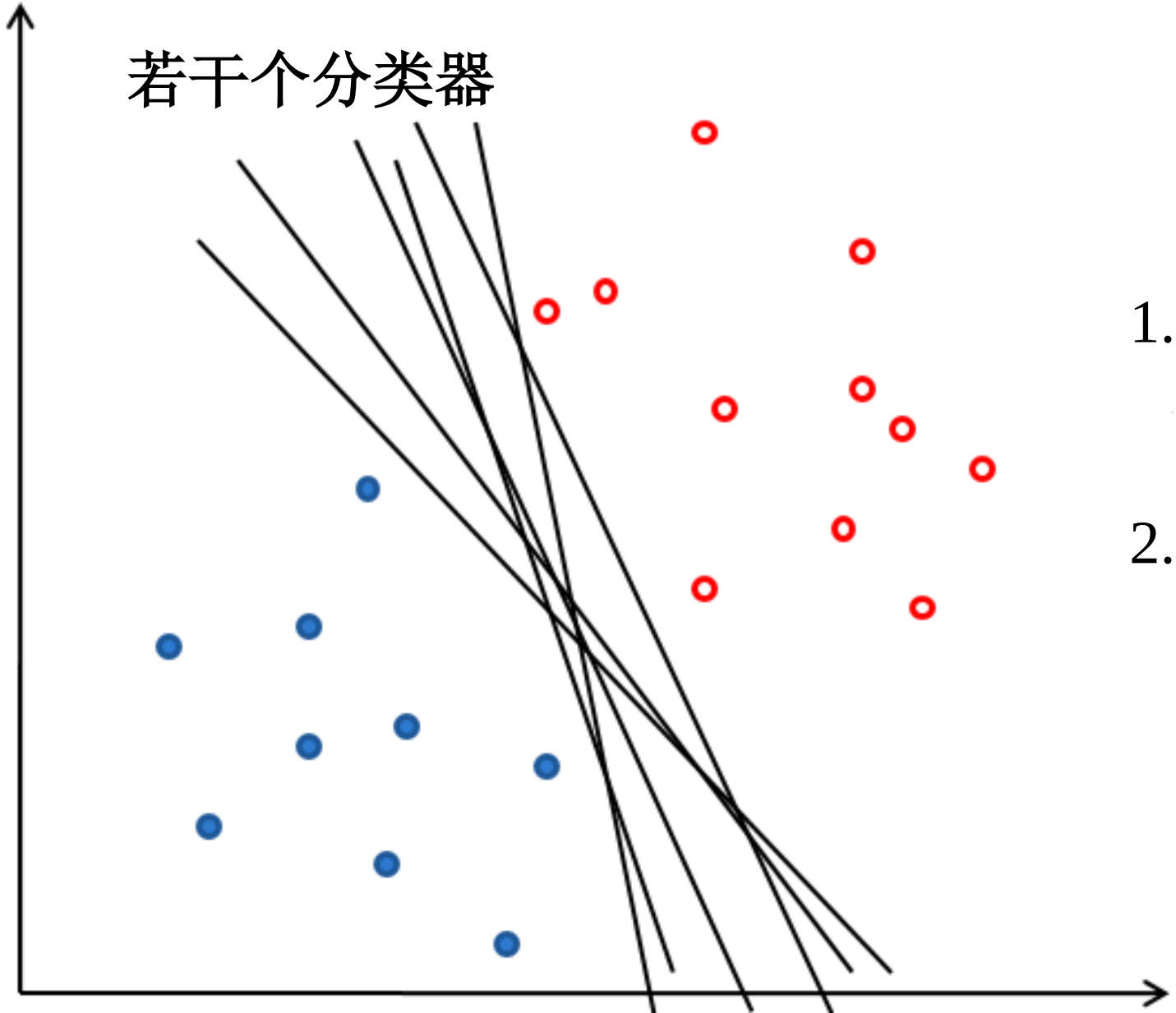
问题：如何确定 *k* 值？只是个经验值



Machine Learning

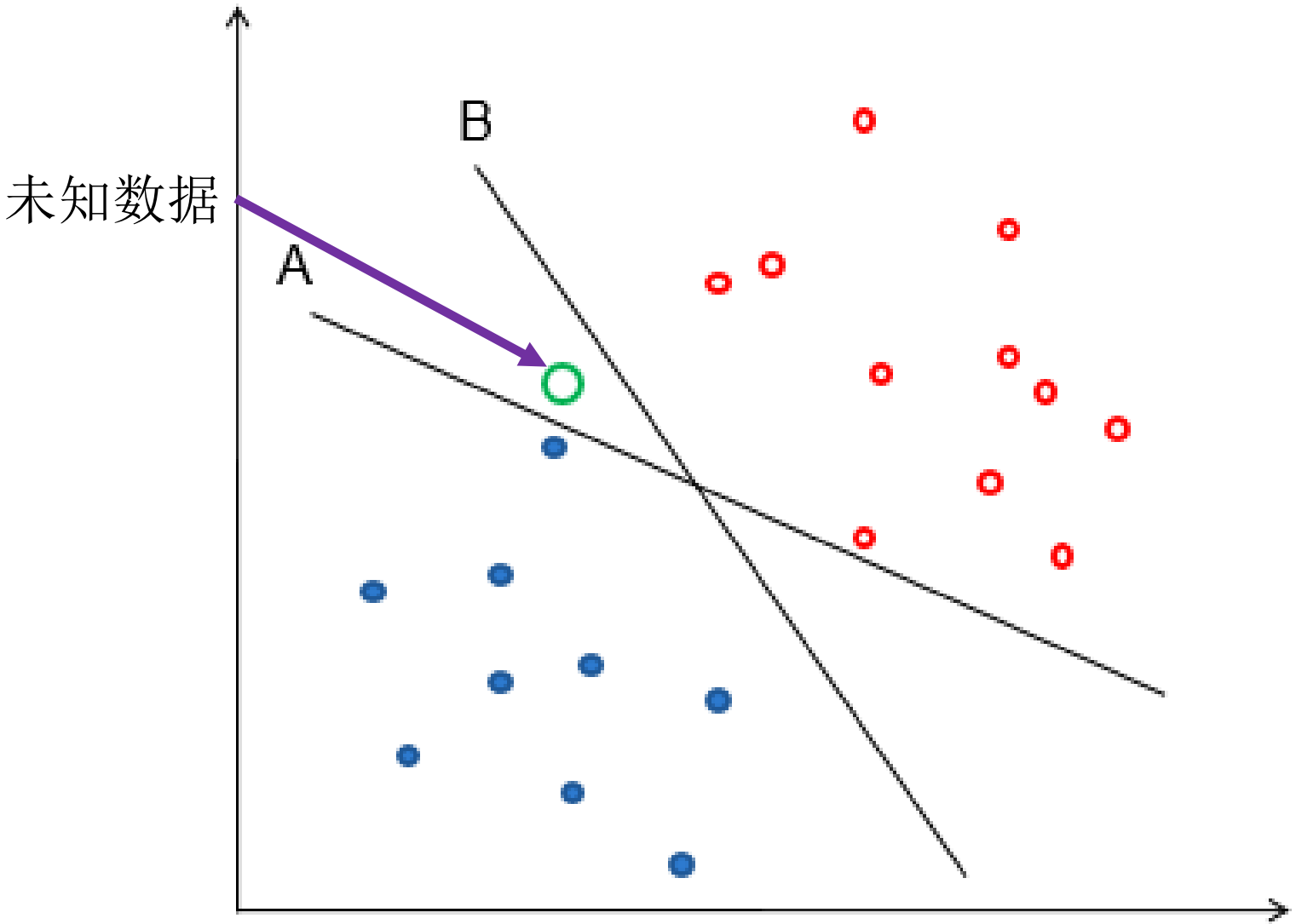
Support Vector Machines （SVM）

分类器的选择



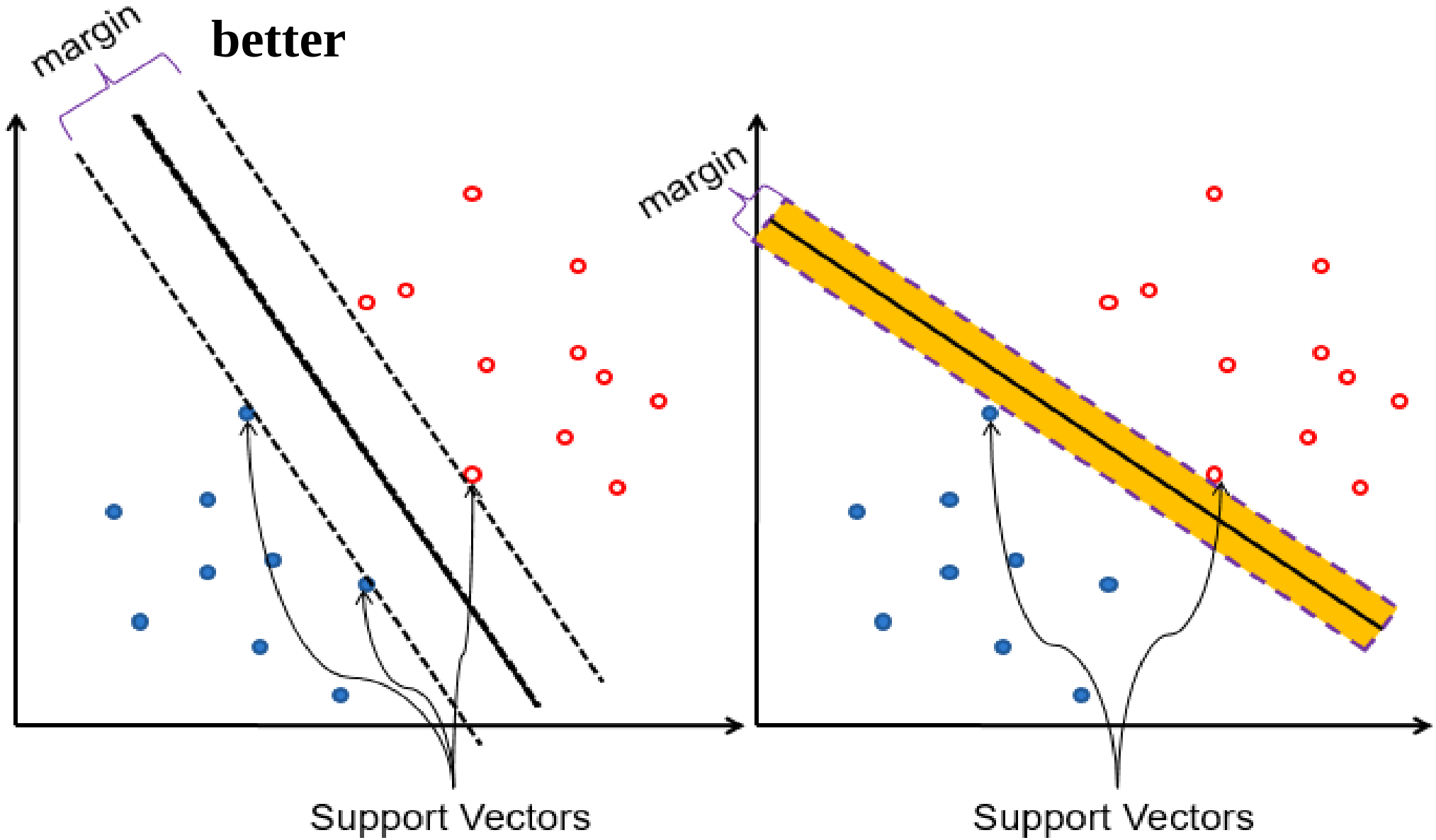
1. 若所有的分类器的错误率系统，哪个最好?
2. 如何泛化?

未知对象



分类器B 更好些，因为它能将未知数据划分正确，这就是泛化能力。

间隔（Margins）



Margins 间隔

- 线性分类器的**间隔**定义为：在碰到数据点之前边界的宽度。
- 直观上，**间隔越大大越好**

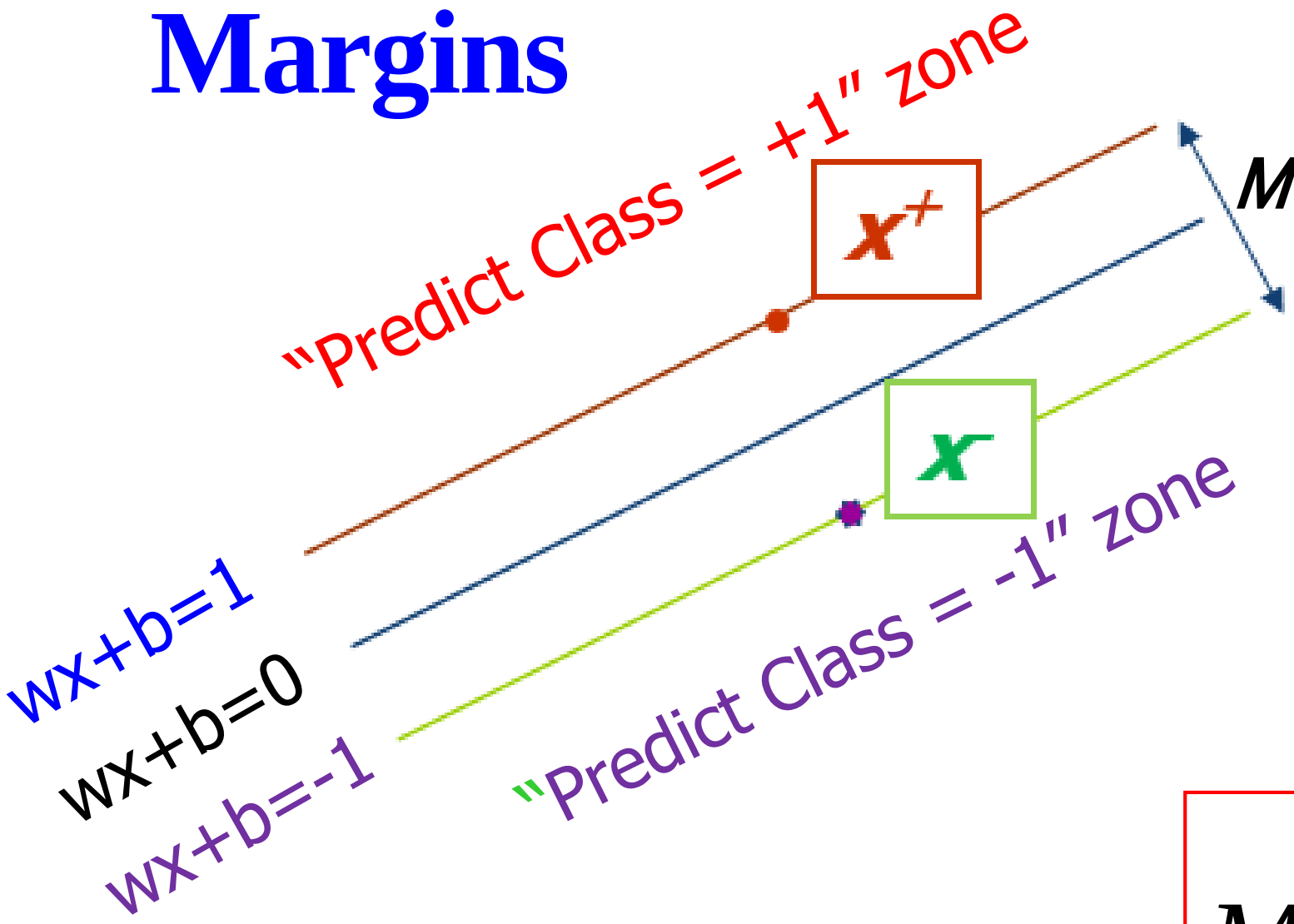
  超平面只由几个数据点确定：

  – Support Vectors（这几个数据点称为**支持向量**）

  – 其余的数据可被忽略
- 选择间隔最大的分类器。

  – Linear Support Vector Machines (LSVM)
- 怎么正式定义“间隔”？

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

Distance between two parallel lines

两条平行线：$Ax+By+c1=0$

$Ax+By+c2=0$

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

Two parallel lines：$wx+b-1=0$ **and** $wx+b+1=0$

Distance between two lines（**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

4.3.2. 无监督学习

◆接收**未标注数据**，并对所有的未知点做出预测。

◆其目标是发现数据中共性的东西，或者减少正在考虑的随机变量的数量。

◆无监督学习则没有训练过程，给定一些对象数据，让机器学习算法直接对这些数据进行分析，得到数据的某些知识。

◆ 例如，抓取了1万个网页，要对这些网页归类，保证同一类网页有相同主题，不同类的网页，主题不同。

◆ 采用聚类算法：事先并未定义好类别，也没有已训练好的分类模型。

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

这是一种“自学”的方式，没有“老师”。

有监督学习与无监督学习

Supervised learning 有监督学习

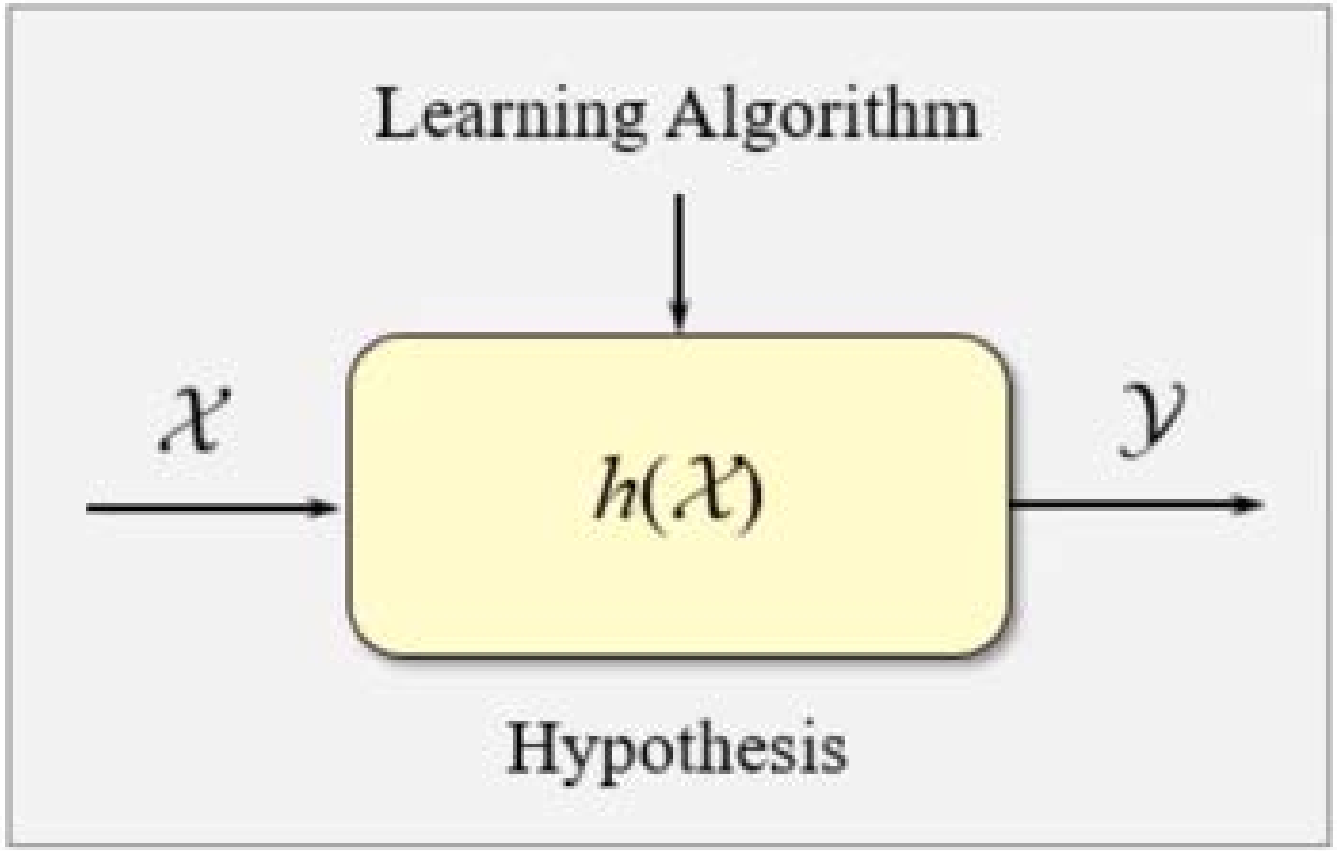
- 给予学习器的对象是**已标注的**
- 对象用于**训练该算法**。

Unsupervised learning 无监督学习

- 给予学习器的对象是**未标注的**
- **没有训练过程**。

与无监督学习相关的任务

- **Clustering** 聚类
- Density estimation 密度估计
- Dimensionality reduction 降维



聚类分析

◆找到对象的分组

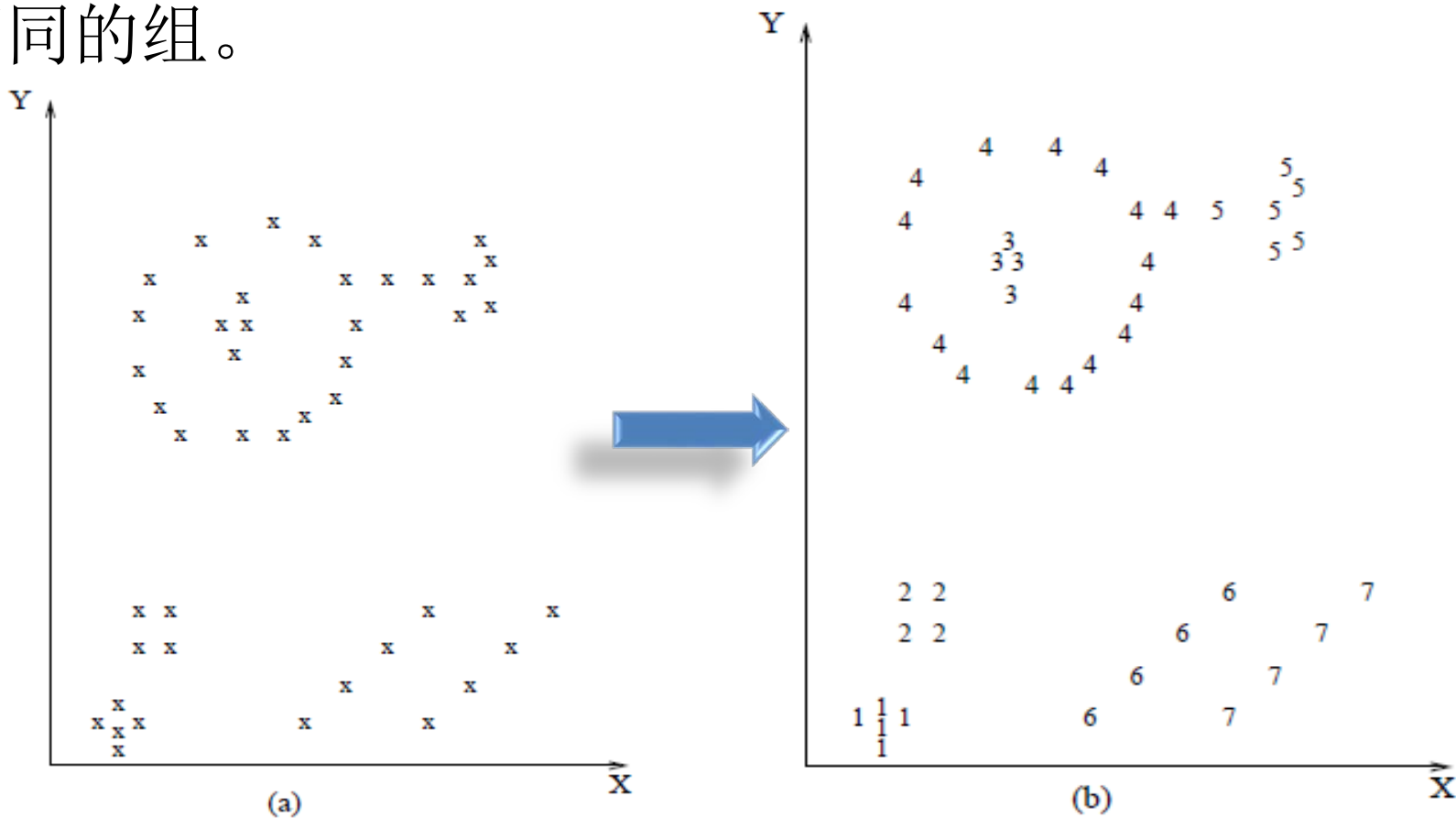
– 相似的对象被分在同一组。

– 不同的对象被分在不同的组。

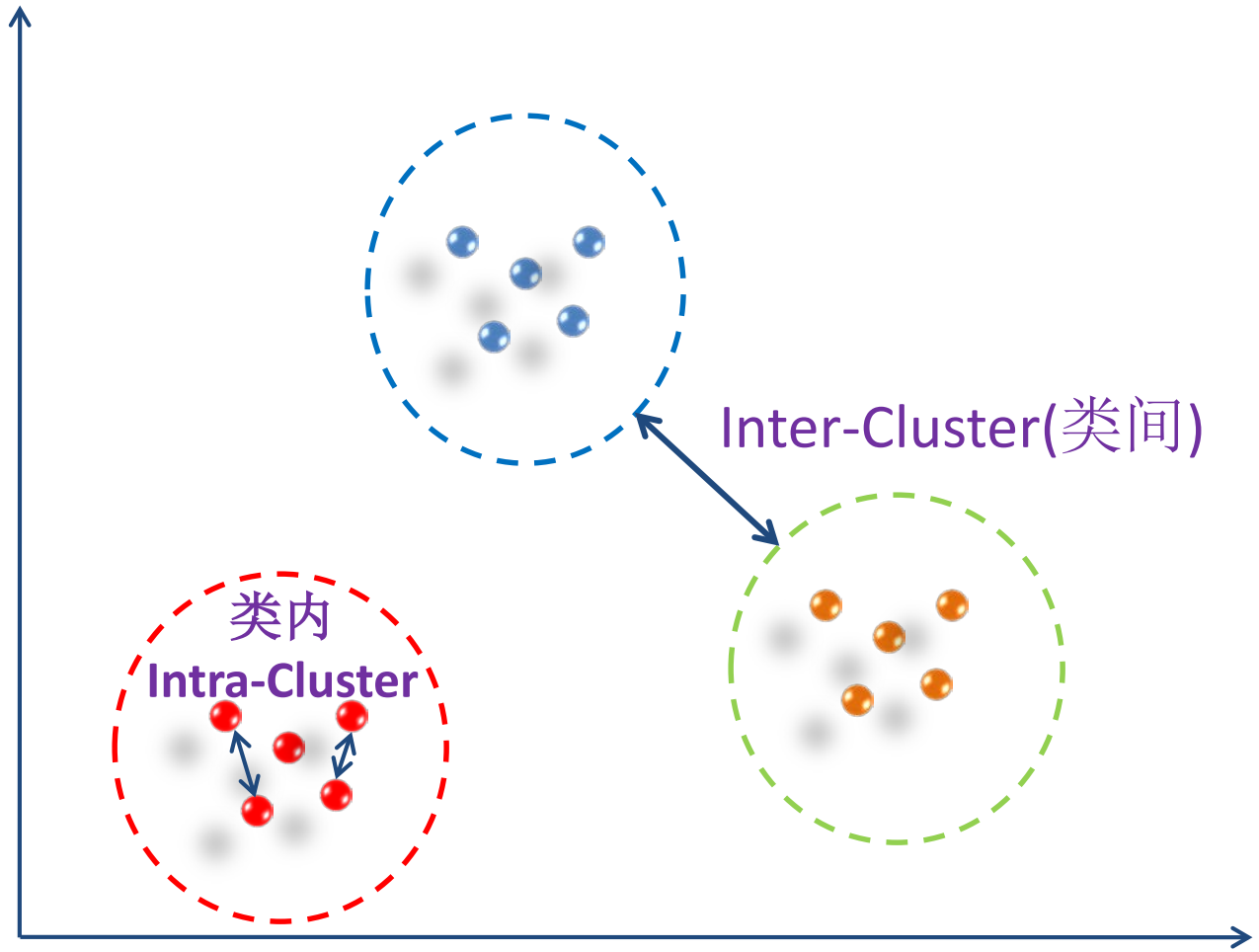
◆**Unsupervised Learning**

– 数据没有标签

– 数据驱动的



Clusters

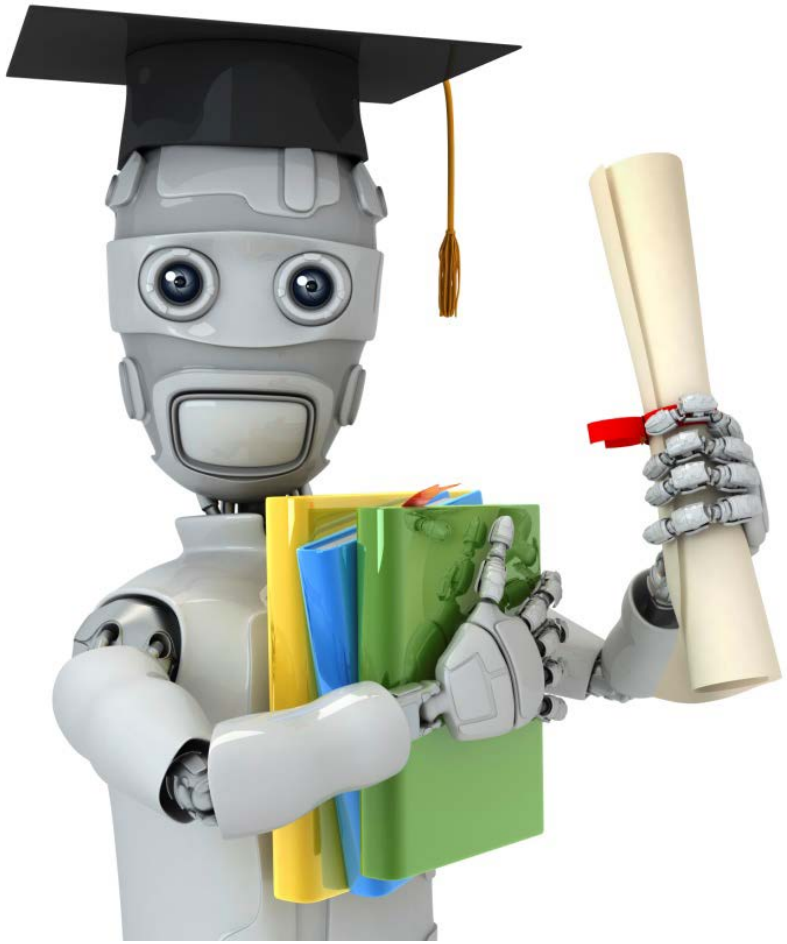


聚类的应用

- Marketing（营销，寻找行为相似的顾客群体）
  - Finding groups of customers with similar behaviours.
- Biology（生物，发现具有相似特征的动物或植物群）
  - Finding groups of animals or plants with similar features.
- Bioinformatics （生物信息学，聚类微阵列数据、基因和序列）
  - Clustering microarray data, genes and sequences.
- Earthquake Studies （聚类观测到地震中心，以识别危险区）
  - Clustering observed earthquake epicenters to identify dangerous zones.
- WWW（聚类博客数据，以发现类似访问模式的组）
  - Clustering weblog data to discover groups of similar access patterns.
- Social Networks（社交网络，发现亲密的个体群）
  - Discovering groups of individuals with close friendships internally.

图像分割

根据颜色（RGB）的相似性分割图像.



Machine Learning

Clustering

K-means algorithm

K-means algorithm

◆基本思想

给定参数**k**（**类别的个数**），把 *m*个对象分为k个组，使组内对象具有较高的相似度，而组间对象具有较低的相似度。

◆ Input:

- K (the number of clusters，组的个数)
- 训练集　D= $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$　$x^{(i)} \in \mathbb{R}^n$

◆ Output: K clusters

K-means algorithm

k-means算法的处理过程如下：

(1) 从数据集D中随机选择k个对象作为初始簇的中心;

(2) 计算每个对象与k个簇中心的距离，并将它划分到距离其最近的簇;

(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，直到簇中的对象不再变化。

K-means algorithm

◆通常，采用平方误差准则，即最小化每个对象到最近质心的欧几里得距离的平方和。

$$J_e = \sum_{i=1}^{k} \sum_{x \in D_i} \|x - C_i\|^2, \qquad C_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

其中，***k*** 簇的个数, C_i is 是第i个簇的中心, J_e 是误差的平方和.

◆**目标：**最后到达“类内的点都足够近，类间的点都足够远”的目标效果。

K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat { //m is the number of samples

for i = 1 to m // find the cluster centroid $c^{(i)}$ closest to the i^{th} object

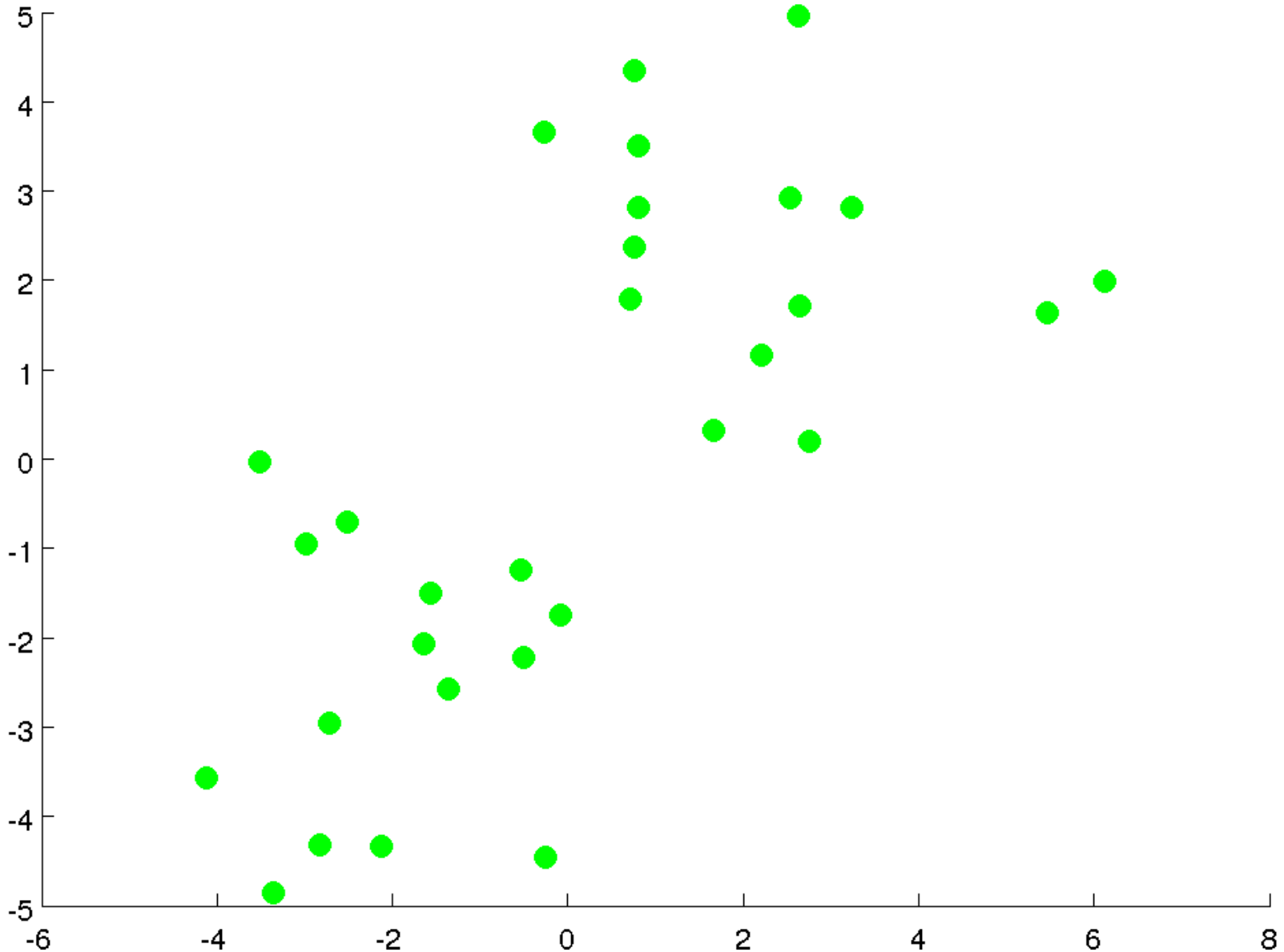
$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

for k = 1 to K //recalculate the centroids of K clusters

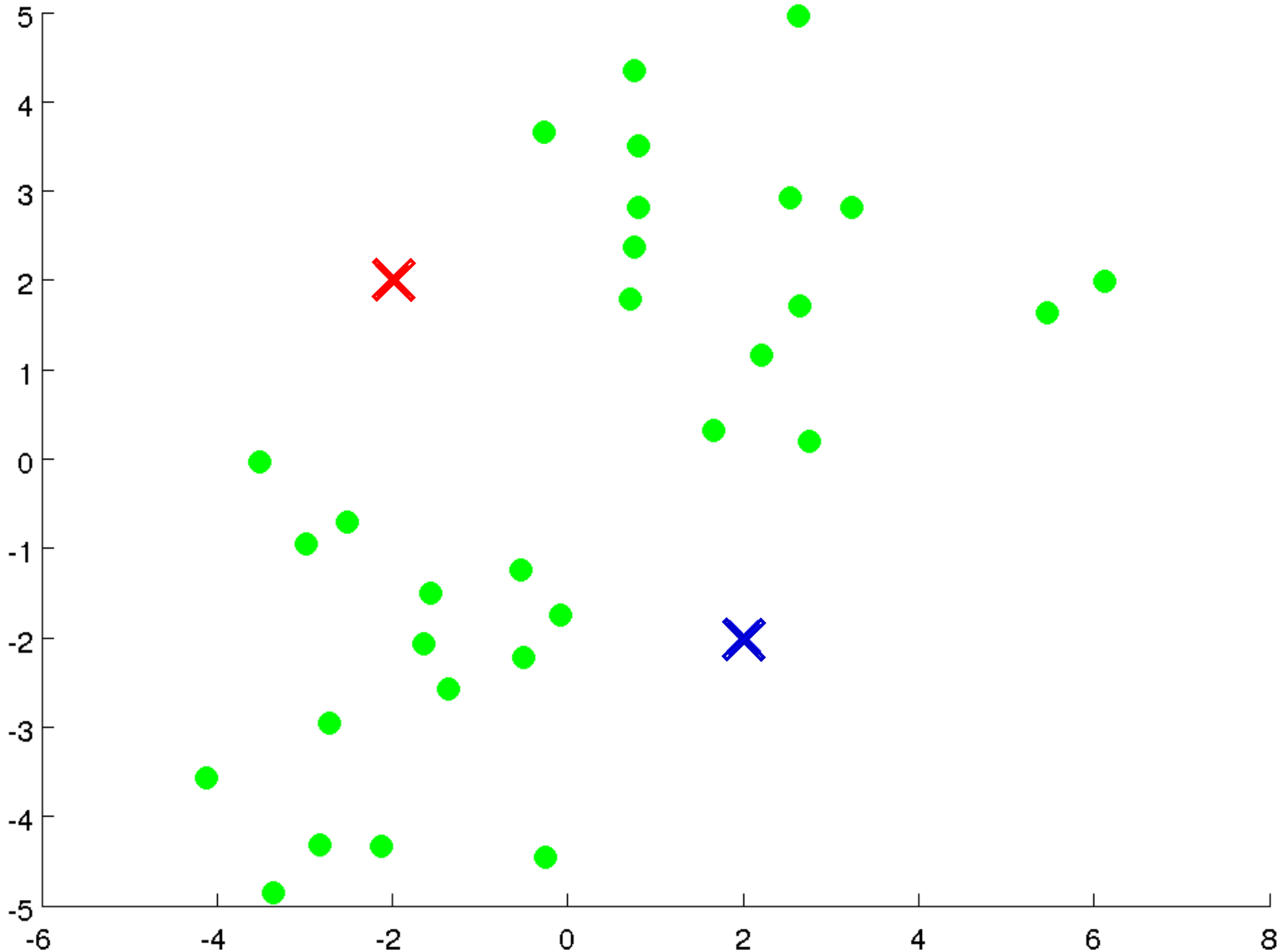
μ_k := average (mean) of points assigned to cluster k

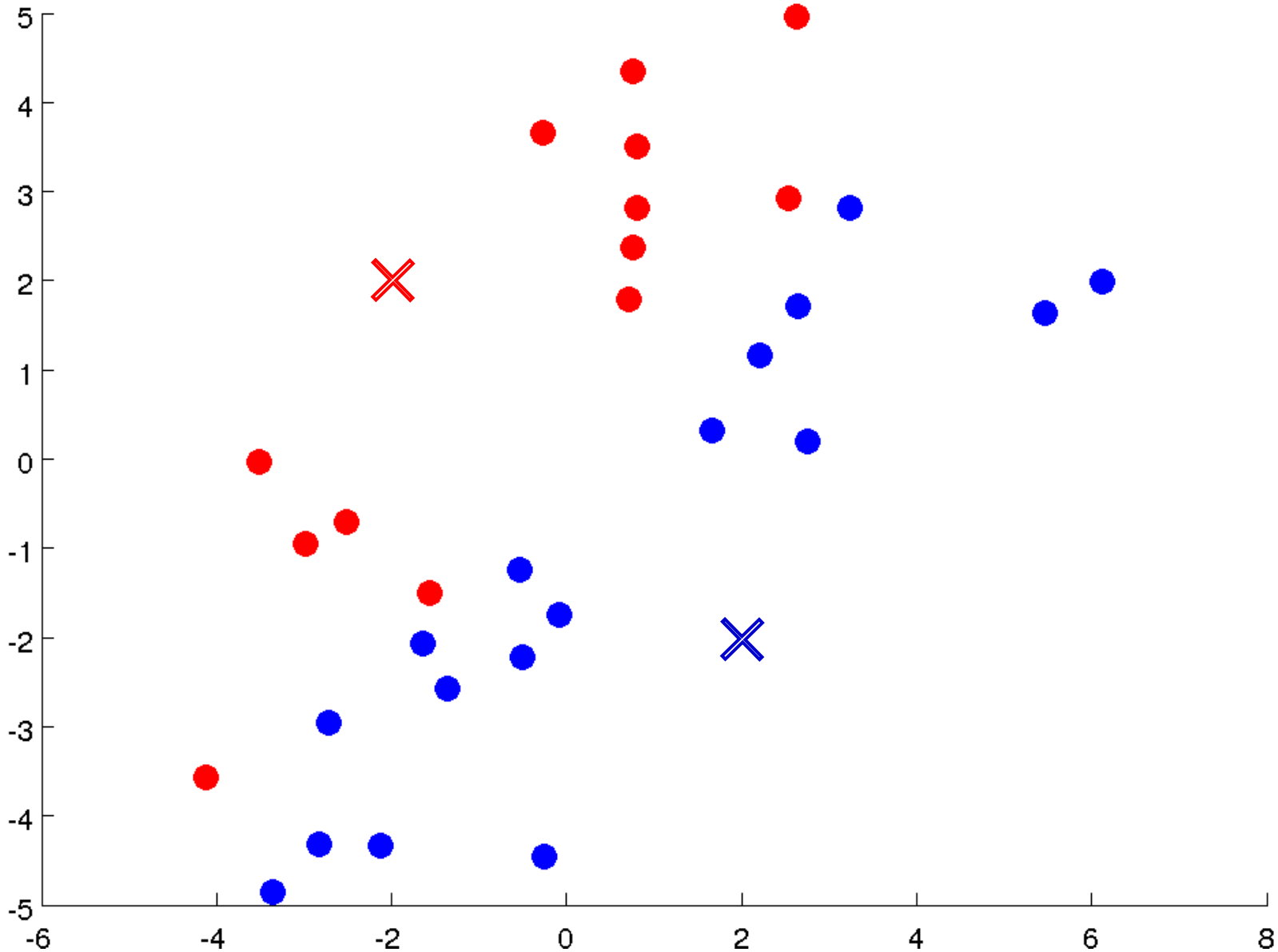
}

K=2

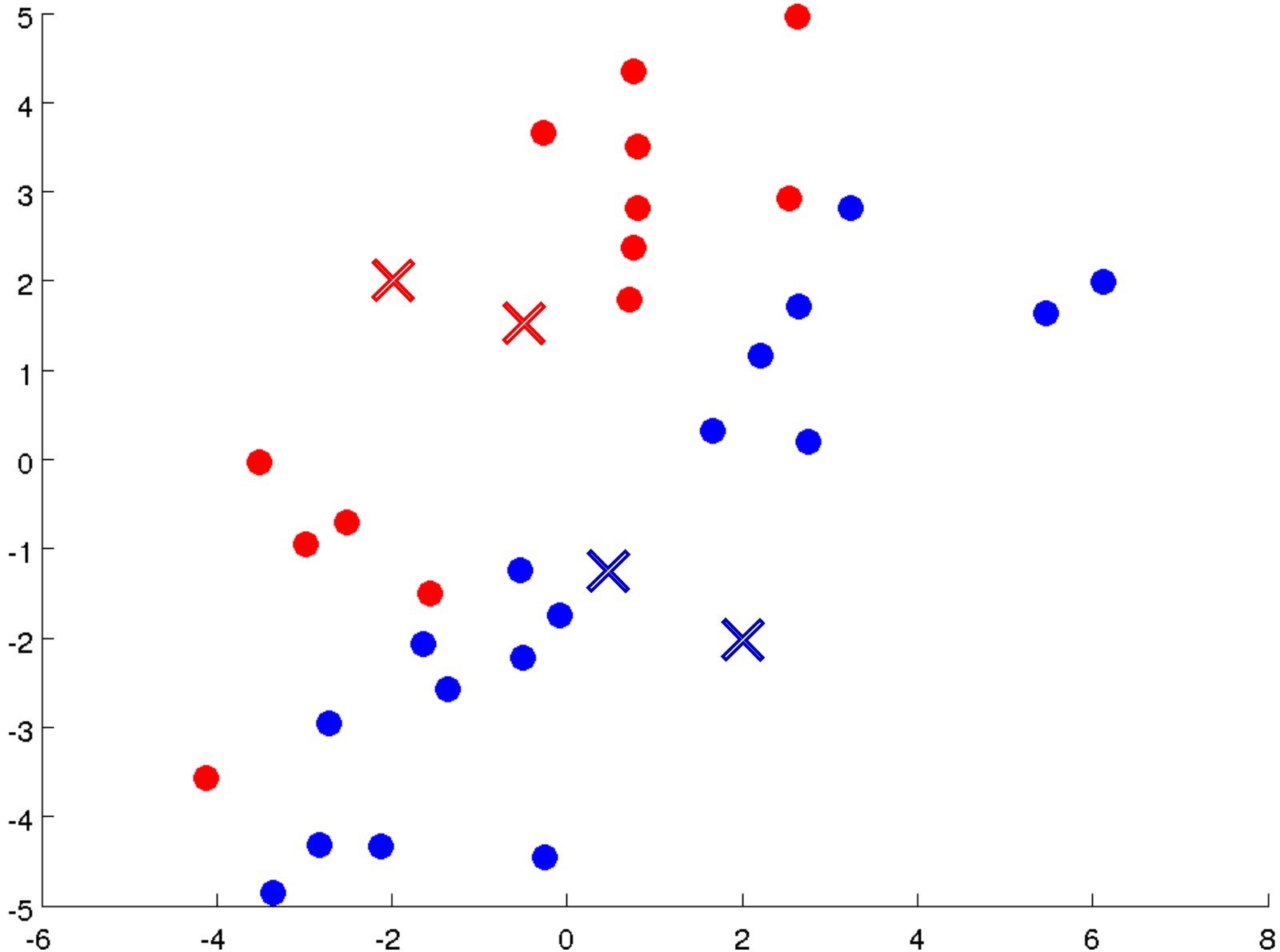


K=2

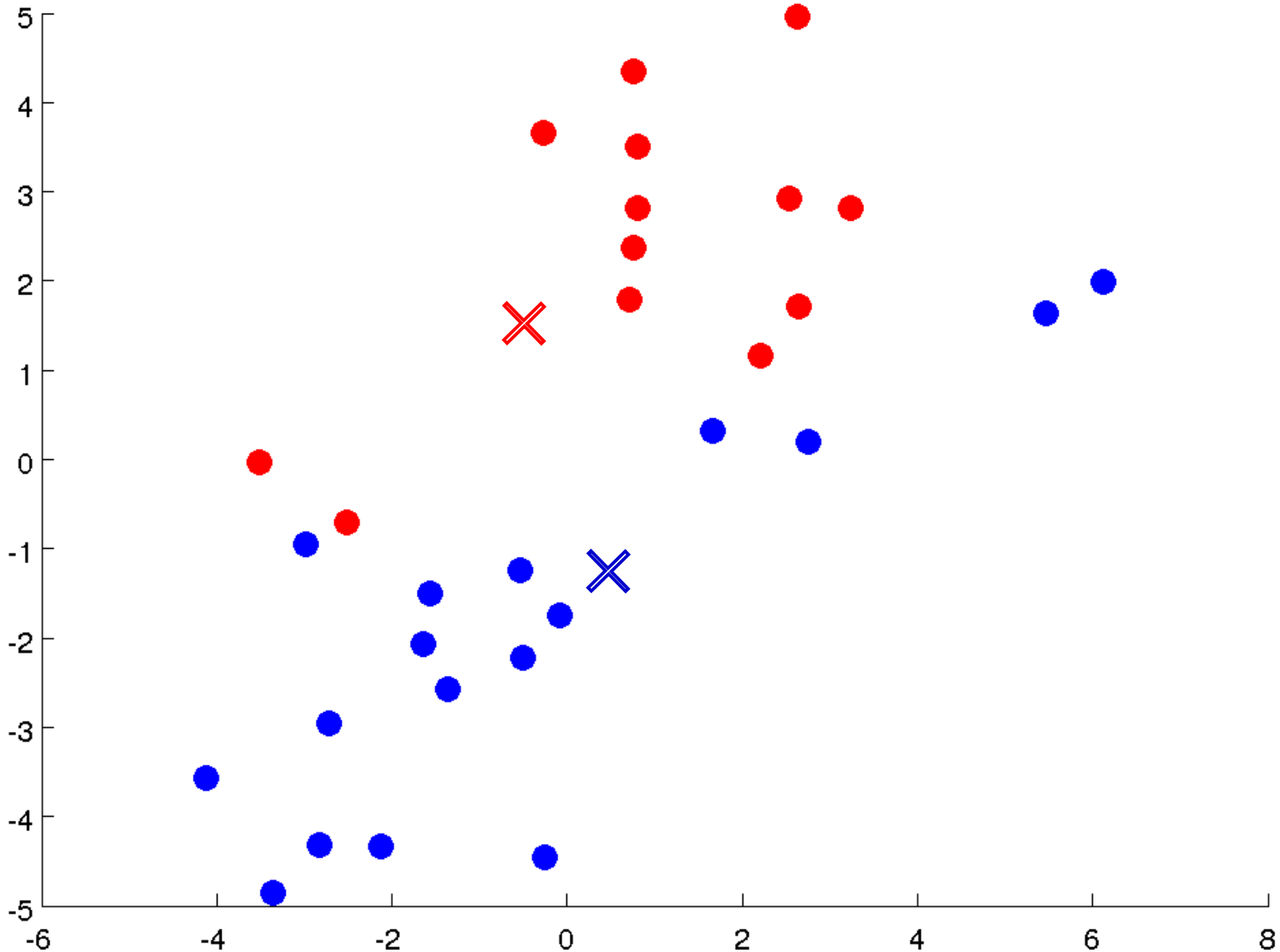



5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

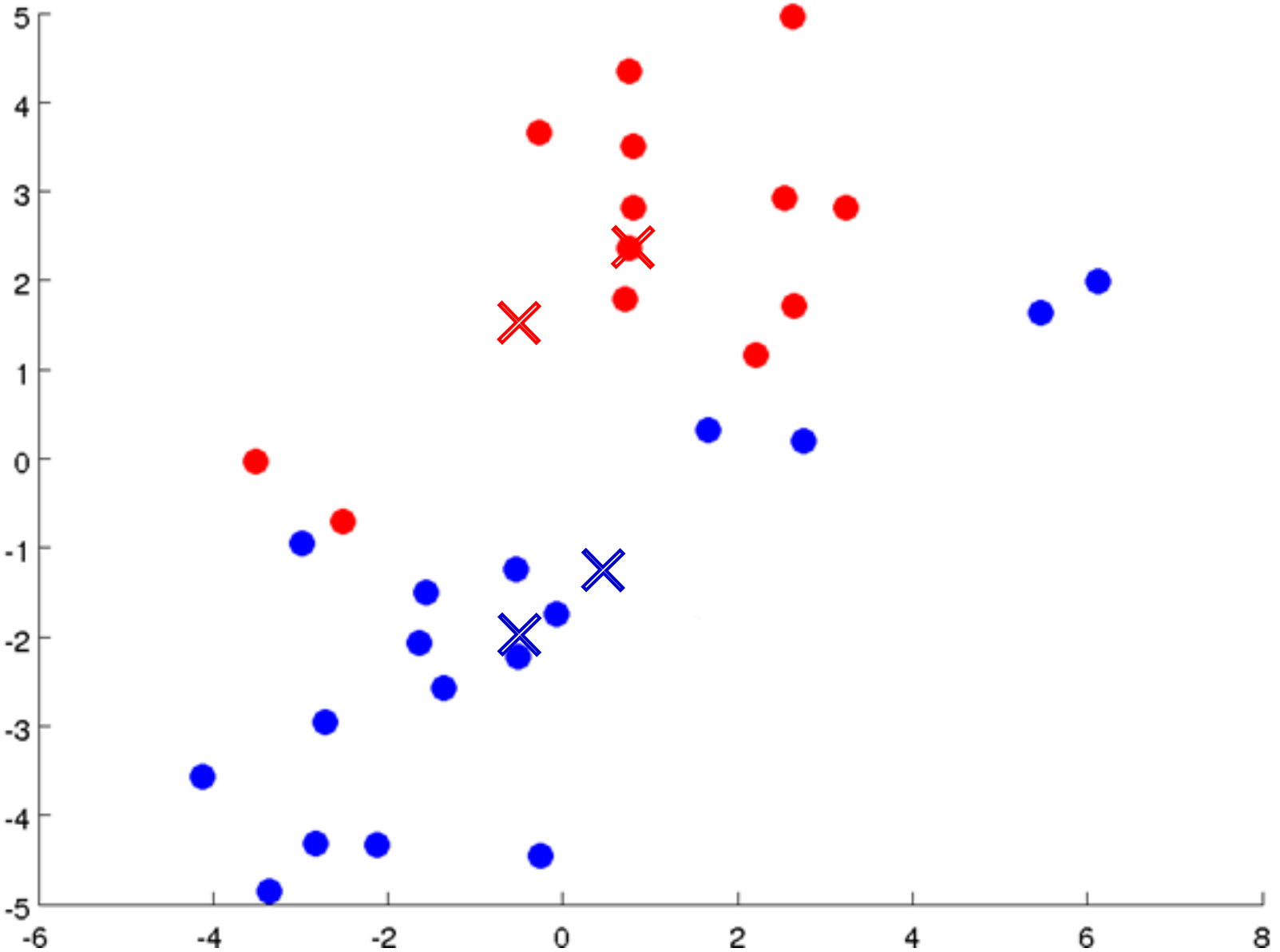
K=2



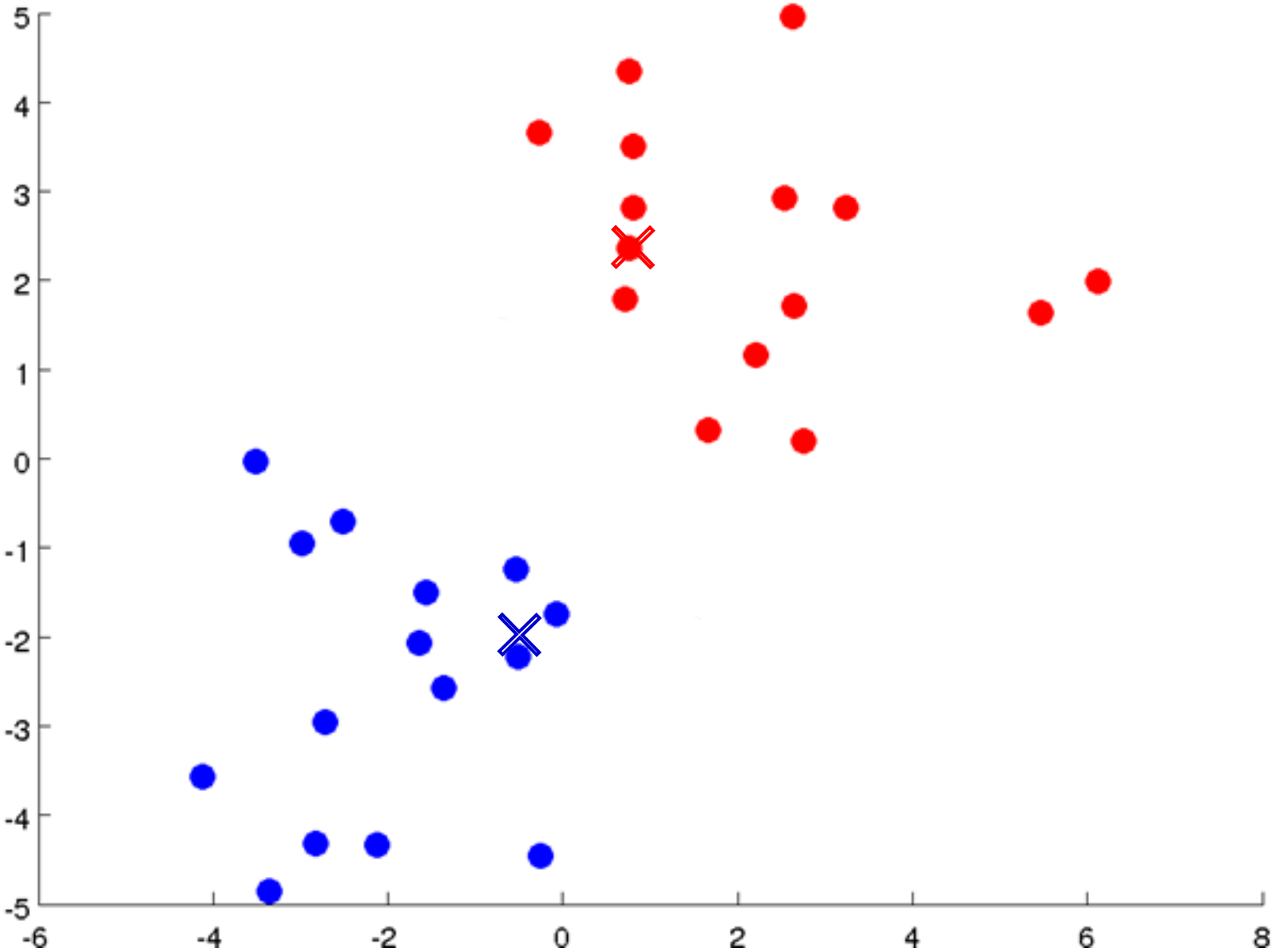
K=2



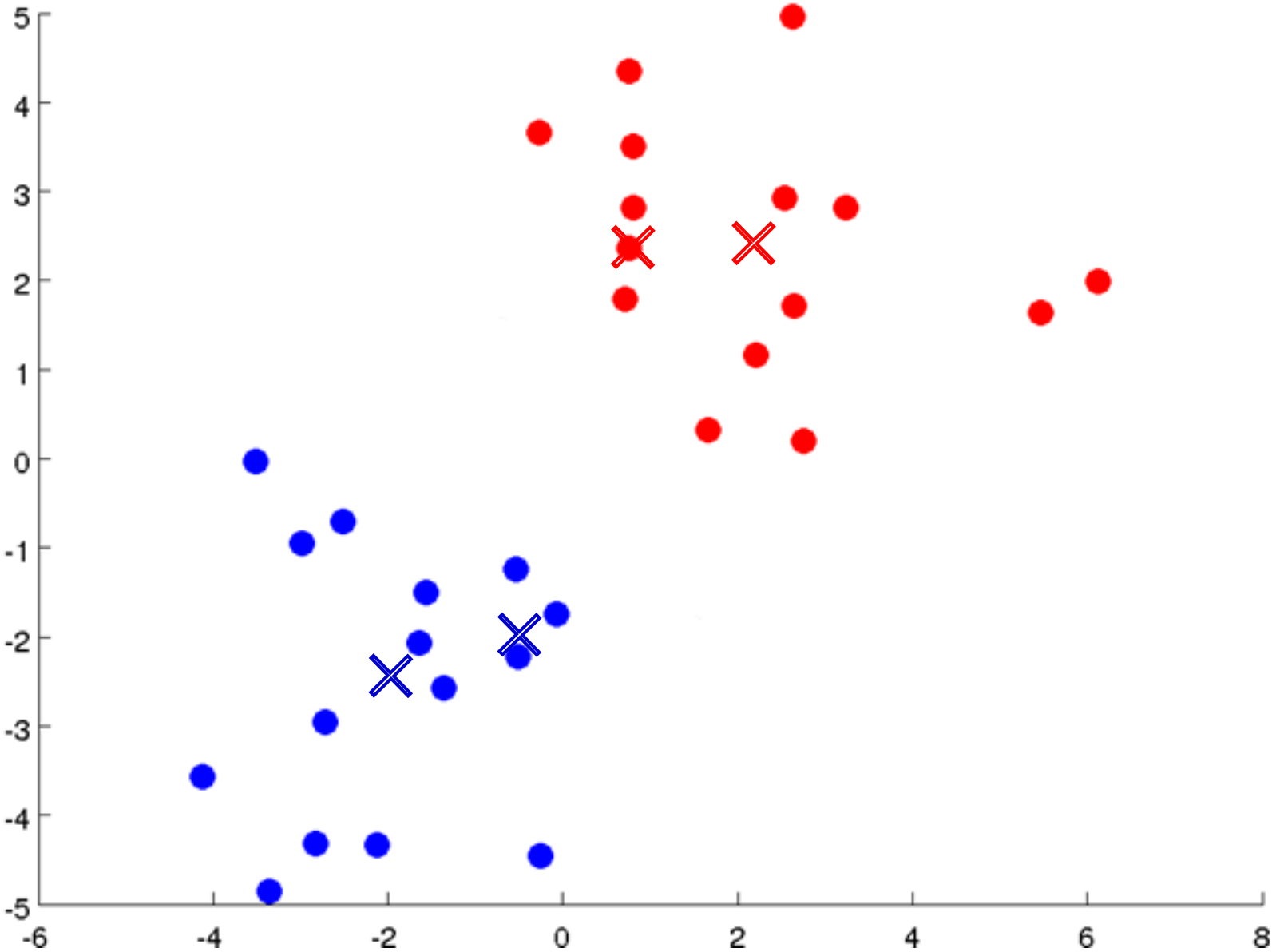
K=2



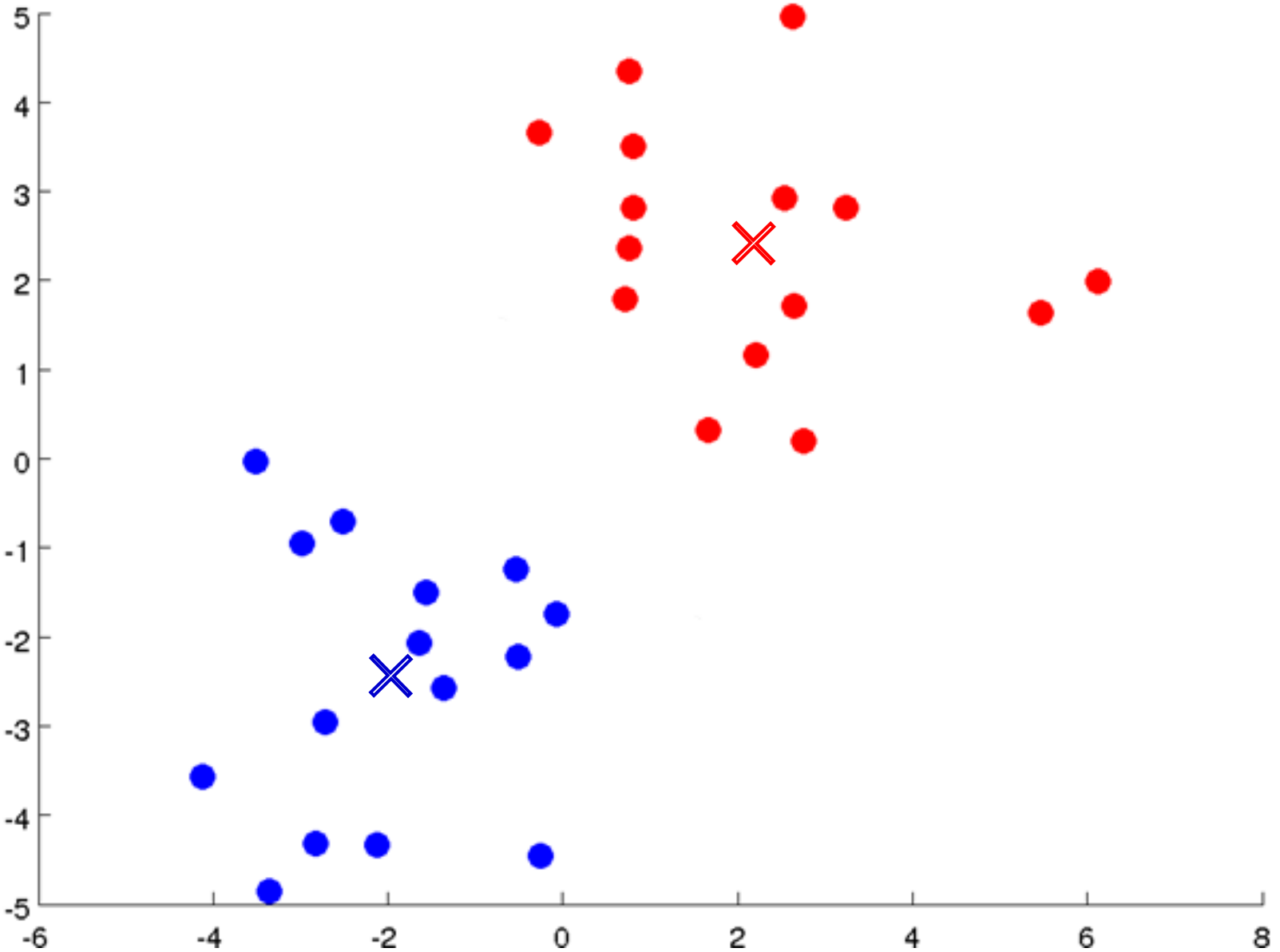
K=2



K=2



K=2



4.3.3 强化学习

- 什么是强化学习？
- 强化学习的类型
- **Q-Learning**
- 强化学习的应用

What is Reinforcement Learning

- 强化学习的灵感来自于行为心理学。
- 关注于智能体如何在环境中采取行动，为了使累积回报最大化。
- 强化学习中由环境提供的强化信号是Agent对所产生动作的好坏作一种评价(通常为标量信号)，而不是告诉Agent如何去产生正确的动作。
- 由于外部环境提供了很少的信息，Agent必须靠自身的经历进行学习。
- 通过这种方式，Agent在行动——评价的环境中获得知识，改进行动方案以适应环境。

强化学习

自学骑车，没人教 (learn by oneself)

Trial and Error
（试错法）

+reward（奖励）：ride stably

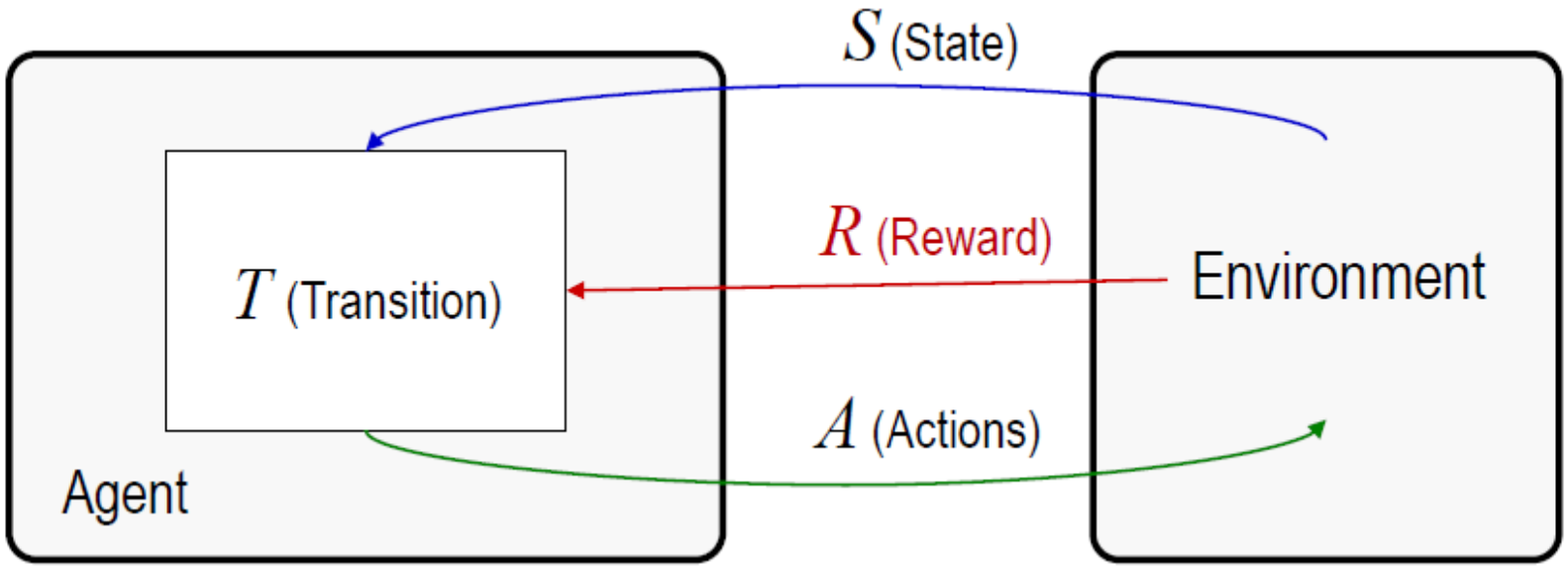
-punish（惩罚）：fall

把奖励最大化的过程

强化学习的思路：通过不断试错，使下一次采取的动作能够得到更多奖励，并且把奖励最大化。

什么是强化学习

- 在强化学习中，其学习器是一个进行决策的智能体，在环境下采取行动并获得这些动作的回报。
- 经过一系列试错运行之后，该智能体能够学到最优策略。
- 该策略是经过一个阶段的动作以及与环境的交互之后，使其回报最大化。



三种范式之比较

- ***Supervised learning*** **有监督学习**
  - 通过标注数据（训练对象，教师信号）提供输入和输出对儿：**有{X，Y}**
  - 从对象中学习
- ***Unsupervised learning*** **无监督学习**
  - 发现无标注数据集中隐藏的结构。**有{X}，但无 {Y}**
  - 自我学习
- ***Reinforcement learning*** **强化学习**
  - 不提供输入和输出对儿，专注于在线的性能优化。**既无{X}，也无{Y}**
  - 反馈学习（经验学习、在线学习）

强化学习的类型

1) Model-based 基于模型学习

构建环境的模型。

- 首先以**马可夫决策过程**方式动作，并学习 T 和 R;
- 然后用学习的 T 和 R 进行数值迭代或策略迭代。

2) Model-free 无模型学习

学习策略而没有任何模型。

- 避开学习 T 和 R 的过程，采用直接评估策略。
- 基于预测的时间差分(TD) 法。

Q-learning

Q-learning is a Model-free Method. Q 表示（状态，动作）对的值。

- Q-Learning是通过Q估计来进行决策，更新学习。
- Q 估计表示为Q(s,a)，是在某状态 s 下(s∈S)，采取 动作a (a∈A)能获得收益的期望;
- 环境会根据agent采取的动作，反馈相应的回报 r;
- 算法的主要思想就是将State与Action构建成一张Q-table来存储Q值，然后根据Q值来选取能够获得最大收益的动作。

强化学习的新算法

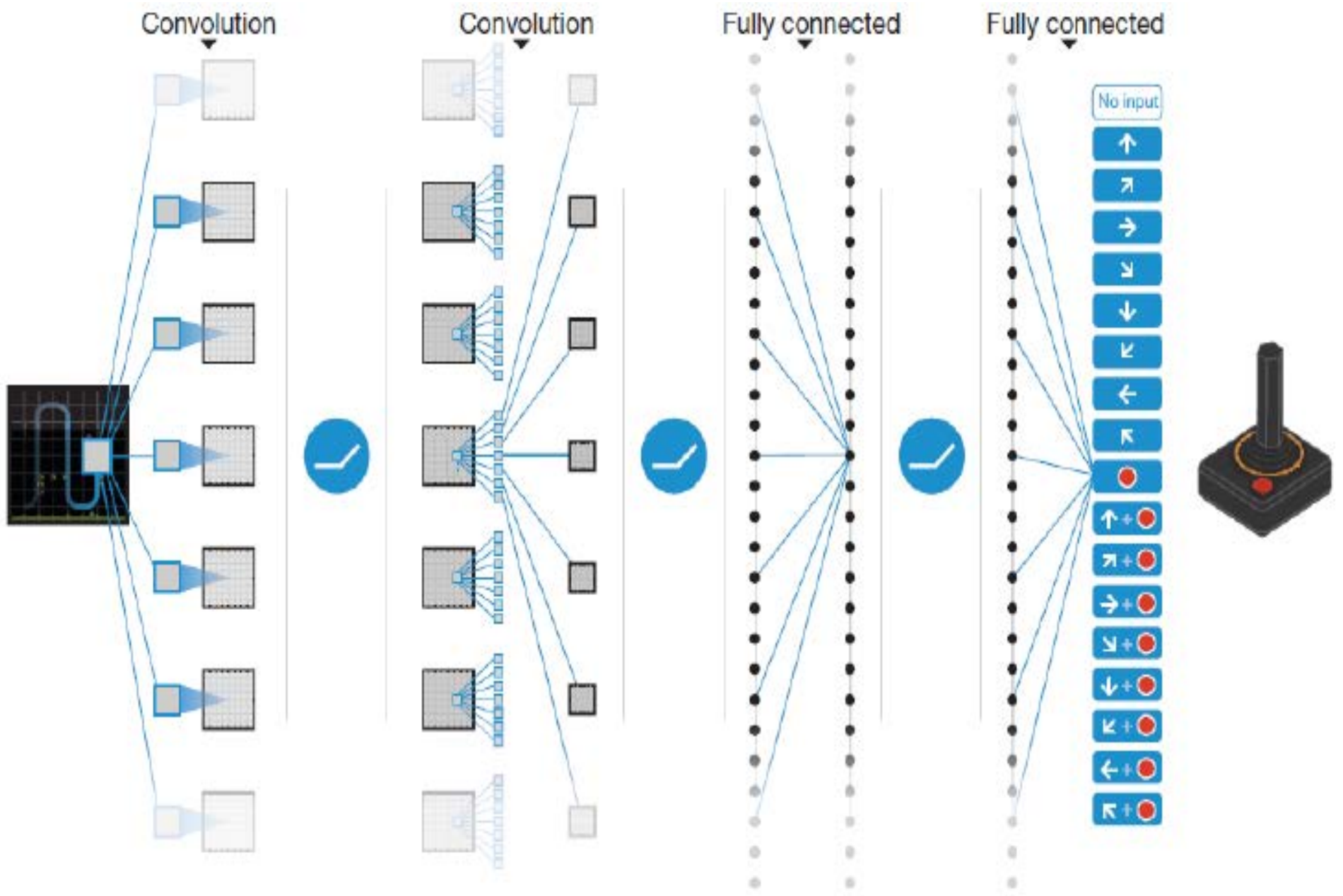
- 2015年2月，谷歌DeepMind发表了**深度Q-网络**，通过深度强化学习达到人类水平的操控。

- **Deep Q-Network (DQN)** 深度Q-Network

  将CNN与Q-学习（一种强化学习的形式）相结合,

Case Study: Deep Reinforcement Learning

输入是原始像素，
输出是估计回报的
价值函数。



强化学习的典型应用

- **Robots** 机器人
  - Robotic arms 机器人手臂

    控制得到最有效的电机组合。
  - Robot navigation 机器人导航

    可通过负反馈来学会碰撞躲避行为。
- **Computer games** 计算机游戏
  - Backgammom, 西洋双陆棋
  - Chess, 国际象棋
  - Go. 围棋

4.4 Models in Machine Learning

- 概率模型
- 几何模型
- 逻辑模型
- 网络模型
  - Artificial Neural Networks (ANN)
  - Convolutional Neural Networks (CNN)
  - Deep Neural Networks (DNN)