Foundation of Artificial Intelligence

人工智能基础

李翔宇
软件学院
Email: lixiangyu@bjtu.edu.cn

Machine Learning

5.1 Perspectives about Machine Learning

5.2 Tasks in Machine Learning

5.3 Paradigms in Machine Learning

5.4 Models in Machine Learning

Tasks in Machine Learning

5.2.1 Classification 分类

5.2.2 Regression 回归

5.2.3 Clustering 聚类

5.2.4 Ranking 排名

5.2.5 Dimensionality Reduction 降维

Classification

What is Classification

- **A longer description**
  Classification is the task of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

- **A shorter description**
  To resolve such problems where the output is divided into two or more categories.
- **A very short description**
  Assign a category to each item.

Classification

- How Classification Works
- Linear and Nonlinear
- Dimensions and Classes
- Applications and Algorithms

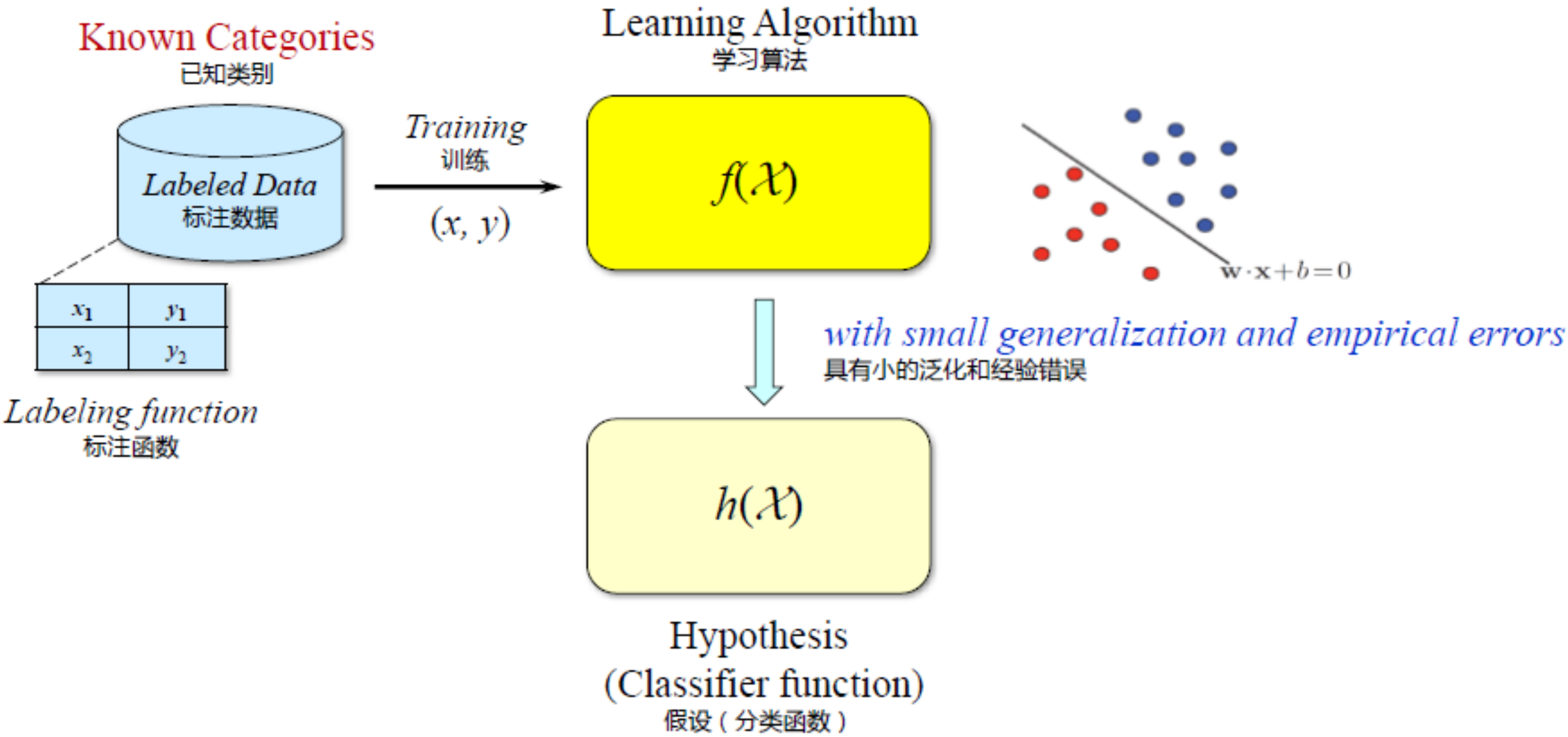
How Classification Works

Classifier 分类器

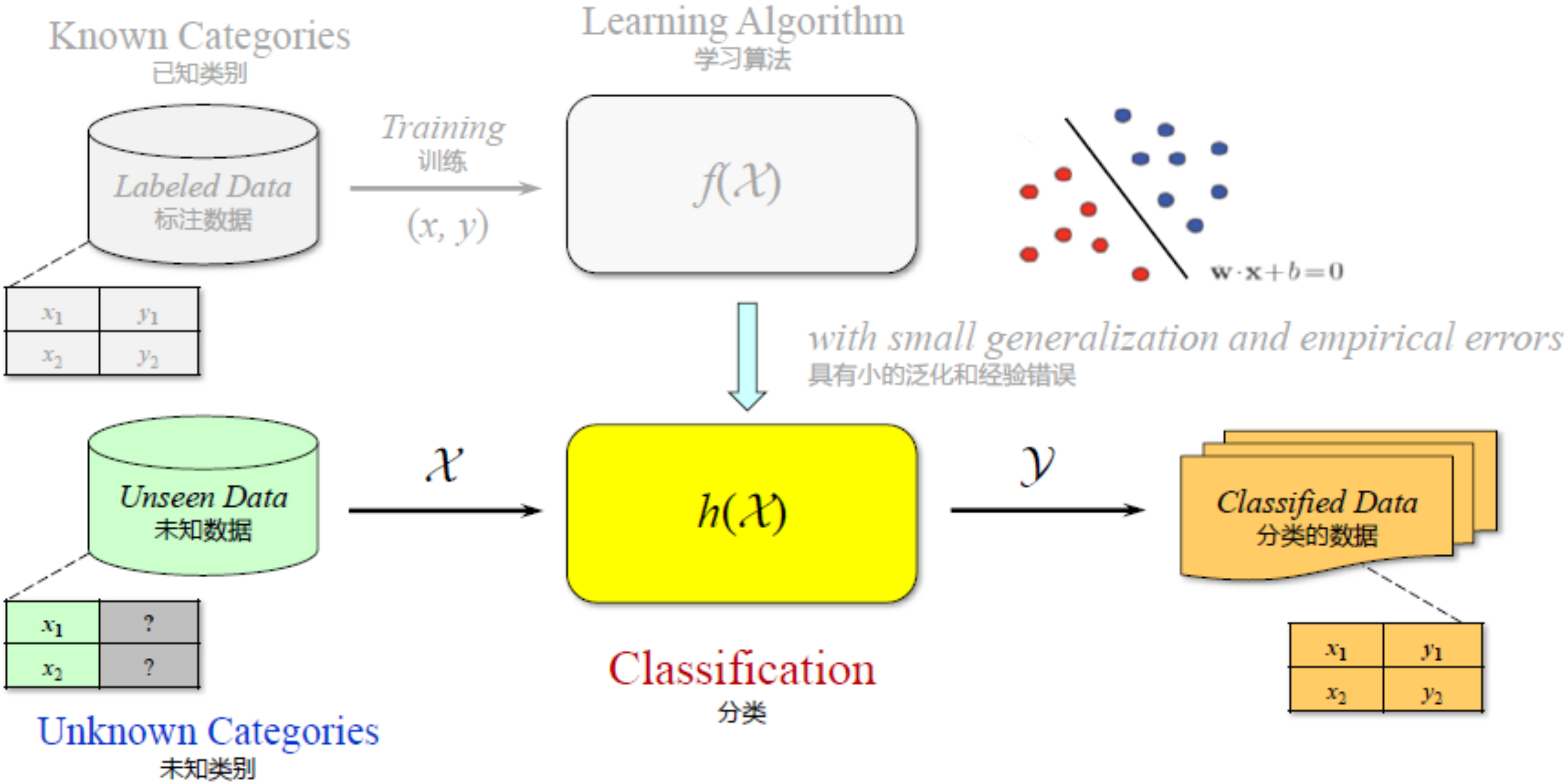
- **About classifier**
  An algorithm that implements classification, especially in a concrete implementation, is known as a classifier.

- **About classifier function**
  The term “classifier” sometimes also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category.

Classification: Training 分类：训练



Classification: Training 分类：训练



A Formal Description of Classification

Let $\mathbb{R}^n$ ($n \geqq 1$) denote a set of n-dimensional real-valued vectors, input space X is a subset of $\mathbb{R}^n$, output space Y is a set of categories, D is an unknown distribution over X$\times$Y, then:

- Let target labeling function: f :X→Y
- Training set (Labeled training sample set):

$$S = \{(x(i), y(j)) \mid (x, y) \in X \times Y, i \in [1, m], j \in [1, n]\}$$

- Classification algorithm:

  Let a hypothesis set H are the mapping X to Y, to determine a hypothesis (classifier function):

$$h: X \rightarrow Y \quad \text{and} \quad h \in H$$

  with small generalization error $R(h) = \Pr x\, [h(x) \neq f(x)]$

A Formal Description of Classification

Classification:

Given a testing data set of unknown categories:

$$X= \{ x^{(i)} \mid x \in X, i \in [1, m] \}$$

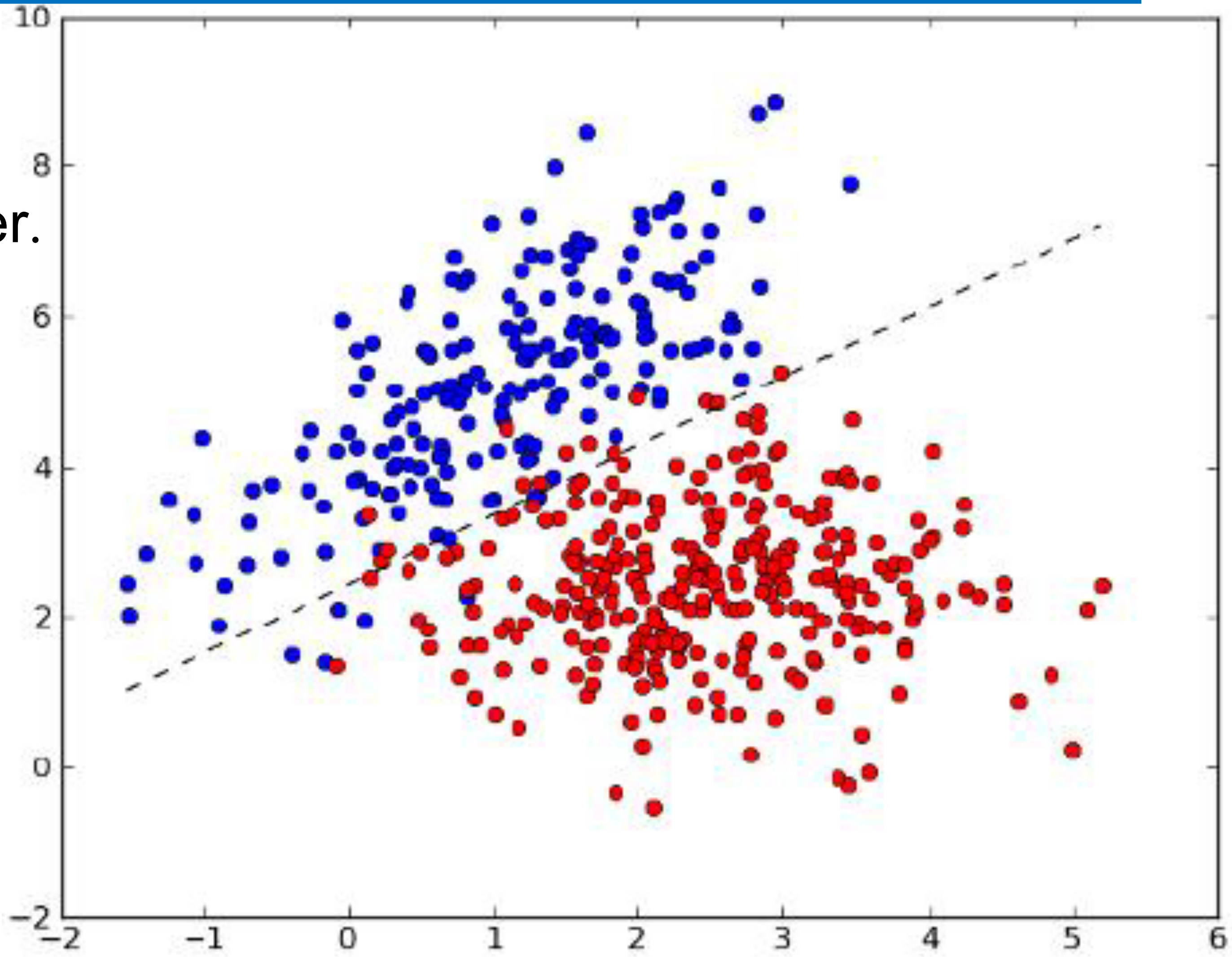
Using the classifier function $h(X) = Y$ determined at above to predicate classifying results:

$$Y = h(X)= \{y^{(j)} \mid y \in Y, j \in [1, n], \ h(x)=y\}$$

where Y is the set of known categories.

Linear and Nonlinear

- **Linear Classification**
  - Linear Classification is doing classification by a linear classifier.



- A linear classifier is a linear discriminant function with a linear decision boundary.

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

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

where,

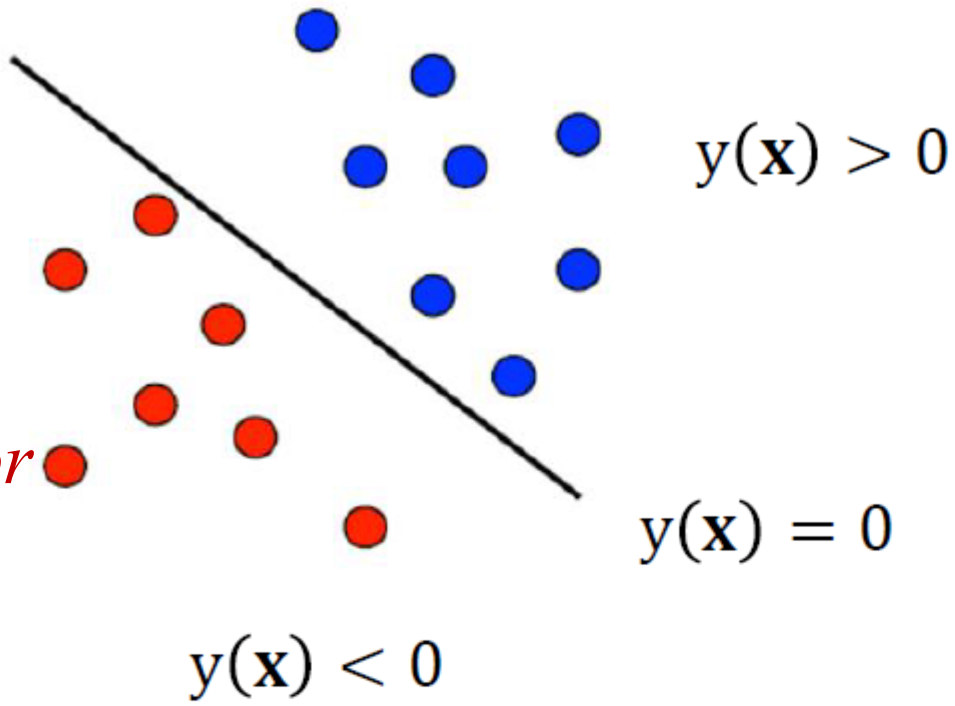
W denotes a row vector, called a *weight vector,*

$$\mathbf{w} = (w_1, \ldots, w_n)$$

X denotes a column vector, called an *input vector*

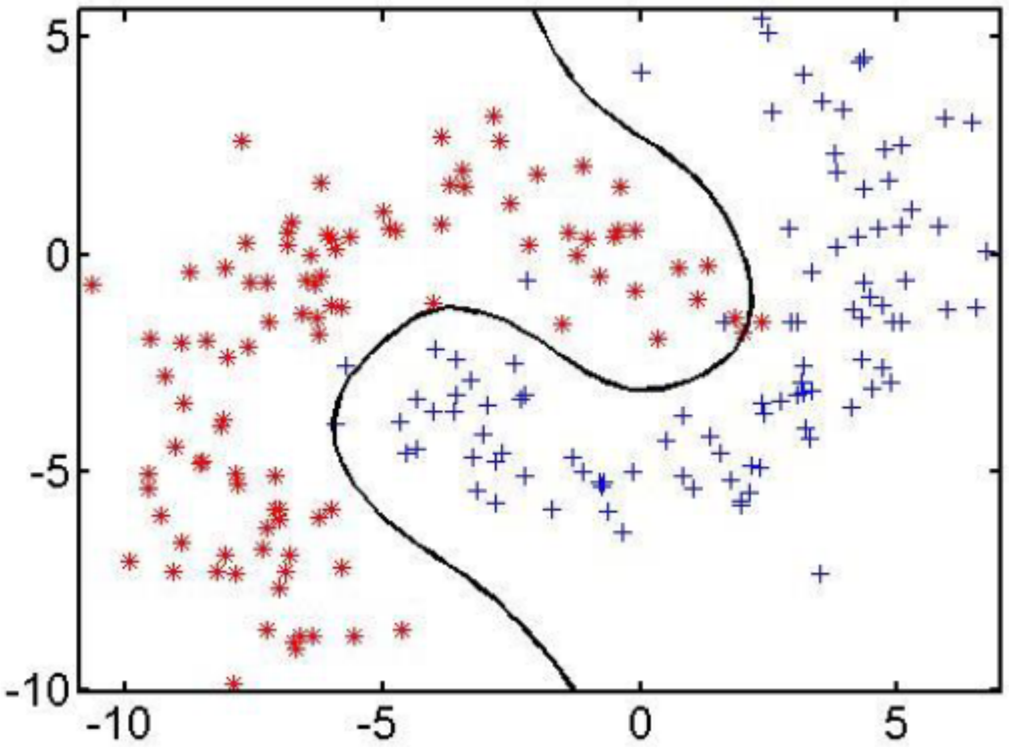
$$\mathbf{x} = (x_1, \ldots, x_n)^T$$

b denotes a bias.



Nonlinear Classification

- Nonlinear Classification is doing classification by a nonlinear classifiers.
- A nonlinear classifiers have nonlinear decision boundaries, and possibly discontinuous decision boundaries.

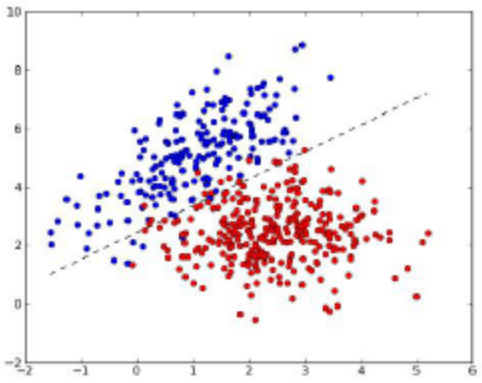


E.g., a nonlinear classifier in SVM is a nonlinear kernel function.

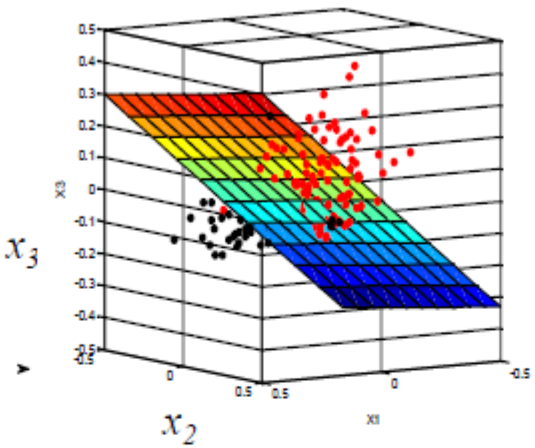
Dimensions and Classes

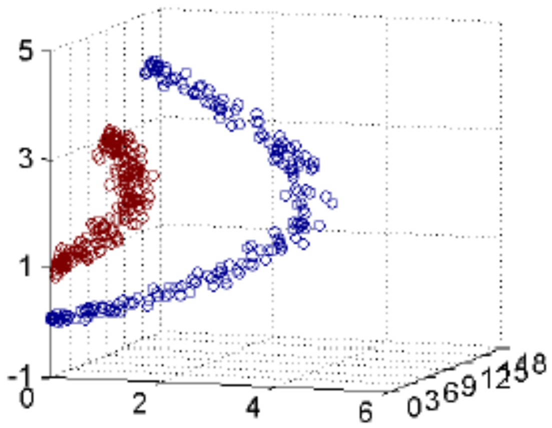
◆ Dimensions

If the problem space is *n* dimensional then its classifier is *n*-1 dimensional hyper-plane. e.g.,



2-dimensions
2维





3-dimensions
3维

in 2-dimensions, the hyper-plane is a line

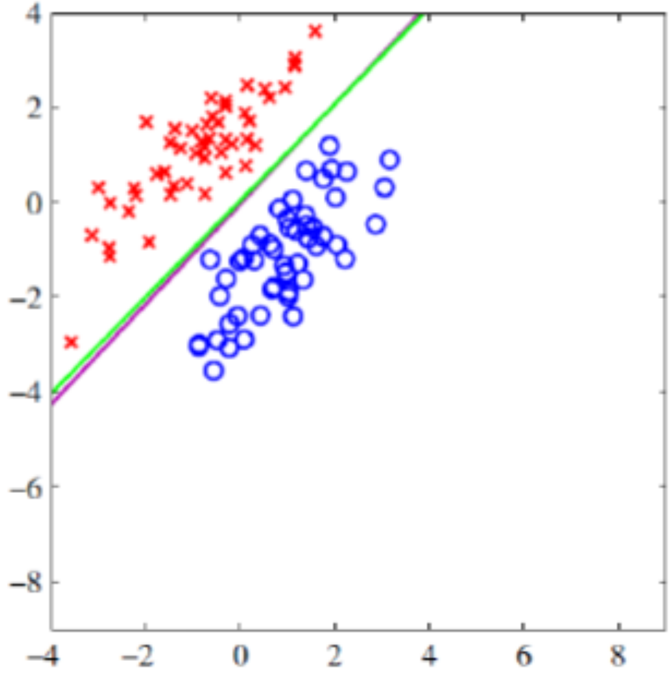
in 3-dimensions, the hyper-plane is a plane

Classes

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

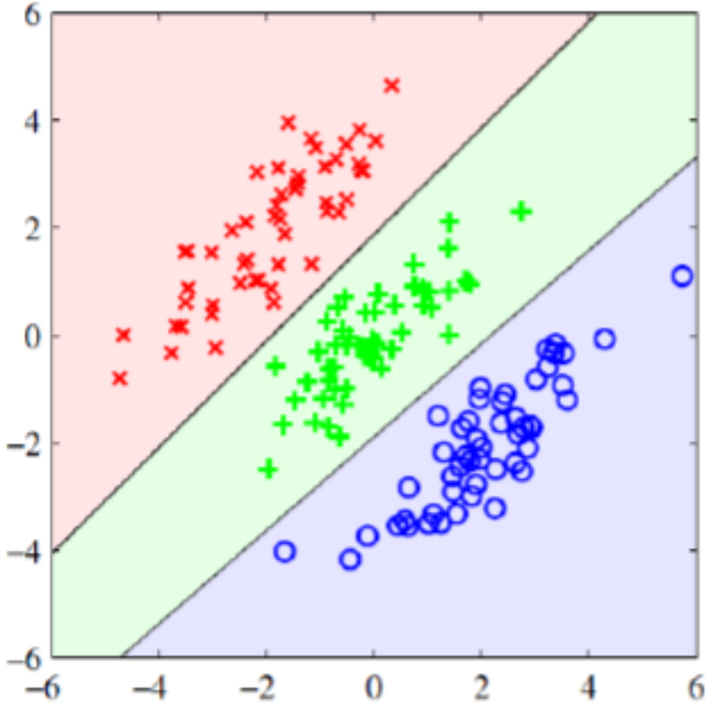
Two classes: $k = 2$

Multiple classes: $k > 2$



Two classes

二元分类



Three classes

三元分类

Case Study: Softmax Classifier

- It is a multiclass classifier, implemented by a softmax function.

- **Softmax function** maps a K-dimensional vector $\mathbf{x}$ of arbitrary real values to a K-dimensional vector $\sigma(\mathbf{x})$ of real values (range 0 to 1, add up to 1).

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

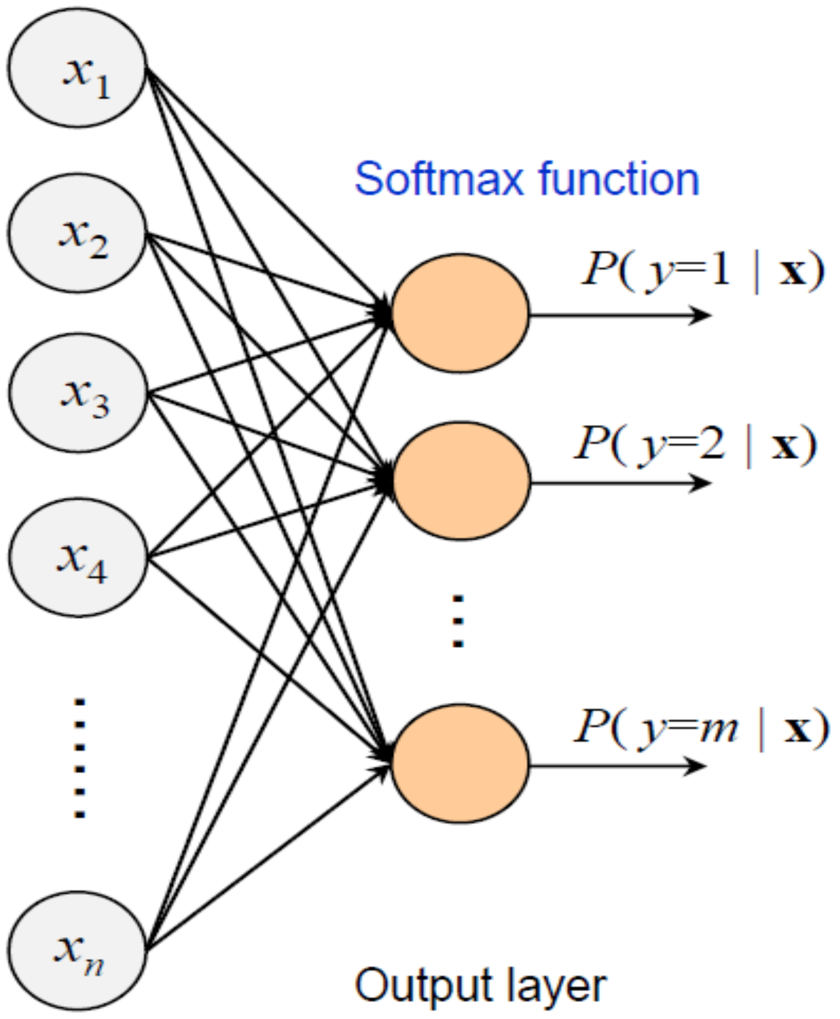
- In probability theory, the output of the softmax function can be represented a categorical distribution, i.e., a probability distribution over K different outcomes.

$$P(y = j \mid \mathbf{x}) = \frac{e^{\mathbf{x}^{\mathrm{T}}\mathbf{w}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}^{\mathrm{T}}\mathbf{w}_k}}$$

Case Study: Softmax Classifier

Softmax function has been used in various **multiclass classification** methods, such as:

- multinomial logistic regression,
- multiclass linear discriminant analysis,
- naive Bayes classifiers,
- artificial neural networks (ANN),
- reinforcement learning.



Softmax function used in ANN as the final layer for multiclass classification.

Typical Applications of Classification

- Computer vision 计算机视觉
  - Face, handwriting recognition 人脸、手写体识别
  - Action recognition 动作识别
  - Medical image analysis 医学图像分析
  - Video tracking 视频跟踪
- Pattern recognition 模式识别
- Biometric identification 生物特征识别
- Statistical natural language processing 统计自然语言处理
- Document classification 文档分类
- Internet search engines 互联网搜索引擎
- Credit scoring 信用评分

Typical Algorithms of Classification 分类的典型算法

- 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
- **K-nearest neighbors** （KNN）K-近邻
- 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
- **Support vector machine** (SVM) 支撑向量机

Regression

- How Regression Works
- Linear and Nonlinear
- Applications and Algorithms

What is Regression 什么是回归

- **A longer description**

  Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.

- **A shorter description**

  To resolve such problems where the output is a **real continuous value**.

- **A very short description**

  Predict a real value for each item.

Regression vs. Classification

- **Similarity**

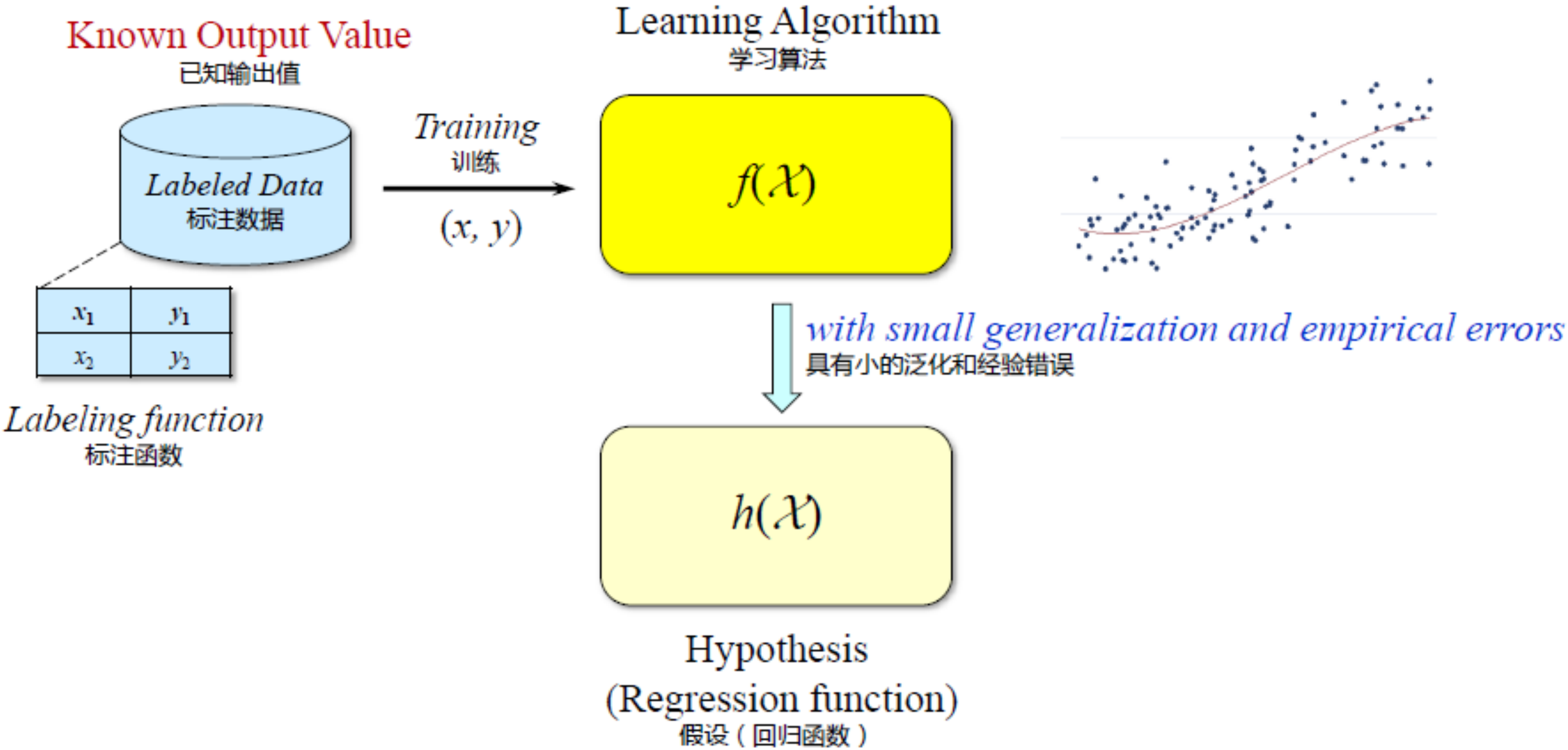
  Need training processing

- **Difference**

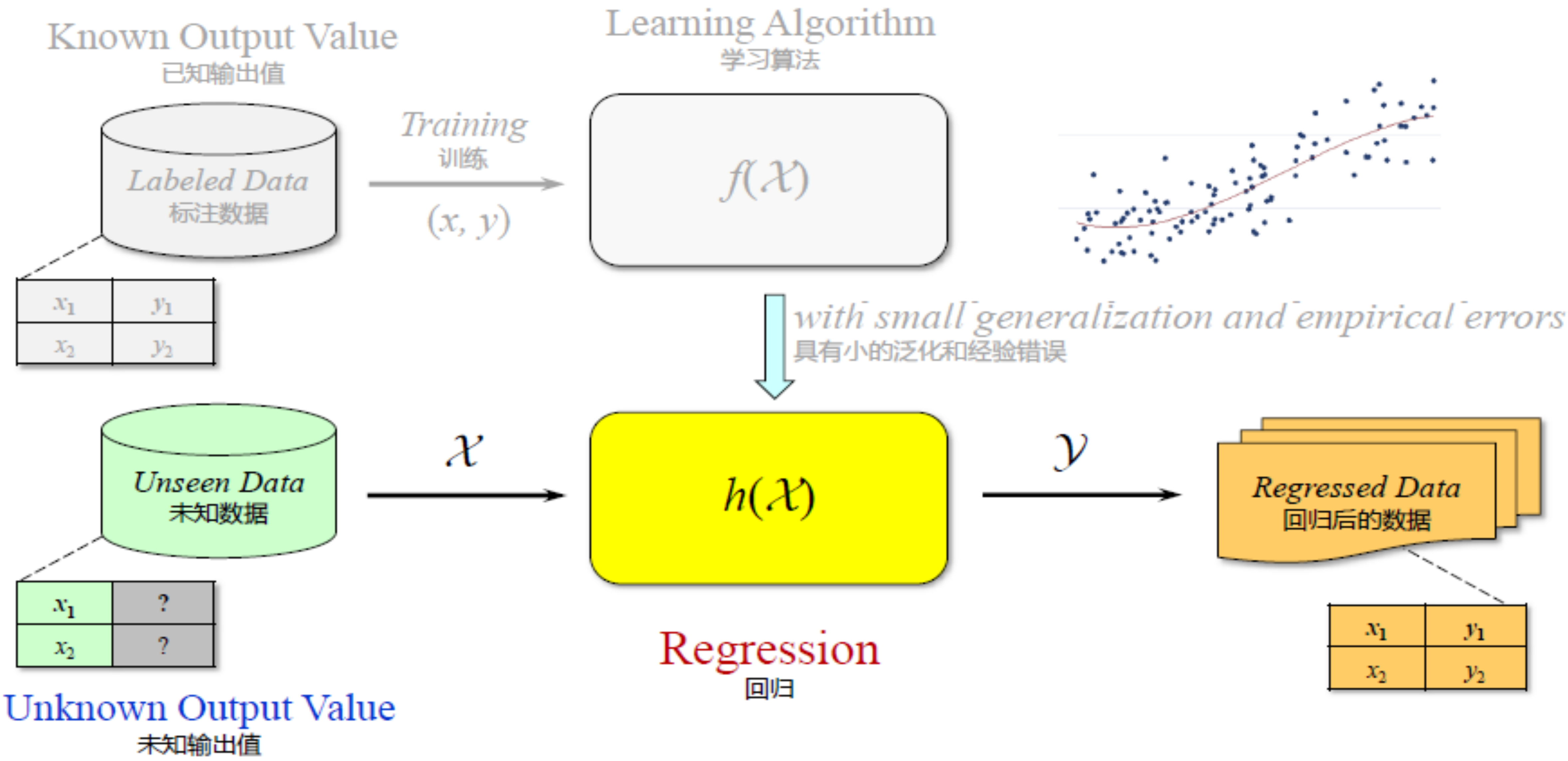
  As shown in the following table

	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



Regression: Testing



A Formal Description of Regression

Let $\mathbb{R}^n$ ($n \geqq 1$) denote a set of n-dimensional real-valued vectors, $\mathbb{R}$+ is a set of non-negative real numbers, input space X is a subset of $\mathbb{R}^n$, output space Y is a set of real numbers $\mathbb{R}$+, D is an unknown distribution over X×Y, then:

- Let target **labeling function**:

$$f : X \rightarrow Y$$

- **Training set** (Labeled training sample set):

$$\mathcal{S} = \{(x^{(i)}, y^{(i)}) \mid (x, y) \in \mathcal{X} \times \mathcal{Y}, i \in [1, m]\}$$

A Formal Description of Regression (cont.)

- **Regression algorithm:**
  Given hypothesis set H, to determine a hypothesis (regression function)

$$h: X \rightarrow Y \text{ and } h \in H$$

With small generalization error $R(h)$:

$$R(h) = \mathrm{E}x\,[L(\,h(x), f(x)\,)]$$

where $L(\,h(x), f(x)\,)]$ is the distance between $h(x)$ and $f(x)$.

A Formal Description of Regression

- **Regression**

  Given a testing data set of unknown output:

$$\mathcal{X} = \{x^{(i)} \mid x \in \mathcal{X},\ i \in [1, m]\}$$

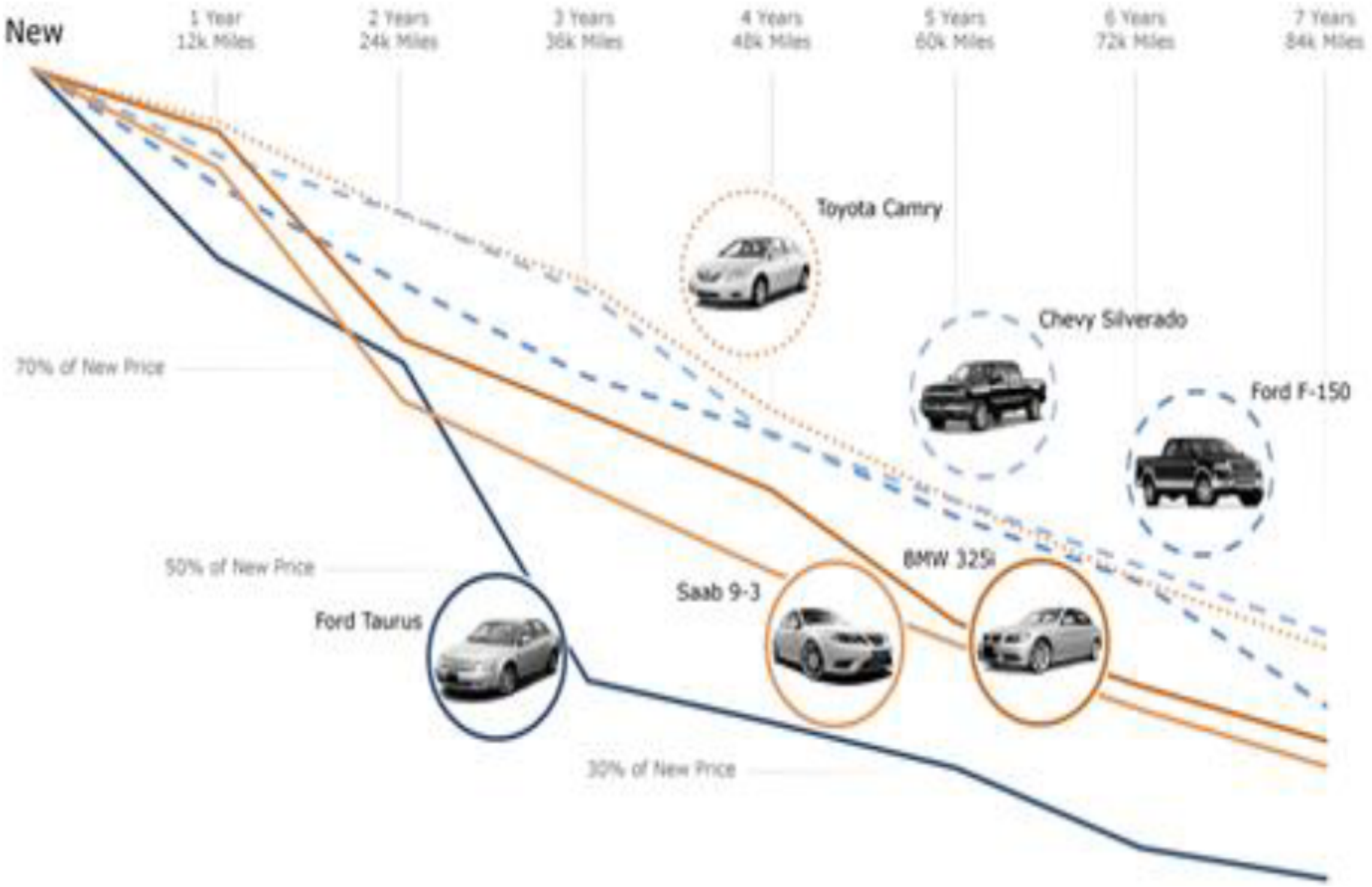
Using the regressive hypothesis *h*(X) = Y determined at above to predicate regressive results:

$$\mathcal{R} = h(\mathcal{X}) = \{y^{(i)} \mid y \in \mathcal{Y},\ i \in [1, n],\ h(x) = y\}$$

Note, in which: Output Y is a set of **continuous real** values.

Example: Used Car Prices

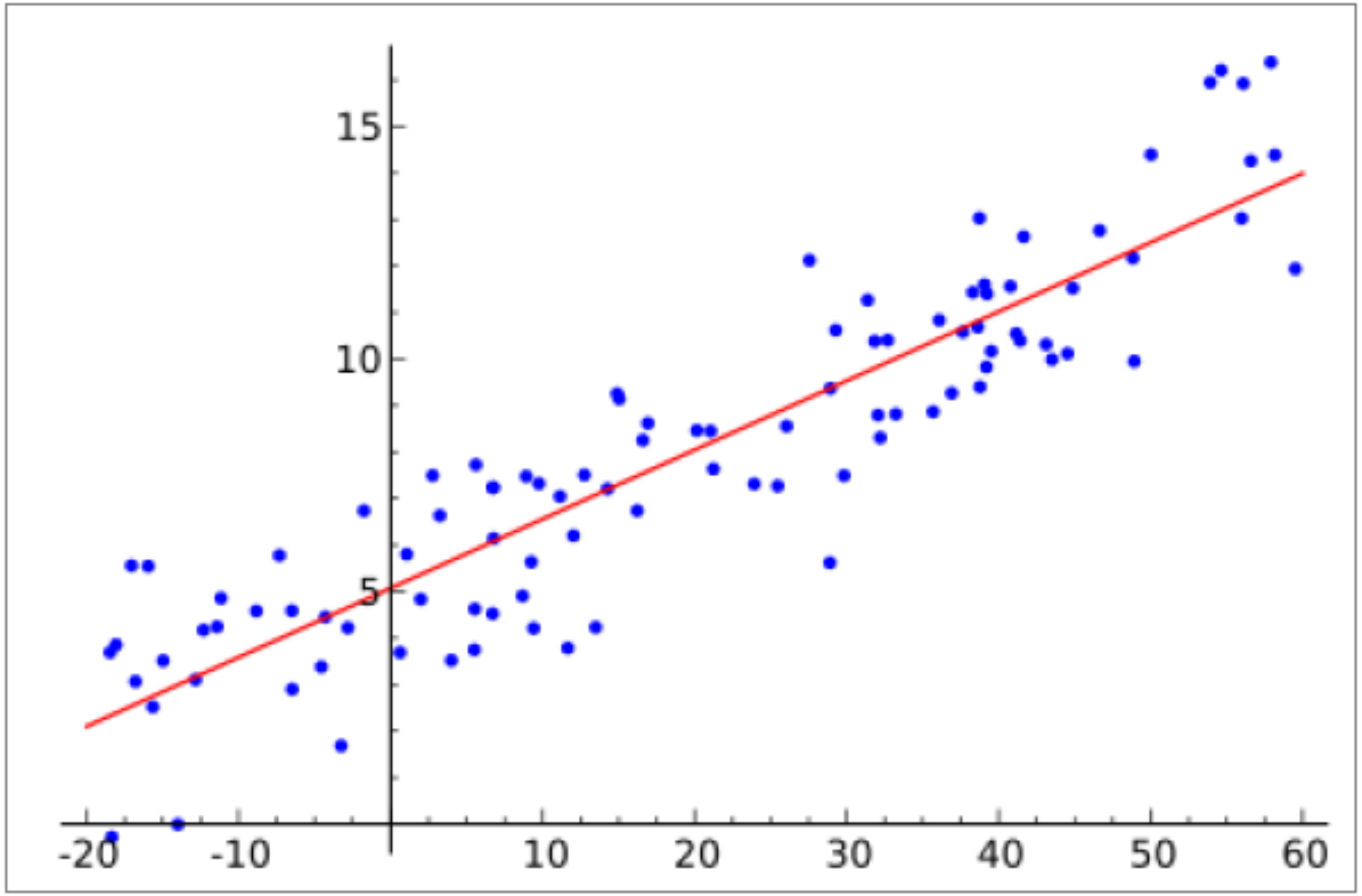
- To have a system that can predict the price of a used car.
- Inputs are the **car attributes**: brand, year, engine capacity, mileage, and other information.
- The output is **the price** of the car.



Linear Regression

- In linear regression, the observational data are modeled by a function with the following features:

  - The function is a linear combination of the model parameters;
  - The function depends on one or more independent variables.



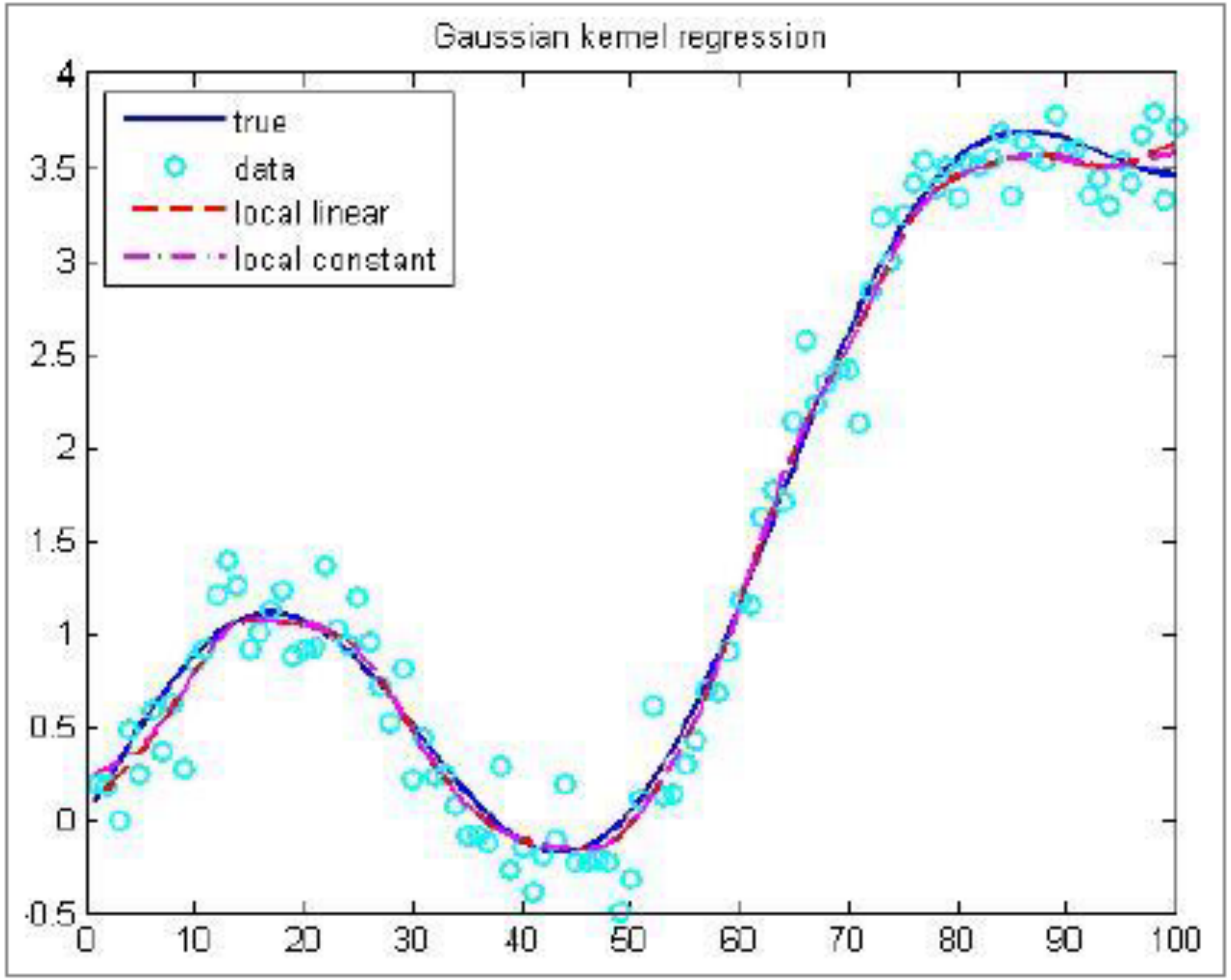
$$y(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b$$

模型表达：$y\,(x,w)=w_1x_1+\ldots+w_nx_n+b$

Nonlinear Regression

- In nonlinear regression, observational data are modeled by a function with the following features:

  - The function is a nonlinear combination of the model parameters;
  - The function depends on one or more independent variables.



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

Logistic Regression

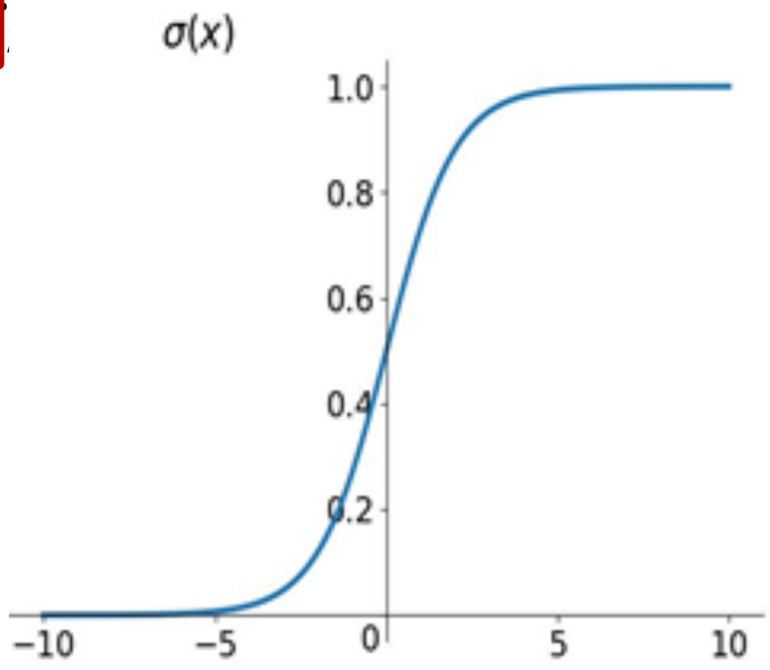
Logistic Regression uses a Sigmoid function ***g*** to map the regression function value y(x) onto [0,1], solve the **classification problem**.

Linear regression: $\boldsymbol{y}\,(\boldsymbol{x},\theta) = \theta^T \boldsymbol{x} \boxed{= \theta_0 + \theta_1 \boldsymbol{x}_1 + \ldots + \theta_n \boldsymbol{x}_n}$

want $0 \leq h_\theta(x) \leq 1$

Sigmoid function/ Logistic function : $g(\boldsymbol{y}) = \frac{1}{1+e^{-\boldsymbol{y}}}$

Logistic Regression: $h_\theta(x) = g(\theta^T x) = \boxed{\frac{1}{1+e^{-\theta^T x}}}$



Output: a real value ***y*** in [0,1]--- probability

Meaning: Tell that the unknown sample belongs to a certain class with a calculated probability ***y***.

Logistic Regression vs. Linear Regression

method	Independent variable	dependent variable	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

Logistic Regression： Tell patient that 70% chance of tumor being malignant

Typical Applications of Regression 回归的典型应用

Be widely used for **prediction** and **forecasting**. 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

  analyzing and quantifying the systematic risk of an investment.
  分析与量化投资的系统性风险。
- Economics 经济

  predicting consumption spending, fixed investment spending, the demand to hold liquid assets, and etc.
  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

Clustering 聚类

- How Clustering Works
- Major Approaches of Clustering
- Applications and Algorithms

What is Clustering

- **A longer description**
  Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

- **A shorter description**
  The process of organizing objects into groups whose members are similar in some way.

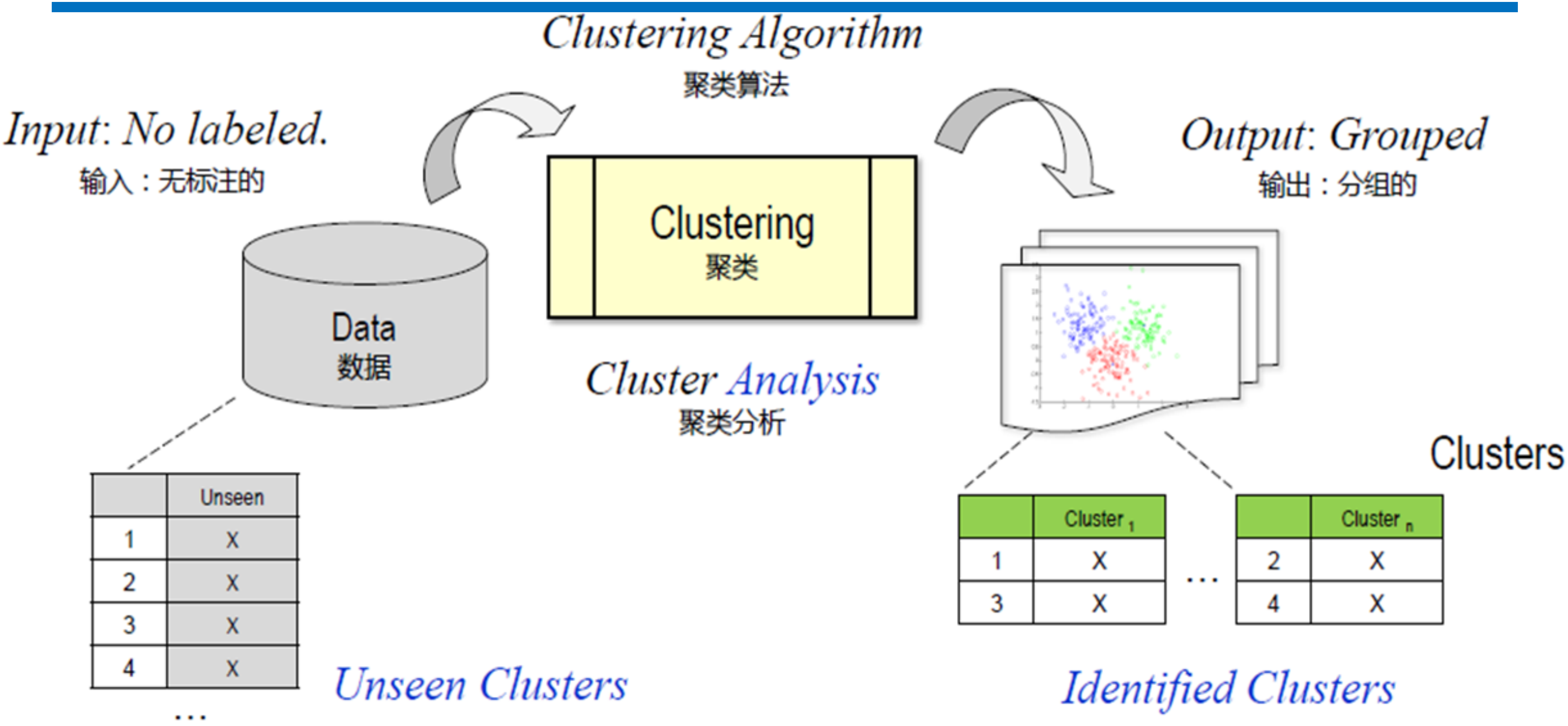
- **A very short description**
  To group data objects.

Clustering vs. Classification

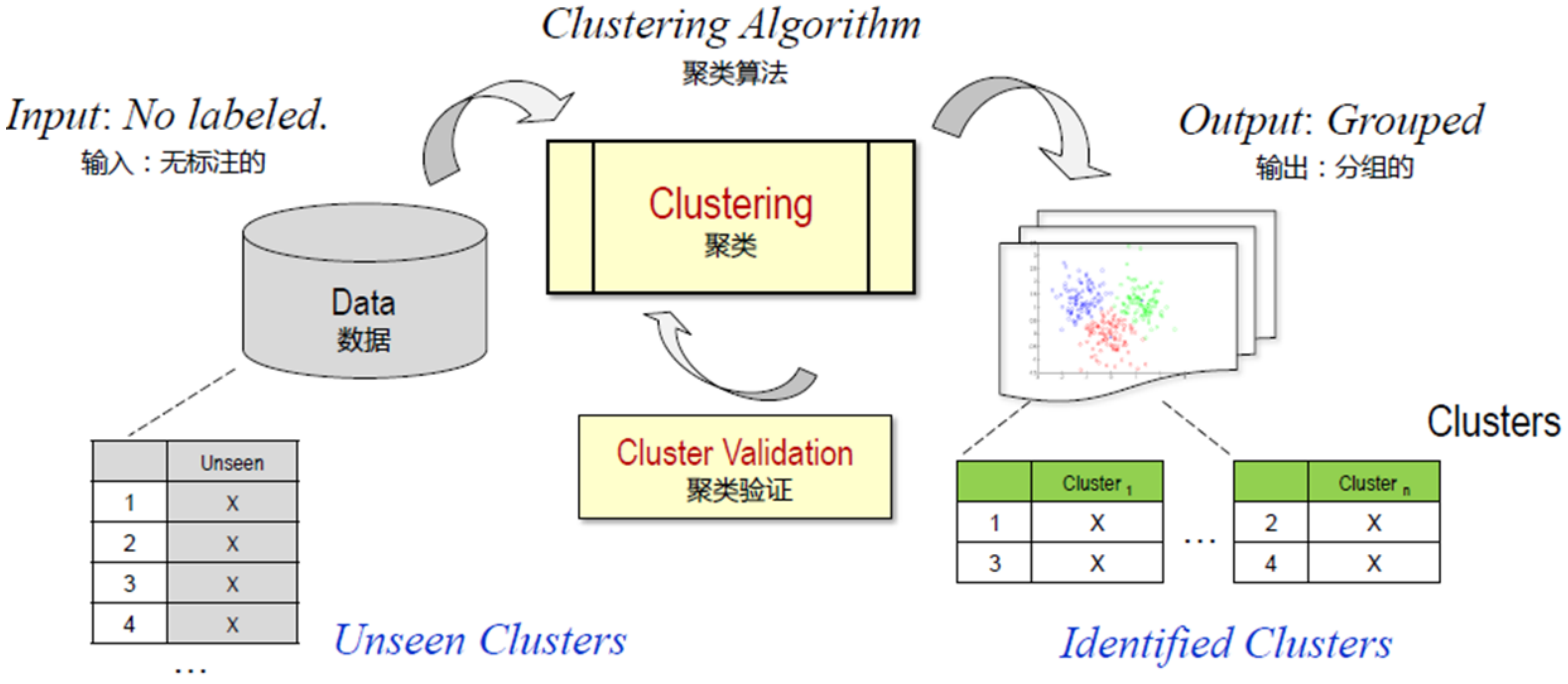
- **Similarity**
  Groups or Classes
- **Difference**
  As shown in the following table

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。

Grouping similar Input Data into Same Cluster



Two Key Steps in Clustering Procedure



A Formal Description of Clustering

Let $\mathbb{R}^n$ ($n \geqq 1$) denote a set of n-dimensional real-valued vectors, input space X is a subset of $\mathbb{R}^n$, output space Y is a set of unknown clusters, D is an unknown distribution over X×Y, then:

- Let a clustering function:

$$h: \mathrm{X} \to \mathrm{Y} \quad \text{and} \quad h \in H$$

- Clustering:
  Given a testing set of unknown clusters:

$$\mathcal{X} = \{x^{(i)} \mid x \in \mathcal{Y}, i \in [1, m]\}$$

- Using the clustering function determined at above to analyze the clustering results:

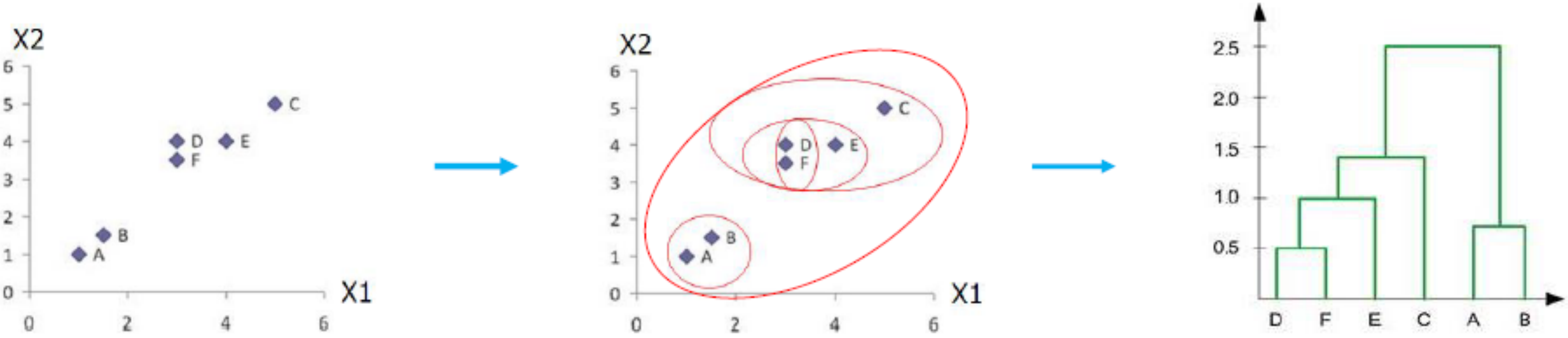
$$\mathcal{Y} = h(\mathcal{X}) = \{y^{(i)} \mid y \in \mathcal{Y}, i \in [1, n], h(x) = y\}$$

Typical Approaches of Clustering Algorithm

1) **Connectivity-based clustering**
   Also known as **hierarchical clustering**, based on the distance between objects.
2) **Centroid-based clustering**
   To find the *k* cluster centers and assign the objects to nearest cluster center.
3) **Distribution-based clustering**
   Clusters can be defined as objects belonging most likely to the same distribution.
4) **Density-based clustering**
   To group objects into one cluster if they are connected by densely populated area.。

Connectivity-based clustering

- Based on the core idea of objects being more related to nearby objects than to objects farther away.
- Creating a hierarchical decomposition of the set of data objects using some criterion.。

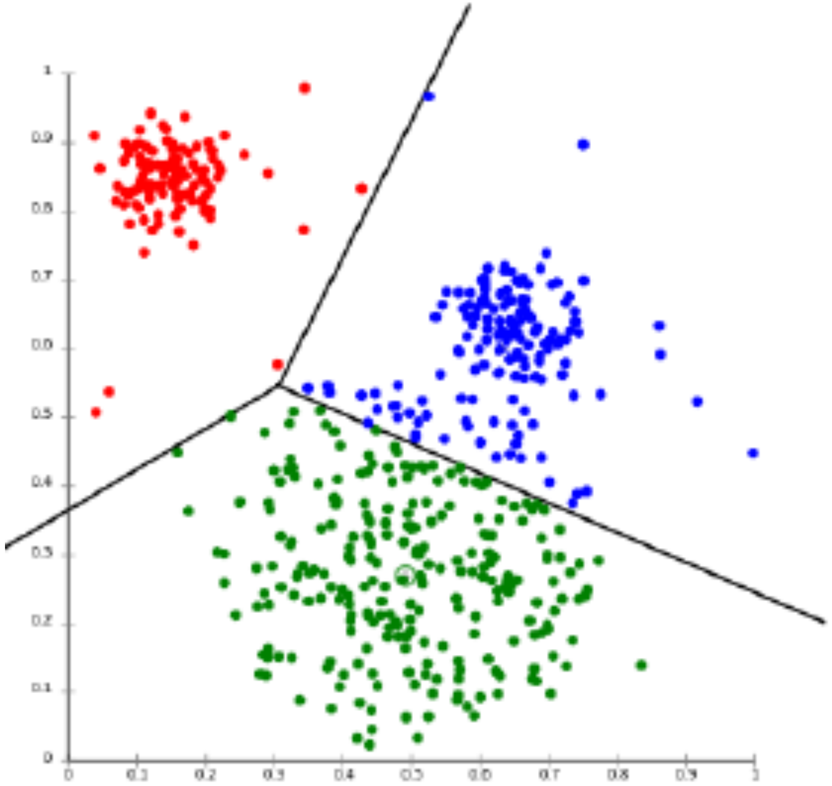


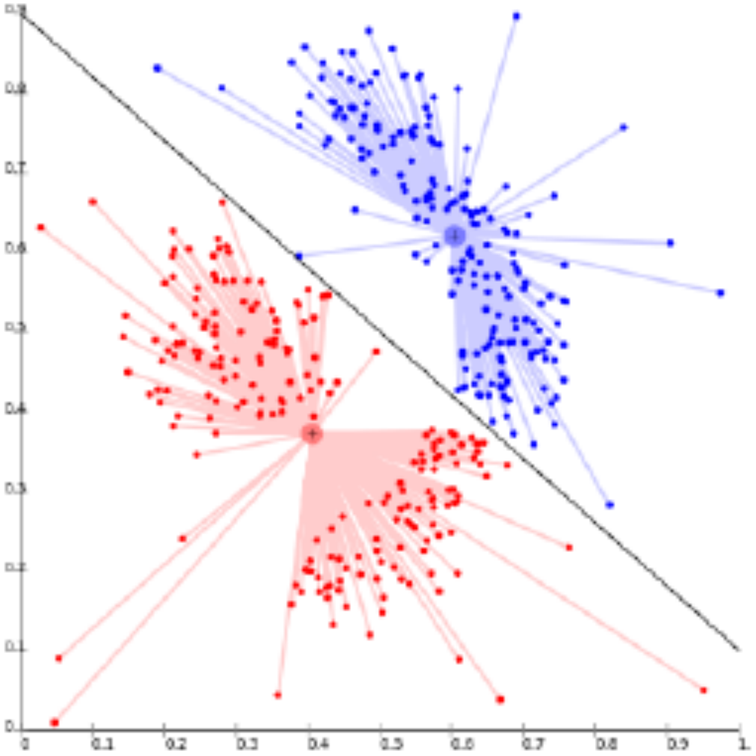
Typical algorithms: AGNES (Agglomerative NESting), DIANA (Divisive Analysis), ……

典型算法：AGNES (集聚嵌套), DIANA (分裂分析), ……

Centroid-based clustering

Constructing various partitions and then evaluating them by some criterion, e.g., minimizing the sum of square distance cost, also called Partition-based clustering.



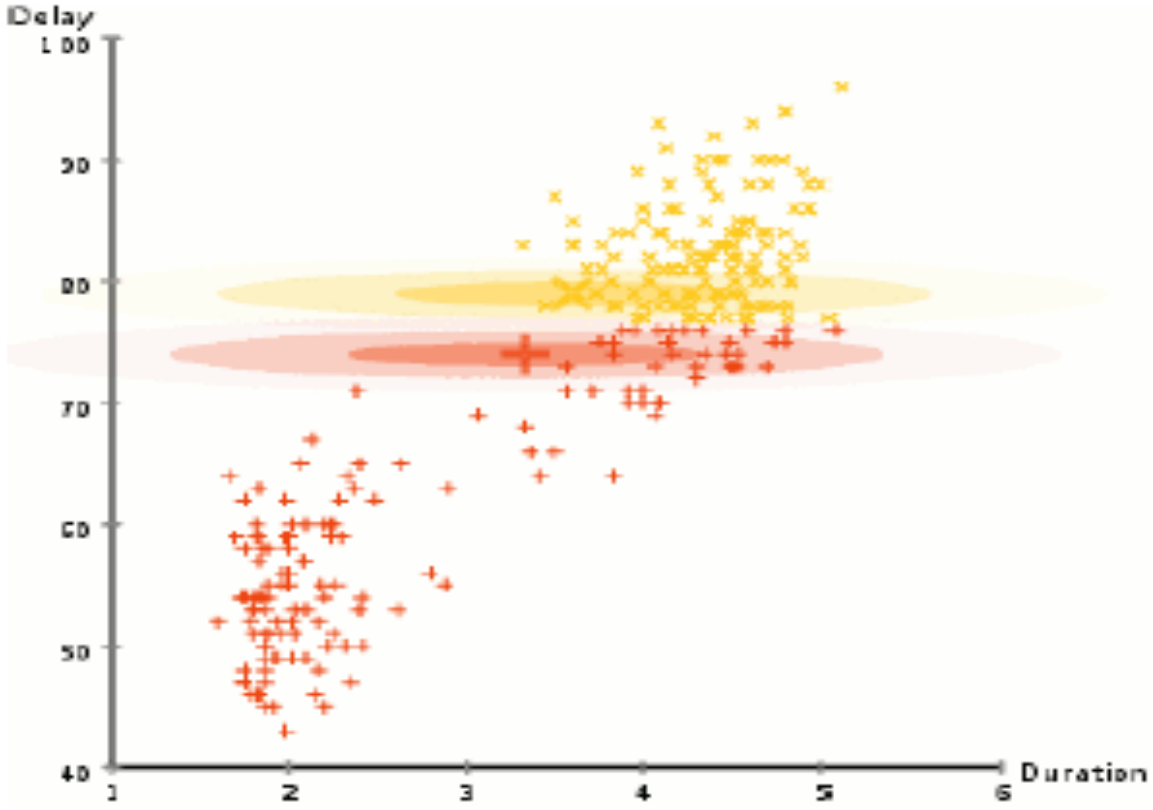


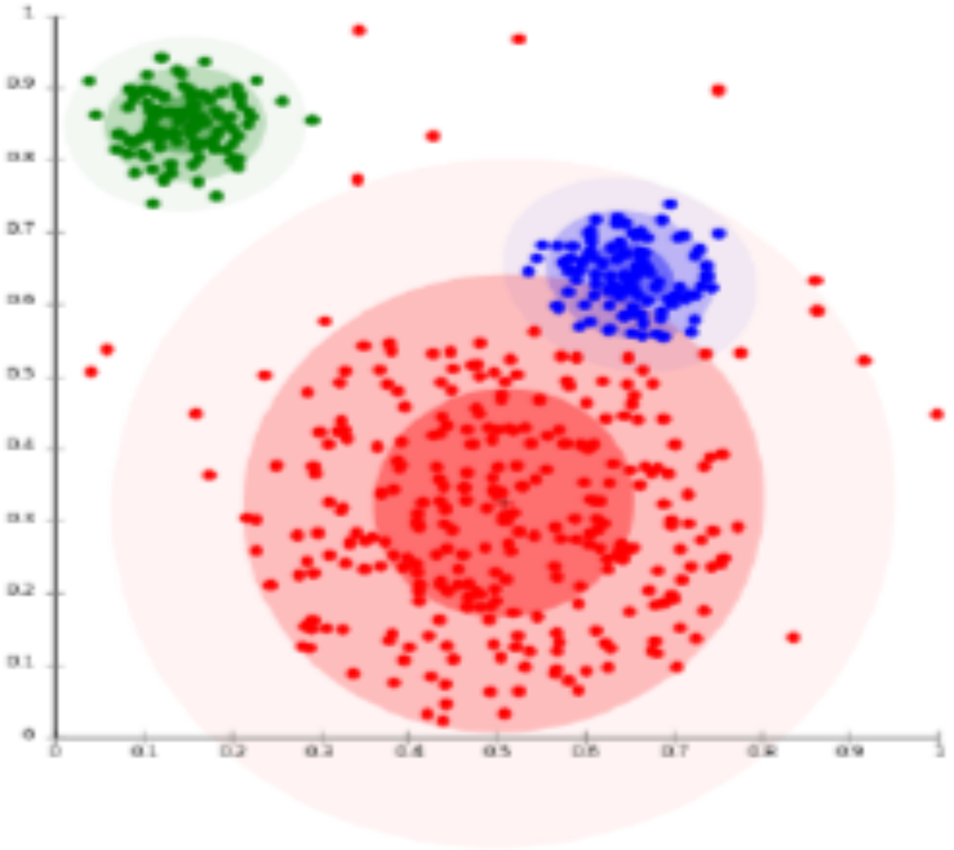
Typical algorithms: k-means, k-medoids, ……

典型算法：k-均值, k-中心点, ……

Distribution-based clustering

Clusters are modeled using statistical distributions, such as multivariate normal distributions.



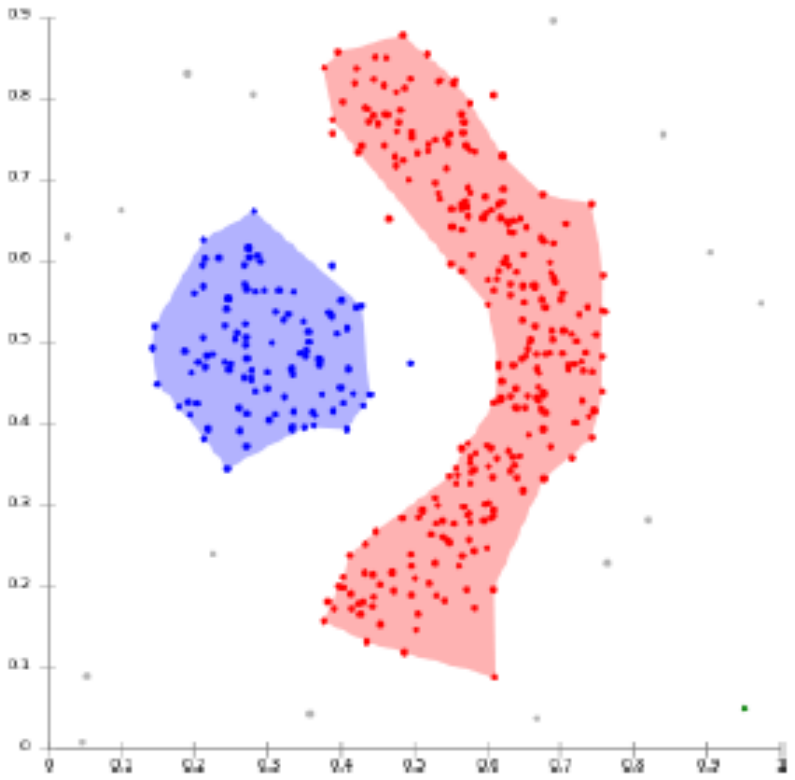


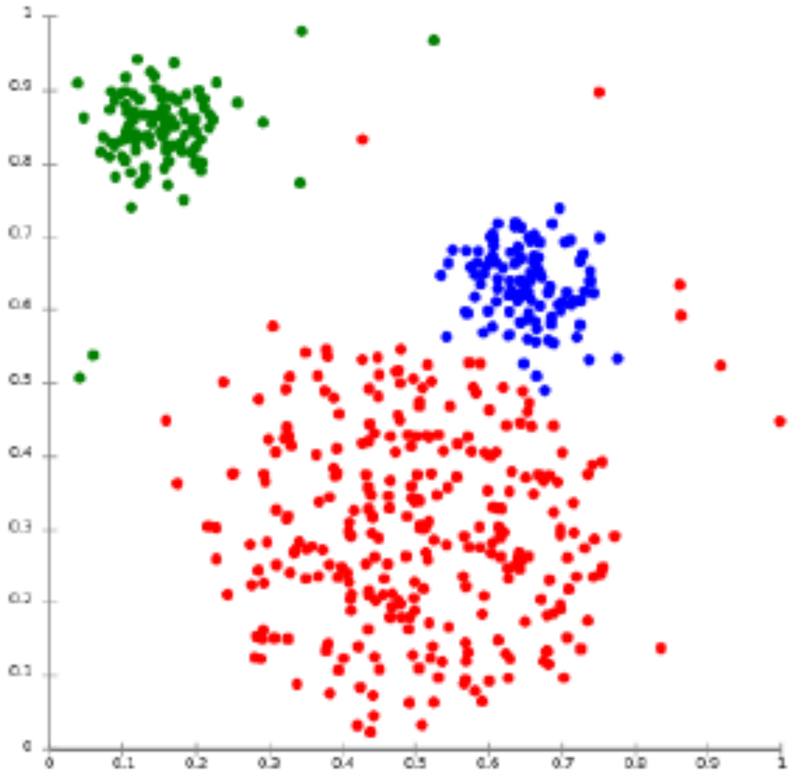
Typical algorithms: Expectation-maximization, ……

典型算法：期望最大化, ……

Density-based clustering

Clusters are defined as areas of higher density than the remainder of the data set.





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

- Cluster centers are characterized by

  1) a higher density than their neighbors,

  2) a larger distance from points with higher densities.

- The features of the clustering method are:
  - the number of clusters arises intuitively,，
  - outliers are automatically spotted and excluded,
  - clusters are recognized regardless of their shape, and space dimensionality.。

Case Study: Clustering by density peaks

Local density: 局部密度:

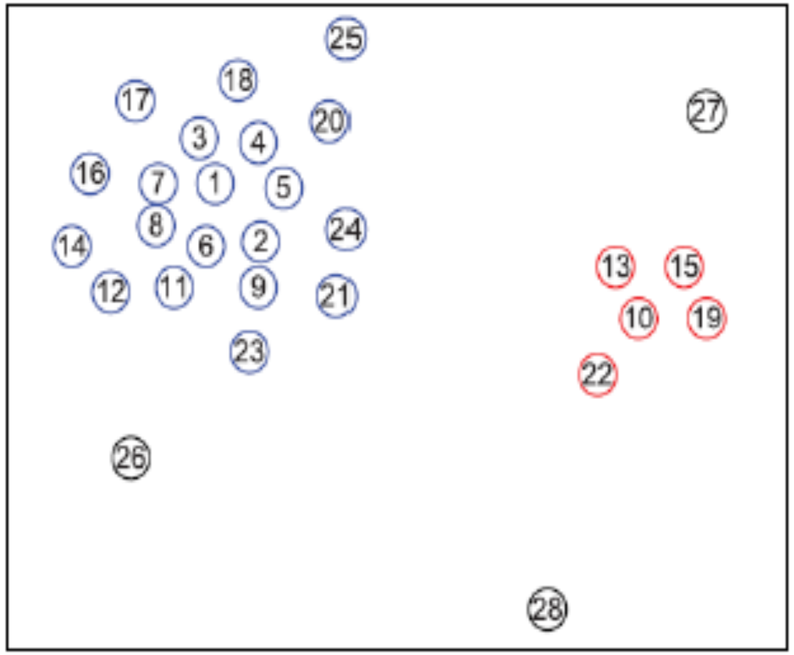
$$\rho_i = \sum_j \chi(d_{ij} - d_c) \qquad \chi(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{otherwise} \end{cases}$$

Minimum distance: 最小距离:

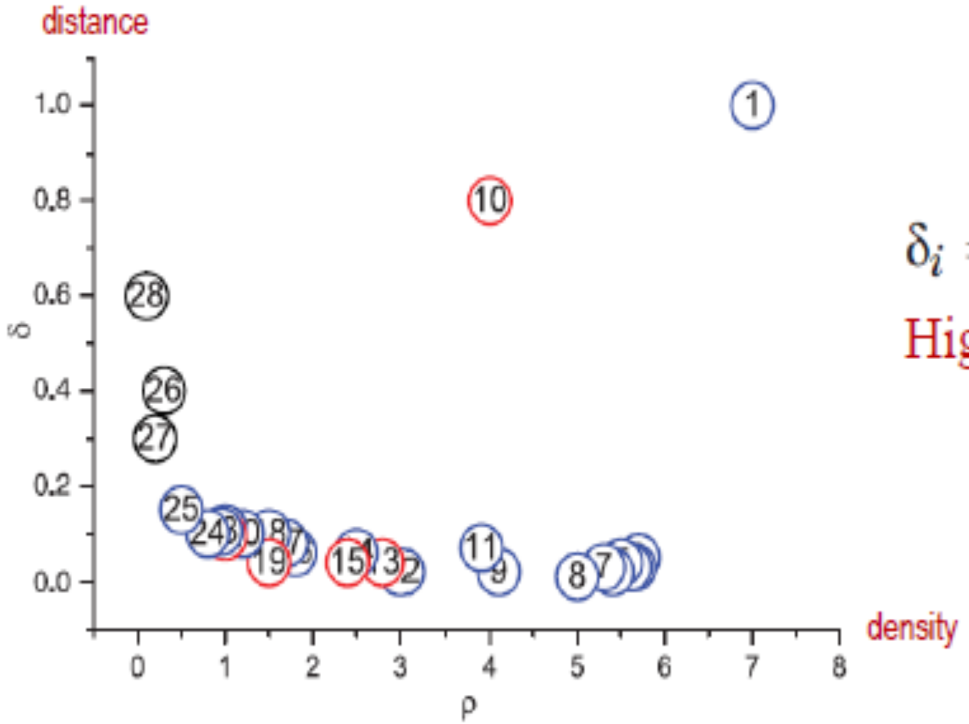
$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

where, d_{ij} : the distances between data points 数据点之间的距离

d_c : cutoff distance. 截断距离



Data (28 points) in decreasing density.
密度降排表示的数据（28个点）



$$\delta_i = \max_j (d_{ij})$$

Highest density
最高密度

Decision graph calculated local density and distance
计算局部密度和距离后的决策图

Case Study: Clustering by density peaks

Clustering analysis of the Olivetti Face Database.

Pictorial representation of the cluster assignations for the first 100 images. Faces with the same color belong to the same cluster, whereas gray images are not assigned to any cluster. Cluster centers are labeled with white circles.

Typical Applications of Clustering

- Medicine 医学
  - Medical imaging 医学影像
- Business and marketing 商务和营销
  - Grouping of customers 顾客分组
  - Grouping of shopping items 购物商品分组
- World wide web 万维网
  - Social network analysis 社交网络分析
  - Search result grouping 搜索结果分组
- Computer science 计算机科学
  - Image segmentation 图像分割
  - Recommender systems 推荐系统

Typical Algorithms of Clustering

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