Artificial Intelligence

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

Contents:

- □ Part 1. Basics
- □ Part 2. Searching
- □ Part 3. Reasoning
- □ Part 4. Planning
- □ Part 5. Learning

Contents:

- □ 9. Perspectives about Machine Learning
- □ 10. Tasks in Machine Learning
- □ 11. Paradigms in Machine Learning
- □ 12. Models in Machine Learning

Tasks in Machine Learning

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

Objectives 教学目的

- This chapter will discuss in detail about the tasks that can be solved with machine learning.
  本章将详细讨论可以通过机器学习解决的一些任务。

What are Learning Tasks 什么是学习任务

- □ The learning tasks are used to denote the general problems that can be solved by learning with desired output.
  学习任务用于表示可以用机器学习解决的基本问题。

Why Study Learning Tasks 为什么要研究学习任务

- □ Various types of problems arising in applications:
  应用中会产生各种类型的问题：
  - ■ computer vision, 计算机视觉，
  - ■ pattern recognition, 模式识别，
  - ■ natural language processing, 自然语言处理，
  - ■ etc. 等等。

Typical Tasks in Machine Learning 机器学习中的典型任务

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

10. Tasks in Machine Learning

Contents:

- □ 10.1. Classification
- □ 10.2. Regression
- □ 10.3. Clustering
- □ 10.4. Ranking
- □ 10.5. Dimensionality Reduction

Classification

AI

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

What is Classification 什么是分类

- □ A longer description 较长描述

  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.

  分类是基于包含已知类别成员观测值的训练数据集、来辨识新的观测值属于哪一组类别的任务。

- □ A shorter description 较短描述

  To resolve such problems where the output is divided into two or more categories.

  解决输出被分为两个或多个类别的问题。

- □ A very short description 极简描述

  Assign a category to each item.

  为每个项指定一个类别。

10.1. Classification

Contents:

- □ 10.1.1. How Classification Works
- □ 10.1.2. Linear and Nonlinear
- □ 10.1.3. Dimensions and Classes
- □ 10.1.4. Applications and Algorithms

Classifier 分类器

□ About classifier 关于分类器

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

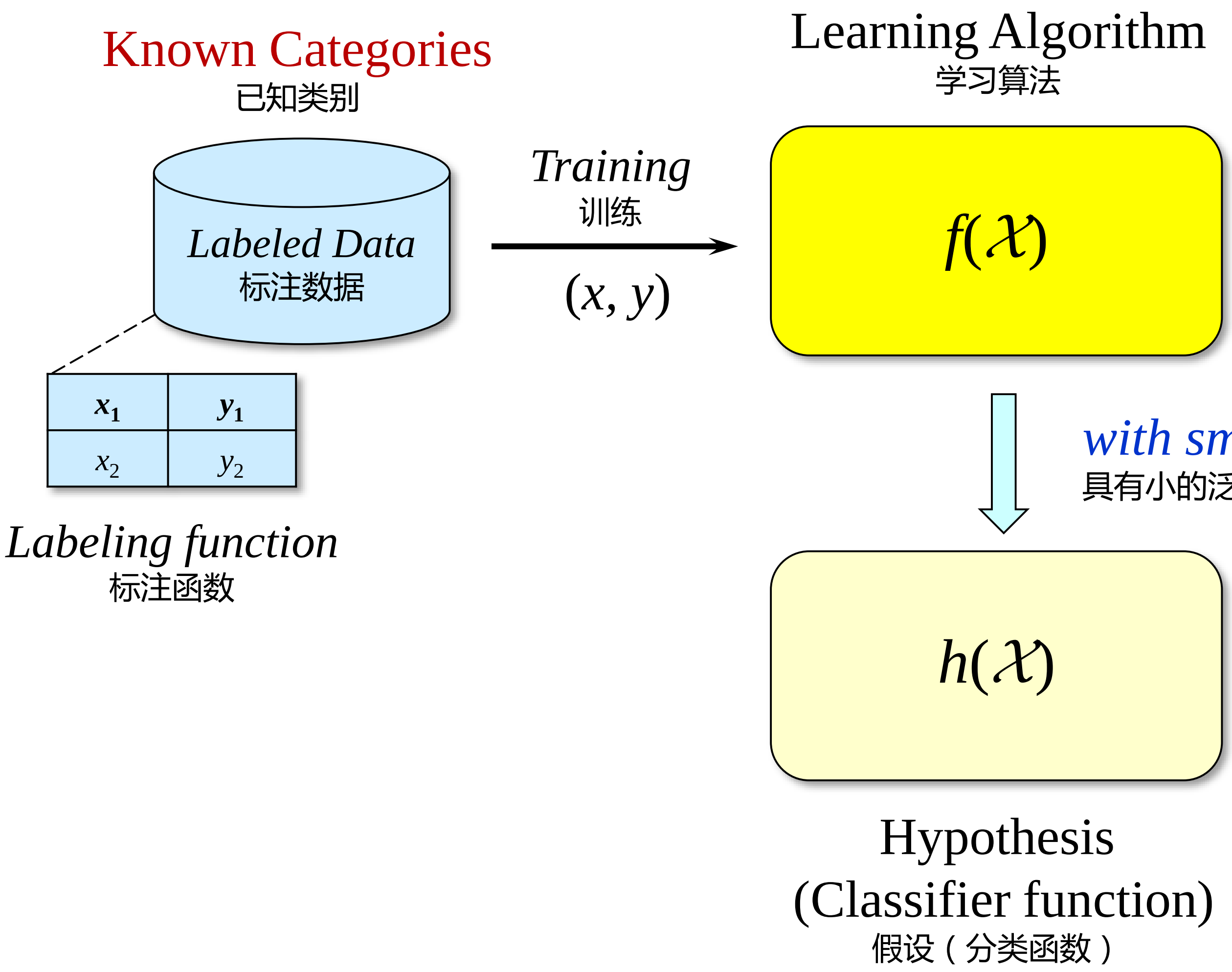
一种实现分类、尤其是构成一种具体实现的算法，被称为一个分类器。

□ About classifier function 关于分类器函数

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

“分类器”这个术语有时还指的是由分类算法所实现的数学函数，它将输入数据映射为一个类别。

Classification: Training 分类：训练



Known Categories
已知类别

Labeled Data
标注数据

$\mathbf{x_1}$	$\mathbf{y_1}$
x_2	y_2

Labeling function
标注函数

Training
训练

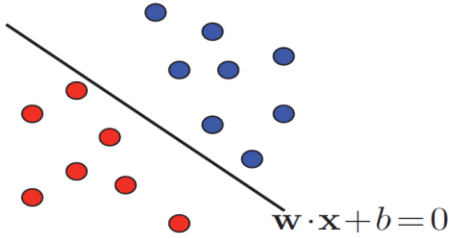
(x, y)

Learning Algorithm
学习算法

$f(\mathcal{X})$

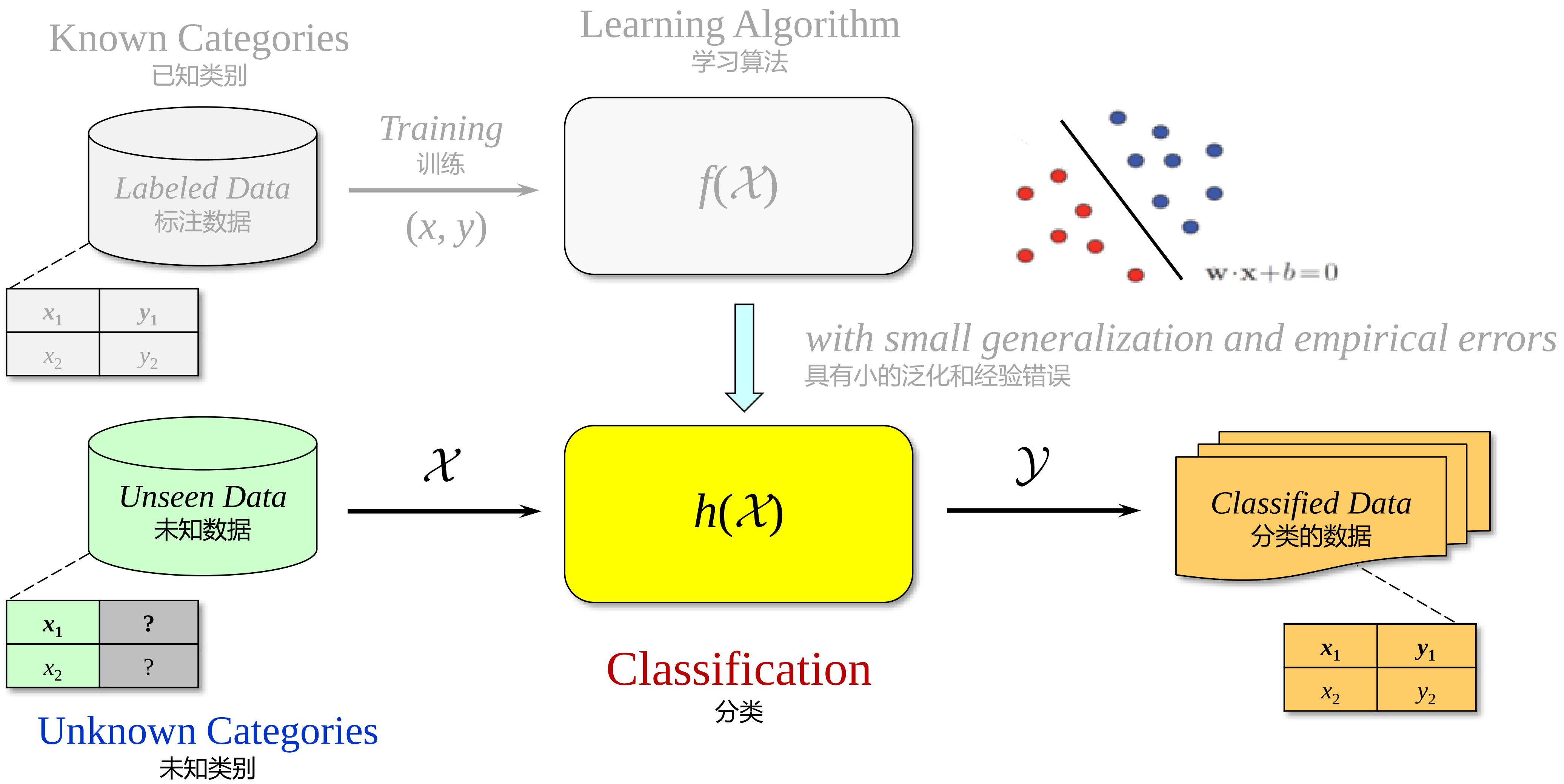
$h(\mathcal{X})$

Hypothesis
(Classifier function)
假设（分类函数）



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

Classification: Testing 分类：实测



Known Categories
已知类别

Labeled Data
标注数据

x_1	y_1
x_2	y_2

Training
训练

(x, y)

Learning Algorithm
学习算法

$f(\mathcal{X})$

$\mathbf{w} \cdot \mathbf{x} + b = 0$

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	?

$\mathcal{X}$

$h(\mathcal{X})$

$\mathcal{Y}$

Classified Data
分类的数据

x_1	y_1
x_2	y_2

Classification
分类

Unknown Categories
未知类别

A Formal Description of Classification 一种分类的形式化描述

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

设$\mathbb{R}^n$ ($n \geqq 1$)表示一个n维实值向量集合，输入空间$\mathcal{X}$是$\mathbb{R}^n$的一个子集，输出空间$\mathcal{Y}$是一组类别，D是$\mathcal{X} \times \mathcal{Y}$的一个未知分布，则：

☐ Let target labeling function: 设目标标注函数

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

☐ Training set (Labeled training sample set): 训练集（标注的训练样本集）

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

☐ Classification algorithm: 分类算法

Let a hypothesis set H are the mapping $\mathcal{X}$ to $\mathcal{Y}$, to determine a hypothesis (classifier function):

设一个假设函数集H是$\mathcal{X}$到$\mathcal{Y}$的映射，来决定一个假设（分类器函数）：

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

with small generalization error: 具有小的泛化错误

$$\mathrm{R}(h) = \Pr_x[h(x) \neq f(x)]$$

A Formal Description of Classification 一种分类的形式化描述

- □ Classification: 分类

  Given a testing data set of unknown categories:

  给定一个未知类别的实测数据集：

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

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

使用前面训练好的分类函数$h(\mathcal{X}) = \mathcal{Y}$来预测分类结果：

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

where 其中

$\mathcal{Y}$ is the set of known categories.

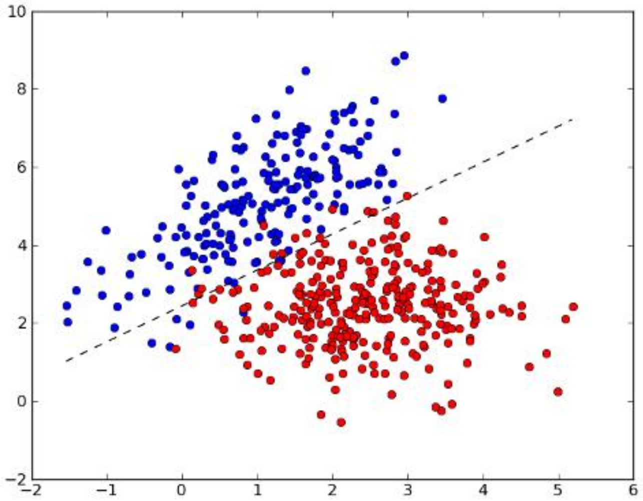
Y是该已知类别的集合。

Contents:

- □ 10.1.1. How Classification Works
- □ 10.1.2. Linear and Nonlinear
- □ 10.1.3. Dimensions and Classes
- □ 10.1.4. Applications and Algorithms

Linear Classification 线性分类

- □ Linear Classification is doing classification by a linear classifier.
  线性分类是通过线性分类器来进行分类。



- □ A linear classifier is 一个线性分类器是
  - ■ a linear discriminant function with a linear decision boundary.
    具有一个线性决策边界的线性判别函数。

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

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

where, 其中

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

w表示行向量、称为权向量，

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

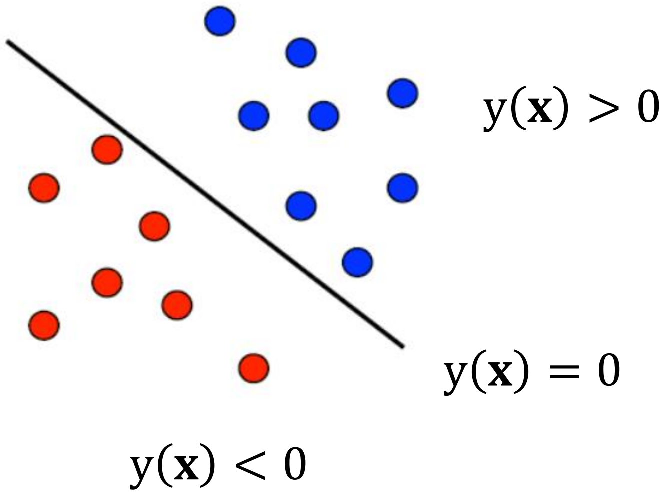
x denotes a column vector,

x表示列向量，

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

b denotes a bias.

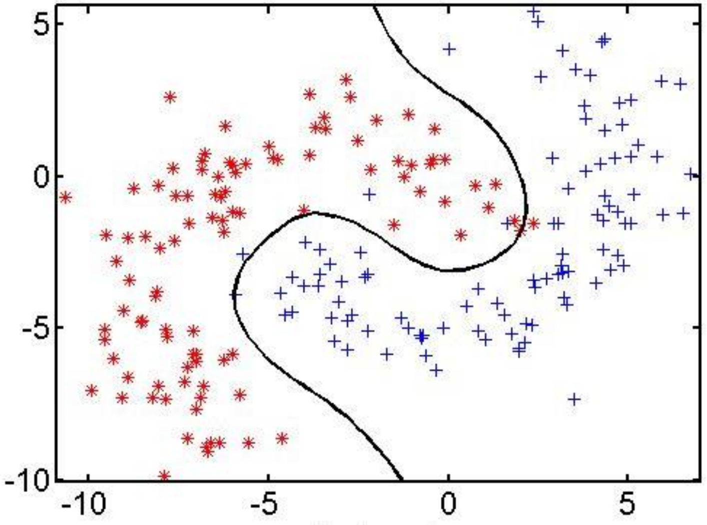
*b*表示偏差。


$\mathrm{y}(\mathbf{x}) > 0$
$\mathrm{y}(\mathbf{x}) = 0$
$\mathrm{y}(\mathbf{x}) < 0$

Nonlinear Classification 非线性分类

□ Nonlinear Classification is doing classification by a nonlinear classifiers.
非线性分类是通过一个非线性分类器来进行分类。

□ A nonlinear classifiers have 一个非线性分类器具有
nonlinear decision boundaries, and possibly discontinous decision boundaries.
若干非线性决定边界，并且可能是非连续决定边界。



E.g., a nonlinear classifier in SVM is a nonlinear kernel function.
例如，在SVM中的非线性分类器是一个非线性核函数。

10.1. Classification

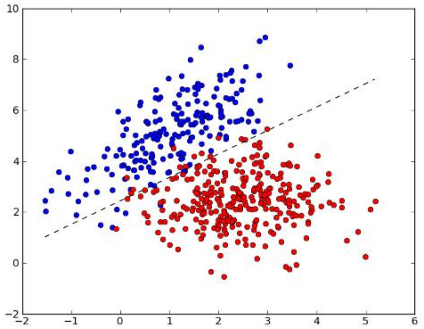
Contents:

- □ 10.1.1. How Classification Works
- □ 10.1.2. Linear and Nonlinear
- □ 10.1.3. Dimensions and Classes
- □ 10.1.4. Applications and Algorithms

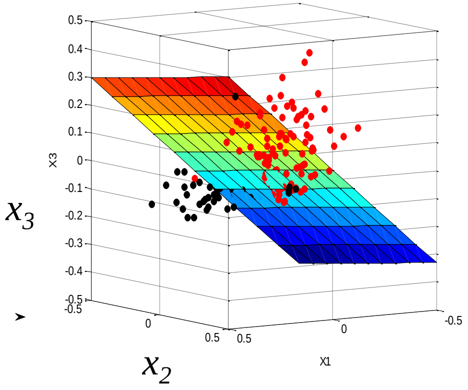
Dimensions 维度

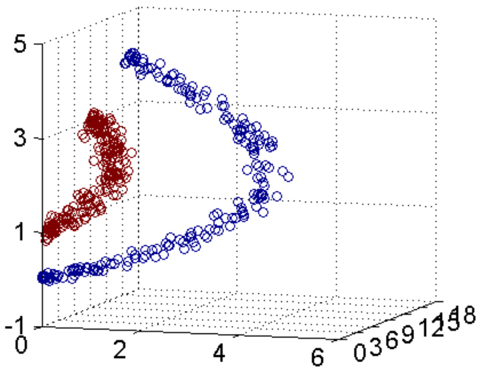
□ If the problem space is n dimensional then its linear classifier is n-1 dimensional hyper-plane. E.g.,

如果问题空间的维度为n，则它的线性分类器的维度为n-1的超平面。例如：



2-dimensions
2维





3-dimensions
3维

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

2维空间中，该超平面为一条线

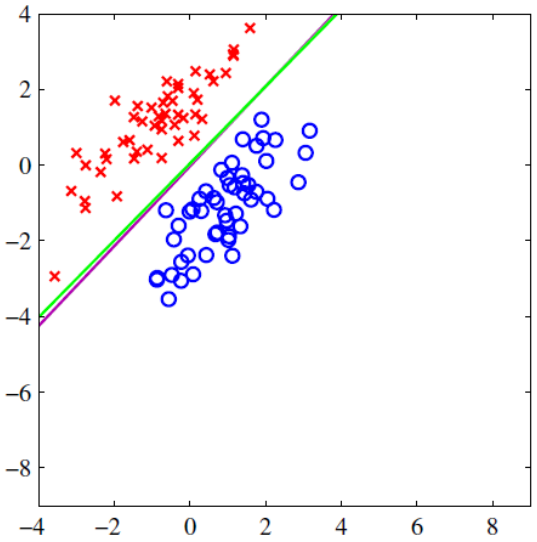
in 3-dimensions, the hyper-plane is a plane

3维空间中，该超平面为一个平面

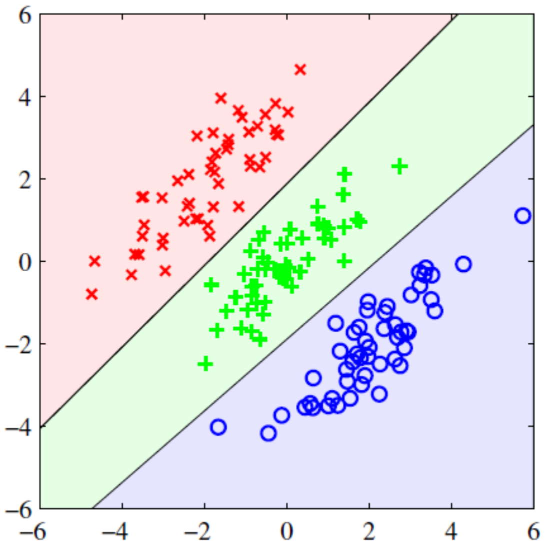
Classes 类别

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

- 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

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

  这是一个多元分类器，由Softmax函数来实现。

- □ 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).

  Softmax函数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$\sigma(\mathbf{x})$ (范围0到1，和为1)。

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

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

  在概率论中，Softmax函数的输出可以被用来表示一个类别分布，即，一个涵盖K个不同结果的概率分布。

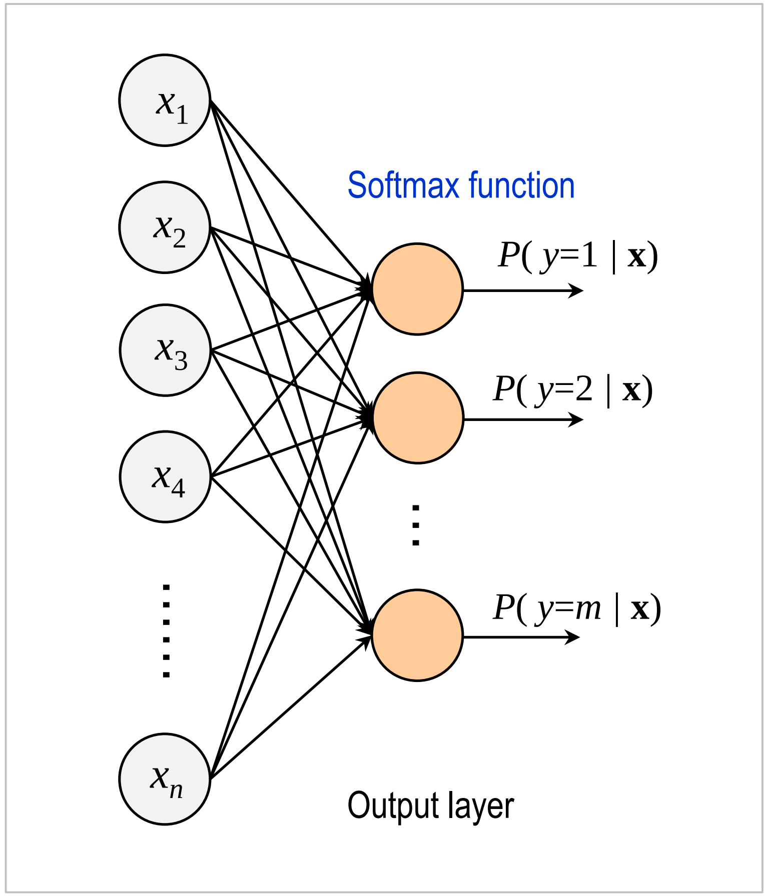
$$P(y = j \mid \mathbf{x}) = \frac{e^{\mathbf{x}^{\mathrm{T}}\mathbf{w}_j}}{\sum_{k-1}^{K} e^{\mathbf{x}^{\mathrm{T}}\mathbf{w}_k}}$$

Case Study: Softmax Classifier Softmax分类器

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

Softmax函数已经被用于各种多元分类方法，例如：

- multinomial logistic regression,
  多项式逻辑回归，
- multiclass linear discriminant analysis,
  多元线性判别分析，
- naive Bayes classifiers,
  朴素贝叶斯分类器，
- artificial neural networks (ANN),
  人工神经网络 (ANN),
- reinforcement learning.
  强化学习。



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

Softmax函数在ANN的最后一层用于多元分类

10.1. Classification

Contents:

- □ 10.1.1. How Classification Works
- □ 10.1.2. Linear and Nonlinear
- □ 10.1.3. Dimensions and Classes
- □ 10.1.4. Applications and Algorithms

Typical Applications of Classification 分类的典型应用

□ 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	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	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	Softmax
□ Support vector machine (SVM)	支撑向量机 (SVM)

Thank you for your attention!

AI

10. Tasks in Machine Learning

Contents:

- □ 10.1. Classification
- □ 10.2. Regression
- □ 10.3. Clustering
- □ 10.4. Ranking
- □ 10.5. Dimensionality Reduction

Regression

AI

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

What is Regression 什么是回归

- □ A longer description 较长描述

  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.

  回归分析是估计变量间关系的统计过程。它包含对多变量进行建模与分析的许多技术，其焦点是某个自变量与一个或多个因变量之间的关系。

- □ A shorter description 较短描述

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

  要解决输出是真实连续值的问题。

- □ A very short description 极简描述

  Predict a real value for each item.

  预测每个项的真实值。

Regression vs. Classification 回归与分类

□ Similarity 相似性

Need training processing 需要训练过程

□ Difference 差异性

As shown in the following table 如下表所示

	Regression 回归	Classification 分类
Difference 差异性	Output is a real continuous value. 输出是**一个真实连续值。**	Output is a discrete categories. 输出是**一个离散的类别。**

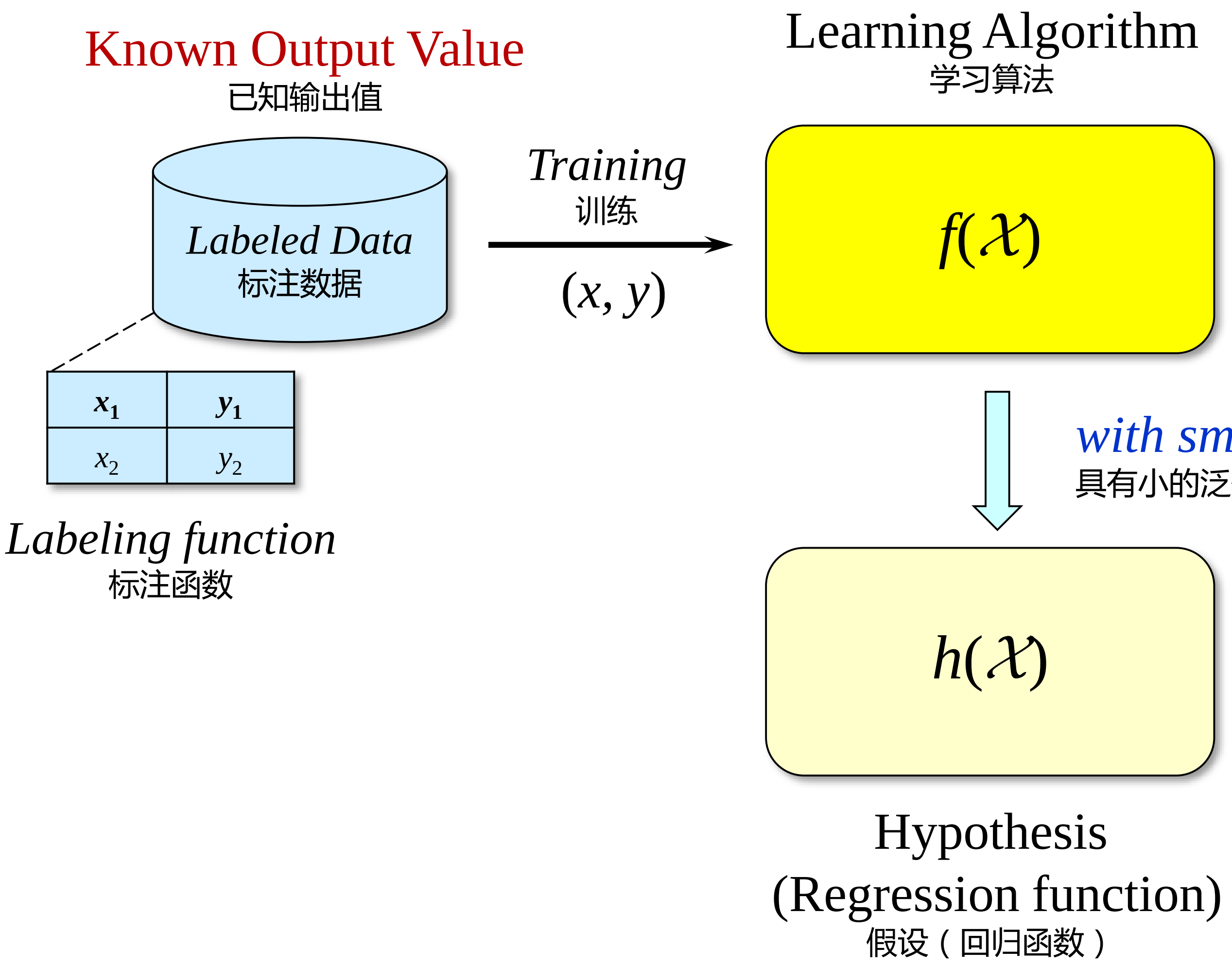
Example
举例
	➢ *Used-car price*
	二手车价格
	➢ *Tomorrow's stock price*
	明天的股票价格
		➢ {*sunny, cloudy, rainy*}
		➢ {0, 1, 2, ..., 9}

10.2. Regression

Contents:

- □ 10.2.1. How Regression Works
- □ 10.2.2. Linear and Nonlinear
- □ 10.2.3. Applications and Algorithms

Regression: Training 回归：训练



Known Output Value
已知输出值

Labeled Data
标注数据

$\mathbf{x_1}$	$\mathbf{y_1}$
x_2	y_2

Labeling function
标注函数

Training
训练

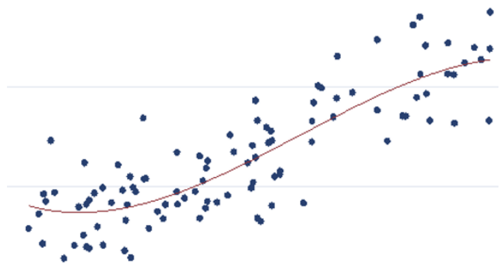
(x, y)

Learning Algorithm
学习算法

$f(\mathcal{X})$

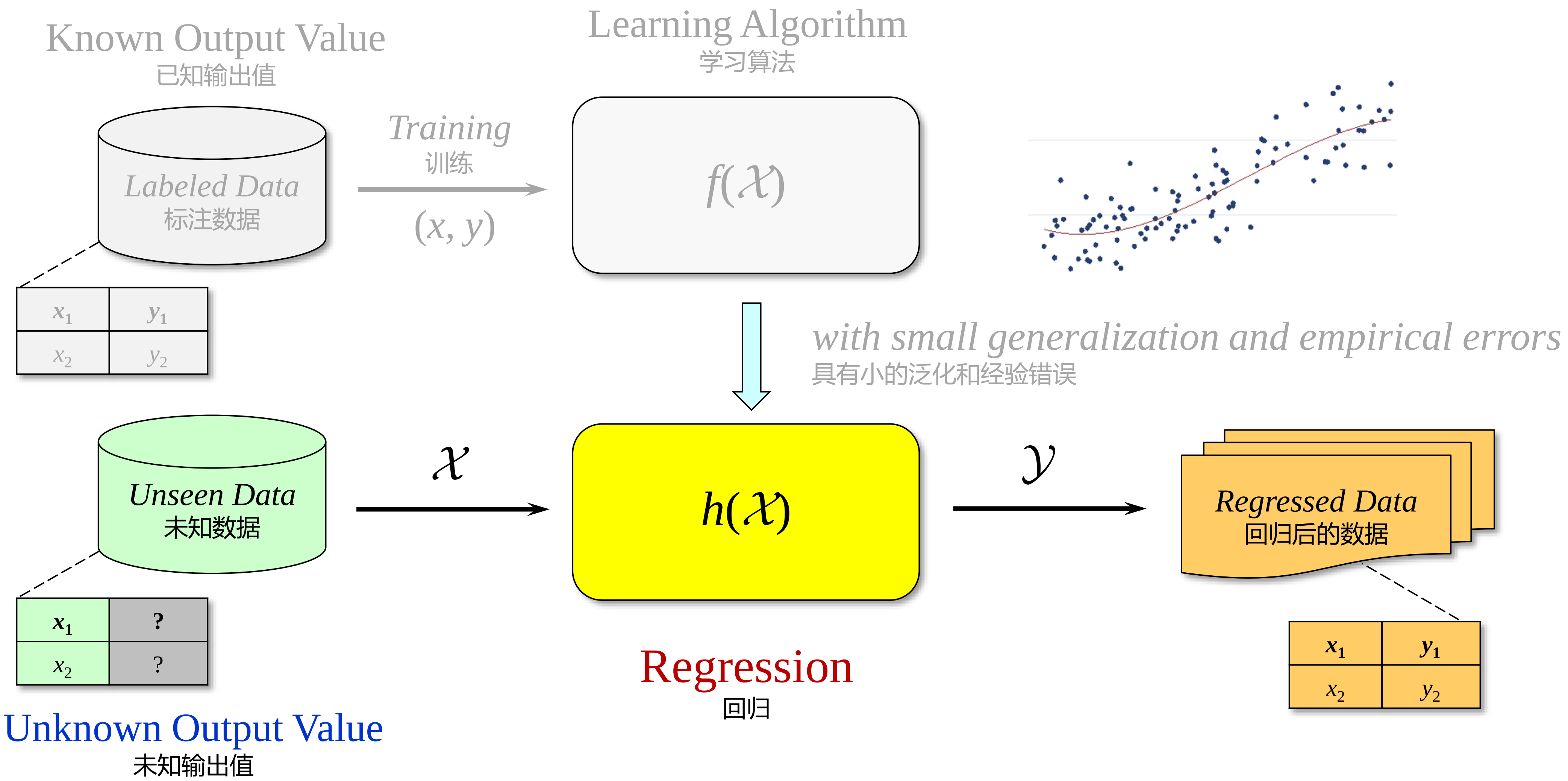
$h(\mathcal{X})$

Hypothesis
(Regression function)
假设（回归函数）



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

Regression: Testing 回归：实测



Known Output Value
已知输出值

Labeled Data
标注数据

x_1	y_1
x_2	y_2

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

Unseen Data
未知数据

$\mathbf{x_1}$	**?**
x_2	?

Unknown Output Value
未知输出值

$\mathcal{X}$

$h(\mathcal{X})$

Regression
回归

$\mathcal{Y}$

Regressed Data
回归后的数据

$\mathbf{x_1}$	$\mathbf{y_1}$
x_2	y_2

A Formal Description of Regression 一种回归的形式化描述

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 $\mathcal{X}$ is a subset of $\mathbb{R}^n$, output space $\mathcal{Y}$ is a set of real numbers $\mathbb{R}_+$, D is an unknown distribution over $\mathcal{X} \times \mathcal{Y}$, then:

设$\mathbb{R}^n$ ($n \geqq 1$)为n维实值向量集， $\mathbb{R}_+$是非负实数集，输入空间$\mathcal{X}$是$\mathbb{R}^n$的子集，输出空间$\mathcal{Y}$是实数集$\mathbb{R}_+$， D是$\mathcal{X} \times \mathcal{Y}$的未知分布，则：

□ Let target labeling function: 设目标标注函数

$$f : \mathcal{X} \to \mathcal{Y}$$

□ Training set (Labeled training sample set): 训练集（标注的训练样本集）

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

□ Regression algorithm: 回归算法

Given hypothesis set H, to determine a hypothesis (regressive function)

给定假设集H，来决定一个假设（回归函数）：

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

With small generalization error $R(h)$: 具有小的泛化错误

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

A Formal Description of Regression 一种回归的形式化描述

- □ Regression 回归

Given a testing data set of unknown output:

给定一个未知输出的实测数据集：

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

Using the regressive hypothesis $h(\mathcal{X}) = \mathcal{Y}$ determined at above to predicate regressive results:

使用前面训练好的回归函数$h(\mathcal{X}) = \mathcal{Y}$来预测回归结果：

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

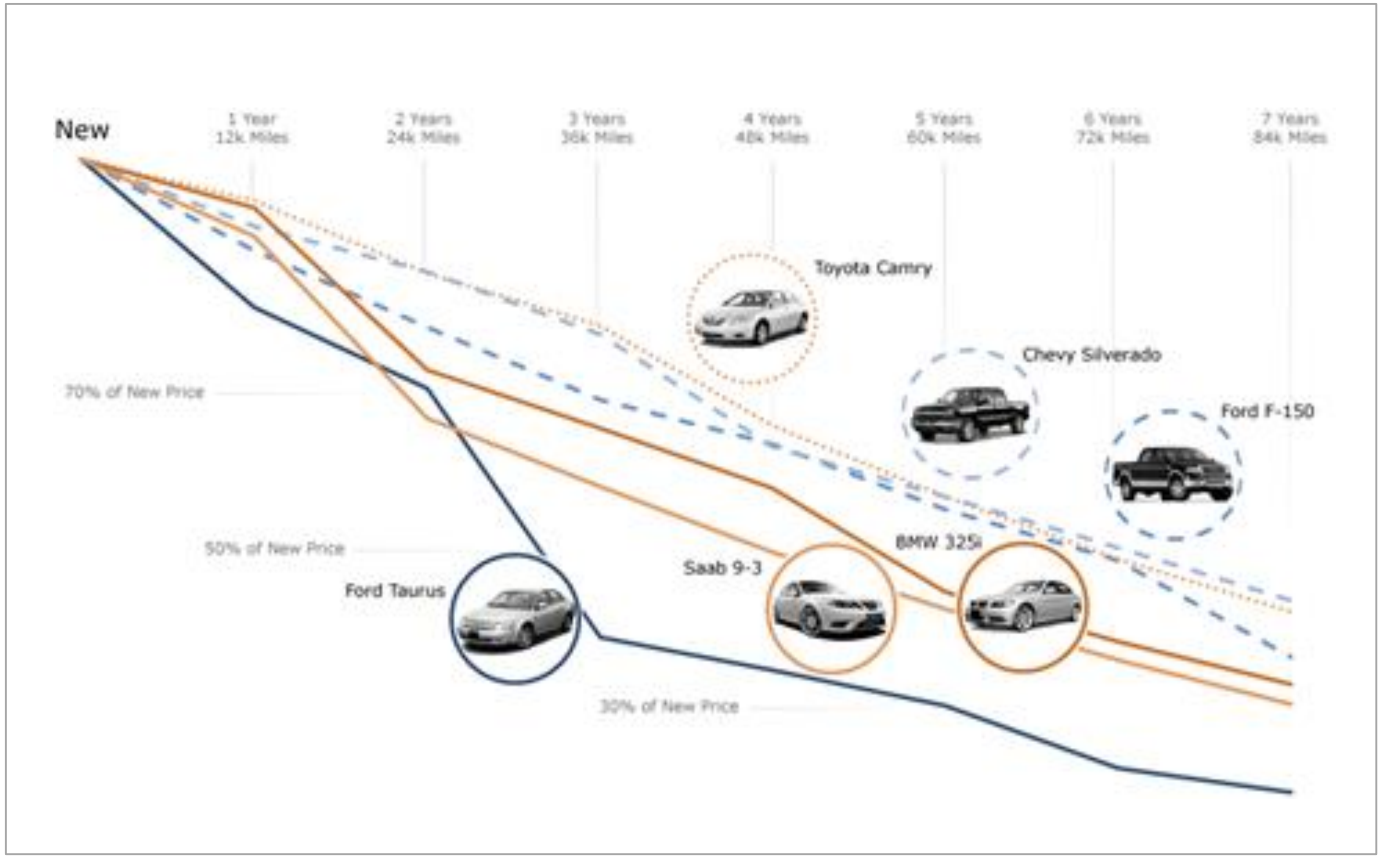
Note, in which: 注意，其中

$\mathcal{Y}$ is a set of real continues numbers.

$\mathcal{Y}$是一个真实连续数值的集合。

Example: Used Car Prices 二手车价格

- □ To have a system that can predict the price of a used car.
  构建一个预测二手车价格的系统。
- □ Inputs are the car attributes: brand, year, engine capacity, mileage, and other information.
  输入是车的属性：品牌、年式、引擎功率、里程、以及其它信息。
- □ The output is the price of the car.
  输出是车的价格。



Used car prices
二手车价格

10.2. Regression

Contents:

- □ 10.2.1. How Regression Works
- □ 10.2.2. Linear and Nonlinear
- □ 10.2.3. Applications and Algorithms

Linear Regression 线性回归

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

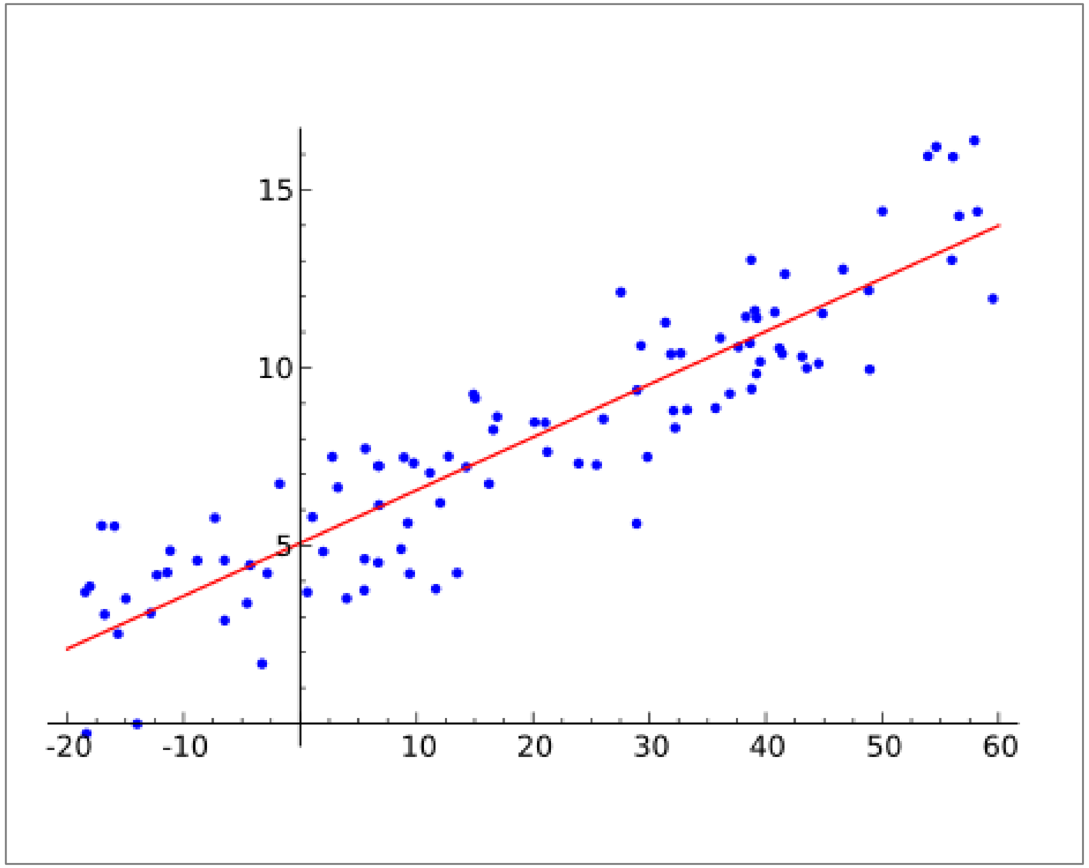
  线性回归中，采用具有如下特征的函数对观测数据进行建模：

  The function is a linear combination of the model parameters;

  该函数是模型参数的线性组合；

  The function depends on one or more independent variables.

  该函数取决于一个或多个独立变量。



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

Nonlinear Regression 非线性回归

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

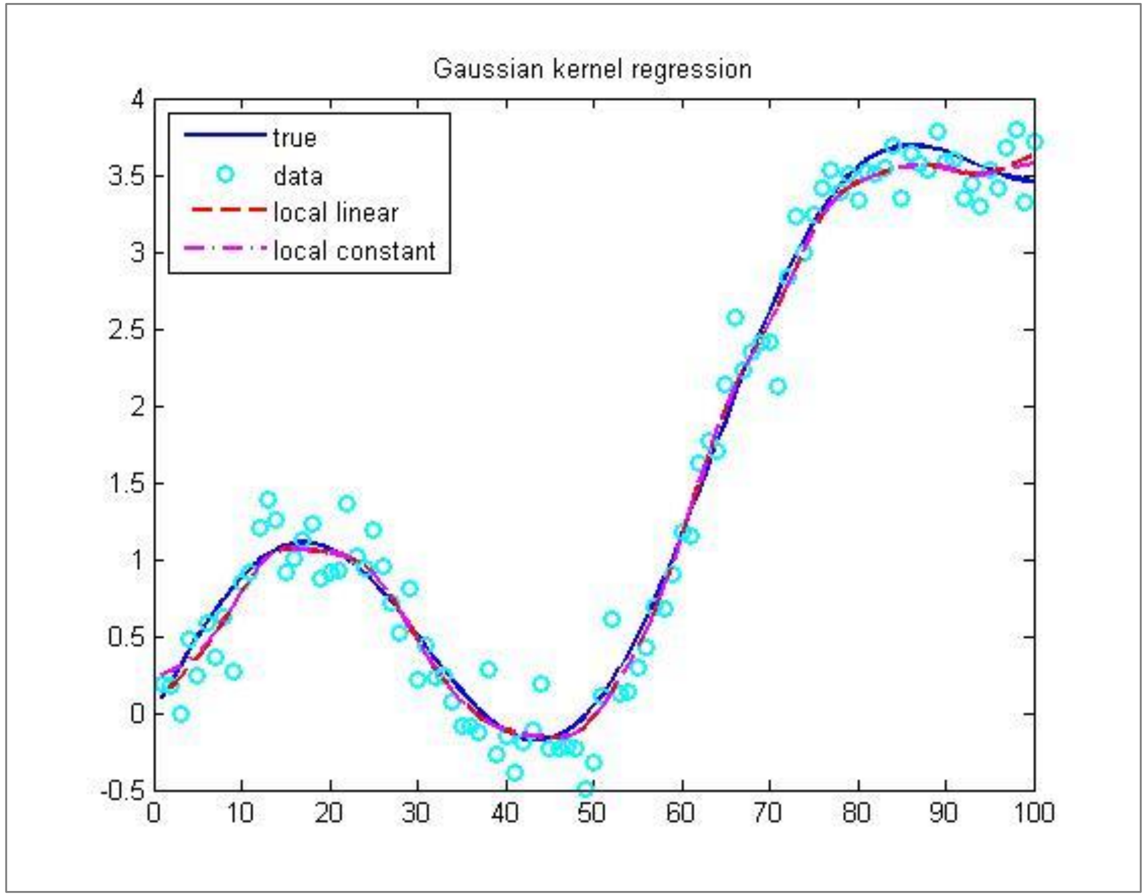
  非线性回归中，采用具有如下特征的函数对观测数据进行建模：

  The function is a nonlinear combination of the model parameters;

  该函数是模型参数的非线性组合；

  The function depends on one or more independent variables.

  该函数取决于一个或多个独立变量。



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

AI

10.2. Regression

Contents:

- □ 10.2.1. How Regression Works
- □ 10.2.2. Linear and Nonlinear
- □ 10.2.3. Applications and Algorithms

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

Thank you for your attention!

AI

10. Tasks in Machine Learning

Contents:

- □ 10.1. Classification
- □ 10.2. Regression
- □ 10.3. Clustering
- □ 10.4. Ranking
- □ 10.5. Dimensionality Reduction

Clustering

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

10.3. Clustering

Contents:

- □ 10.3.1. How Clustering Works
- □ 10.3.2. Major Approaches of Clustering
- □ 10.3.3. Applications and Algorithms

What is Clustering 什么是聚类

- □ A longer description 较长描述

  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.

  聚类是以这样的一种方式将对象进行分组的任务，即同一组中的对象彼此之间比其他组中的对象更相似。

- □ A shorter description 较短描述

  The process of organizing objects into groups whose members are similar in some way.

  将对象进行分组的过程，组内成员具有某种方式的相似性。

- □ A very short description 极简描述

  To group data objects.

  将数据对象分组。

Clustering vs. Classification 聚类与分类

- □ Similarity 相似性

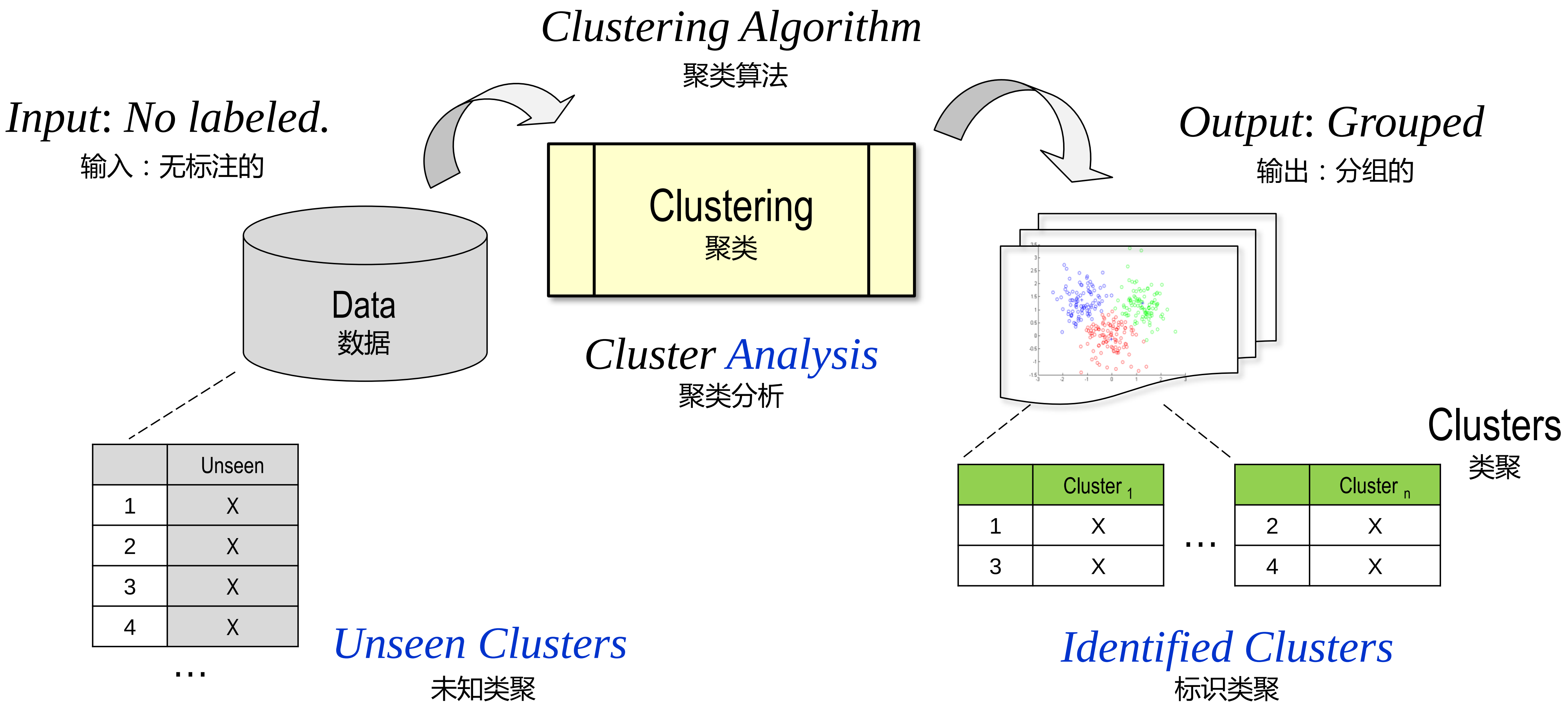
  Groups or classes

- □ Difference 差异性

  As shown in the following table 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、密度等发现类聚。	Classifiers need to have a high accuracy for classification. 分类器需要具有较高的分类精度。

Grouping Input Data into Same Cluster 将输入数据分成相同的类聚



Clustering Algorithm

聚类算法

Input: No labeled.

输入：无标注的

Output: Grouped

输出：分组的

Cluster Analysis

聚类分析

Clusters

类聚

	Unseen
1	X
2	X
3	X
4	X

…

Unseen Clusters

未知类聚

	Cluster $_1$
1	X
3	X

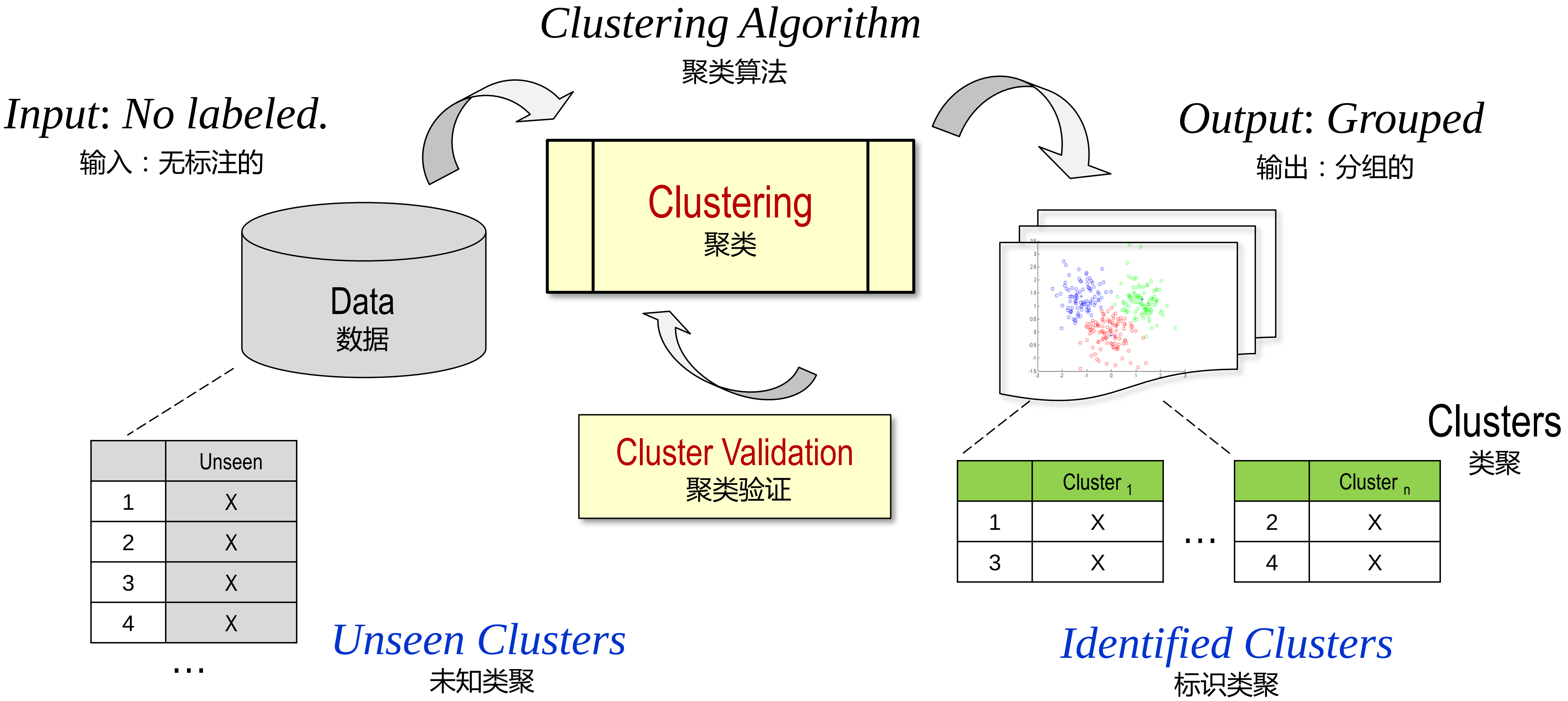
…

	Cluster $_n$
2	X
4	X

Identified Clusters

标识类聚

Two Key Steps in Clustering Procedure 聚类过程中的两个重要步骤



Clustering Algorithm
聚类算法

Input: No labeled.
输入：无标注的

Output: Grouped
输出：分组的

Data
数据

Clustering
聚类

Cluster Validation
聚类验证

	Unseen
1	X
2	X
3	X
4	X

…

Unseen Clusters
未知类聚

Clusters
类聚

	Cluster $_1$
1	X
3	X

…

	Cluster $_n$
2	X
4	X

Identified Clusters
标识类聚

A Formal Description of Clustering 一种聚类的形式化描述

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

设$\mathbb{R}^n$ $(n \geqq 1)$ 表示一个n维实数向量集，输入空间$\mathcal{X}$是$\mathbb{R}^n$的子集，输出空间$\mathcal{Y}$是一组未知的类聚， D是$\mathcal{X} \times \mathcal{Y}$笛卡尔乘积上的未知分布，则：

- □ Let a clustering function: 设聚类函数

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

- □ Clustering: 聚类

  Given a testing set of unknown clusters:

  给定一个未知类聚的测试集：

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

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

  采用上述确定的聚类函数来分析聚类结果：

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

10.3. Clustering

Contents:

□ 10.3.1. How Clustering Works

□ 10.3.2. Major Approaches of Clustering

□ 10.3.3. Applications and Algorithms

Typical Approaches of Clustering Algorithm 聚类算法的典型方法

- □ 1) Connectivity-based clustering 基于连接性聚类

  Also known as hierarchical clustering, based on the distance between objects.

  也被称为基于对象间距离的层次聚类。

- □ 2) Centroid-based clustering 基于中心点聚类

  To find the *k* cluster centers and assign the objects to nearest cluster center.

  发现*k*个类聚中心并将对象分配到最近的类聚中心点。

- □ 3) Distribution-based clustering 基于分布聚类

  Clusters can be defined as objects belonging most likely to the same distribution.

  类聚可被定义为恰好属于同一分布的对象群。

- □ 4) Density-based clustering 基于密度聚类

  To group objects into one cluster if they are connected by densely populated area.

  将稠密区域连接的对象组成一个类聚。

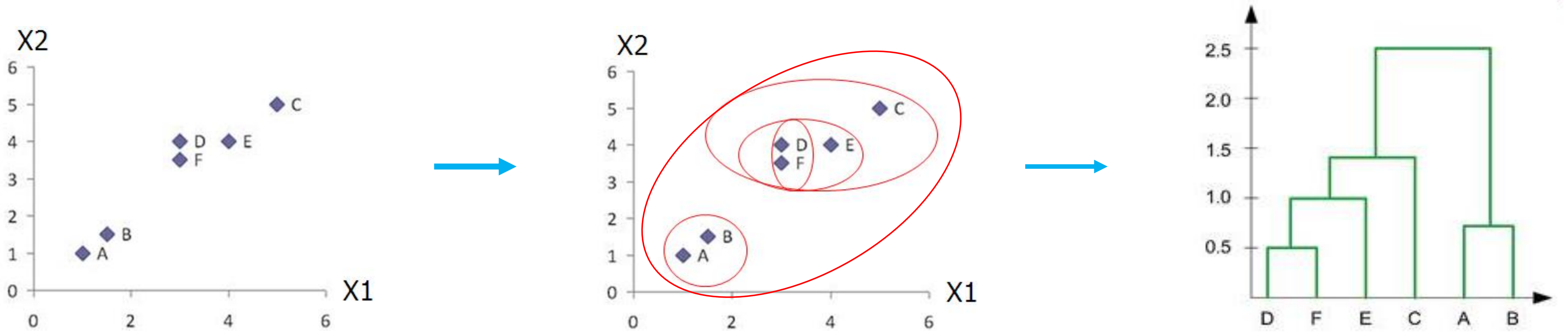
1) Connectivity-based clustering 基于连接性聚类

- Based on the core idea of objects being more related to nearby objects than to objects farther away.

  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。

- Creating a hierarchical decomposition of the set of data objects using some criterion.

  采用某种准则来创建数据对象集的层次分解。



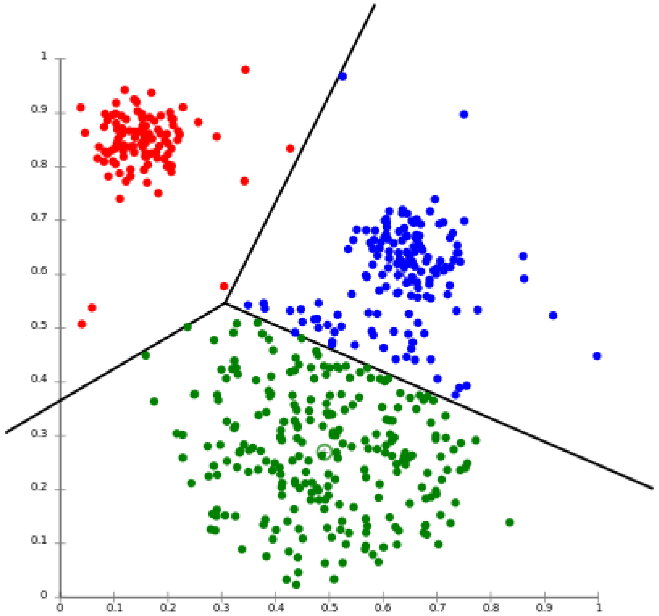
Typical algorithms: AGNES (Agglomerative NESting), DIANA (Divisive Analysis), ……

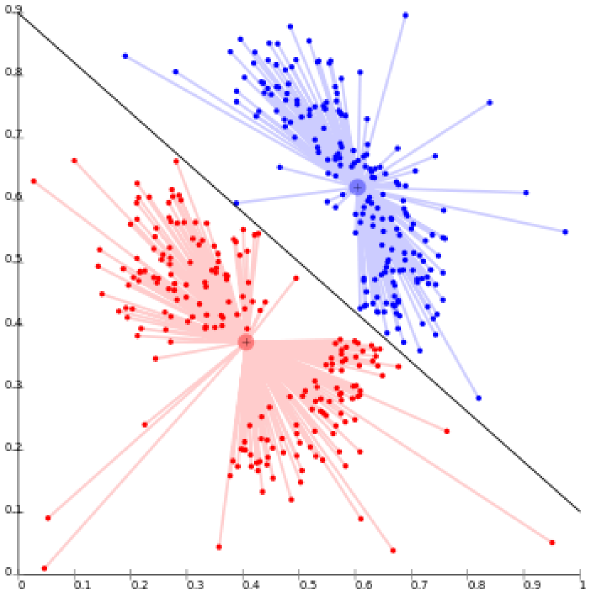
典型算法：AGNES (集聚嵌套), DIANA (分裂分析), ……

2) Centroid-based clustering 基于中心点聚类

- □ Constructing various partitions and then evaluating them by some criterion, e.g., minimizing the sum of square distance cost.

  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。



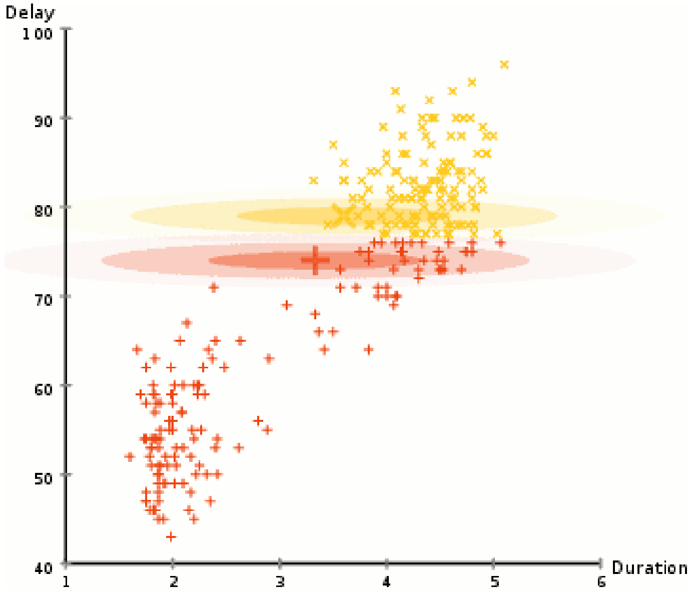


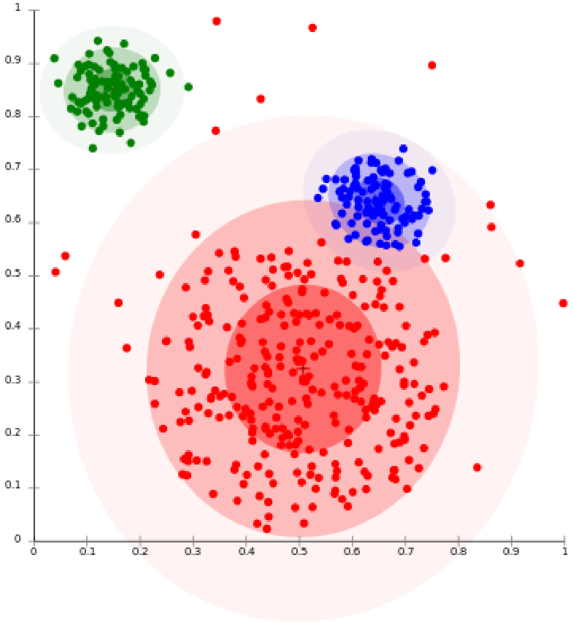
Typical algorithms: *k*-means, *k*-medoids, ……

典型算法：*k*-均值, *k*-中心点, ……

3) Distribution-based clustering 基于分布聚类

- □ Clusters are modeled using statistical distributions, such as multivariate normal distributions.
  采用统计分布（诸如多元正态分布）对类聚进行建模。



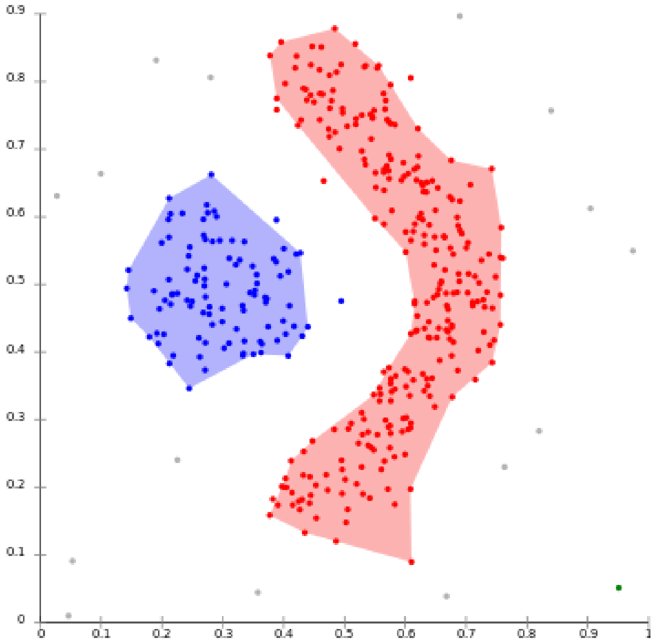


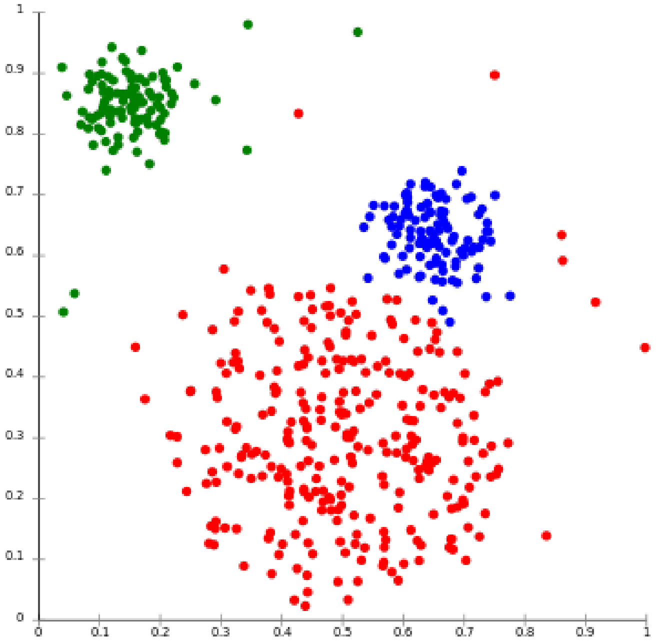
Typical algorithms: Expectation-maximization, ……

典型算法：期望最大化, ……

4) Density-based clustering 基于密度聚类

□ Clusters are defined as areas of higher density than the remainder of the data set.
类聚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 根据密度峰值聚类

Source: “Clustering by fast search and find of density peaks”, *SCIENCE, Vol. 344, Jun. 27 2014.*

- □ Cluster centers are characterized by
  - 1) a higher density than their neighbors,
  - 2) a larger distance from points with higher densities.

  类聚中心点的特性是：1) 密度高于其相邻点，2) 距离大于其它较高密度点。
- □ The features of the clustering method are:

  该聚类方法的特点：
  - the number of clusters arises intuitively,

    直观地得到类聚的个数，
  - outliers are automatically spotted and excluded,

    自动地发现和排除离群点，
  - clusters are recognized regardless of their shape, and space dimensionality.

    无论其形状以及空间的维度，类聚都能被识别。

Case Study: Clustering by density peaks 根据密度峰值聚类

Local density:
局部密度:

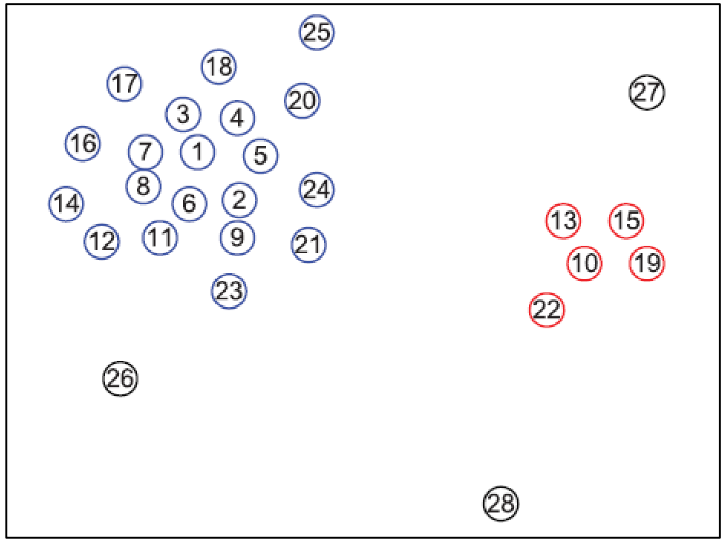
$$\rho_i = \sum_j \chi(d_{ij} - d_c) \qquad \chi(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{otherwise} \end{cases}$$

Minimum distance:
最小距离:

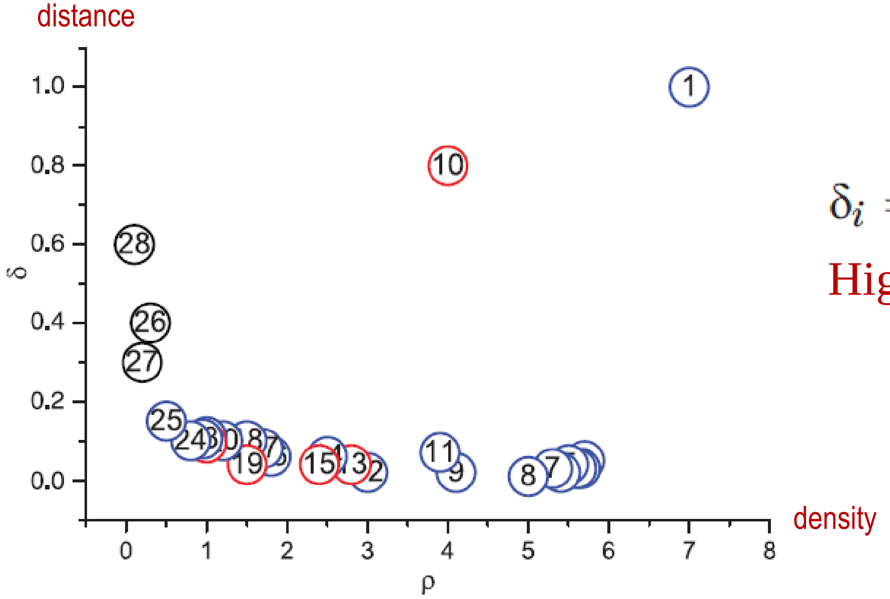
$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

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）



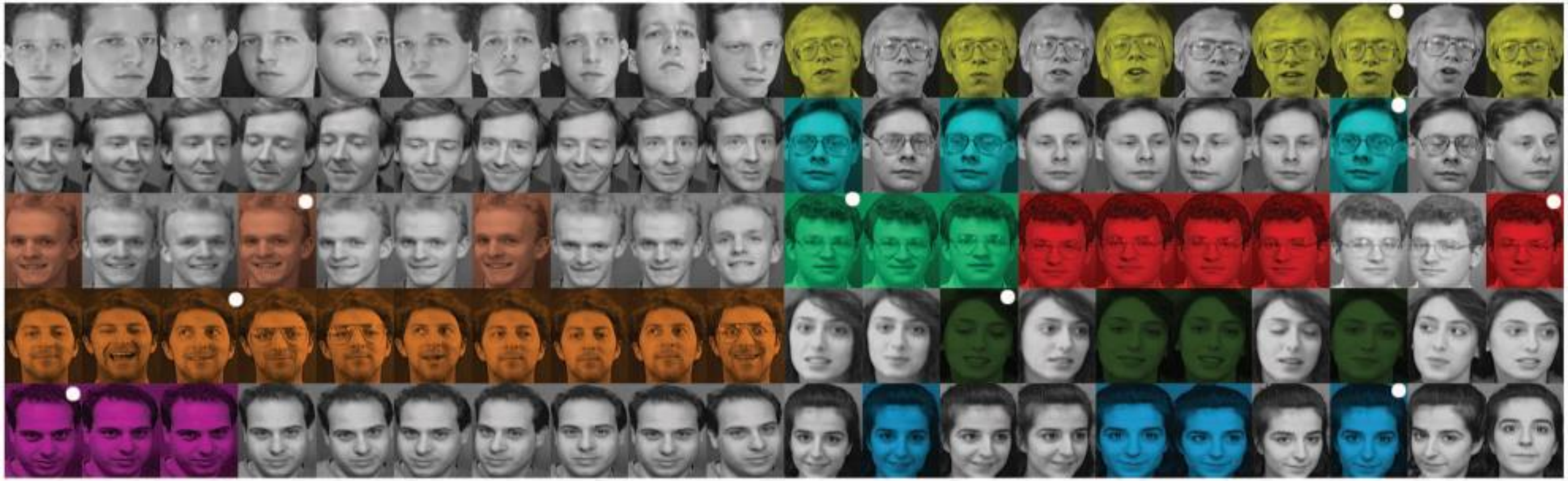
$$\delta_i = \max_j (d_{ij})$$

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 根据密度峰值聚类

- □ Clustering analysis of the Olivetti Face Database. 人脸数据库Olivetti的聚类分析



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.

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

10.3. Clustering

Contents:

- □ 10.3.1. How Clustering Works
- □ 10.3.2. Major Approaches of Clustering
- □ 10.3.3. Applications and Algorithms

Typical Applications of Clustering 聚类的典型应用

□ 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 典型的聚类算法

- □ *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

Thank you for your attention!

AI

10. Tasks in Machine Learning

Contents:

- □ 10.1. Classification
- □ 10.2. Regression
- □ 10.3. Clustering
- □ 10.4. Ranking
- □ 10.5. Dimensionality Reduction

Ranking

AI

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

What is Ranking 什么是排名

- □ A longer description 较长描述

  A ranking is a relationship between a set of items such that, for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second.

  排名是一组项之间的关系，即对于任意两个项，满足第一个“排名高于”、“排名低于”或“排名等于”第二个。

- □ A shorter description 较短描述

  The data transformation in which numerical or ordinal values are replaced by their rank.

  排名是一种数据转换，其中数值或者顺序值由其排名来代替。

- □ A very short description 极简描述

  To order items according to some criterion.

  依据某种准则整理数据项。

Contents:

□ 10.4.1. How Ranking Works

□ 10.4.2. Major Approaches of Ranking

□ 10.4.3. Applications and Algorithms

A Formal Description of Ranking 一种排名的形式化描述

Let $\mathcal{X}$ denote input space, D an unknown distribution over $\mathcal{X} \times \mathcal{X}$.

设$\mathcal{X}$表示输入空间，D是$\mathcal{X} \times \mathcal{X}$上的未知分布。

□ Target ranking function: 目标排名函数：

$$f : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{Y} = \{-1, 0, +1\}$$

where 其中

- $f(x, x') = +1$, if x is ranked higher than x', 若x排名高于x'，
- $f(x, x') = -1$, if x is ranked lower than x', 若x排名低于x'，
- $f(x, x') = 0$, if both x and x' has same ranking. 若x与x'二者排名相同。

□ Training data: 训练数据

$$\mathcal{S} = \{(x^{(i)}, x'^{(i)}, y^{(j)}) \mid y^{(j)} = f(x^{(i)}, x'^{(i)}) \in \mathcal{Y}, i \in [1, m], j \in [1, 3]\}$$

A Formal Description of Ranking 一种排名的形式化描述

- Ranking problem: 排名问题
  Given a hypothesis set H of functions mapping $\mathcal{X} \times \mathcal{X}$ to $\mathcal{Y} = \{-1, 0, +1\}$, to select a hypothesis $h \in H$ with the target function f:

  给定一个将$\mathcal{X} \times \mathcal{X}$映射到$\mathcal{Y} = \{-1, 0, +1\}$的假设函数集$H$，选择一个具有目标函数$f$的假设$h \in H$：

  - small expected generalization error: 最小预期泛化错误：

$$R(h) = \Pr_{(x, x')}[f(x, x') \neq 0 \land (f(x, x')(h(x') - h(x)) \leq 0)]$$

  - empirical pairwise misranking error: 经验性成对误排名错误：

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\,((y^{(i)} \neq 0) \land (y^{(i)}\,(h(x'^{\,(i)}) - h(x^{(i)})) \leq 0)))$$

Contents:

- □ 10.4.1. How Ranking Works
- □ 10.4.2. Major Approaches of Ranking
- □ 10.4.3. Applications and Algorithms

Typical Approaches of Ranking 典型的排名方法

□ 1) Score-based approach 基于分值方法

- The predictor is a real-valued function, called *scoring function*.
  该预测器是一个实数函数，称为分值函数。
- The scores assigned to input points by this function determine their ranking.
  由该函数分派给输入数据点的分值决定其排名。
- This approach is the most widely explored one.
  这种方法是研究得最多的一种。

□ 2) Preference-based approach 基于偏好方法

- The predictor is a *preference function.*
  该预测器是一个偏好函数。

10.4. Ranking

Contents:

- □ 10.4.1. How Ranking Works
- □ 10.4.2. Major Approaches of Ranking
- □ 10.4.3. Applications and Algorithms

Typical Applications of Ranking 排名的典型应用

□ In information retrieval	信息检索领域
■ Search engine	搜索引擎
■ Document retrieval	文档检索
■ Collaborative filtering	协同式过滤
■ Sentiment analysis	情感分析
■ Computational advertising	计算广告学
□ In other areas	其它领域
■ Machine translation	机器翻译
■ Recommender systems	推荐系统
■ Computational biology	计算生物学
■ Proteomics	蛋白质组学

Case Study: PageRank

- □ An algorithm used by Google to rank websites in their search engine, named after Larry Page, one of Google founders.

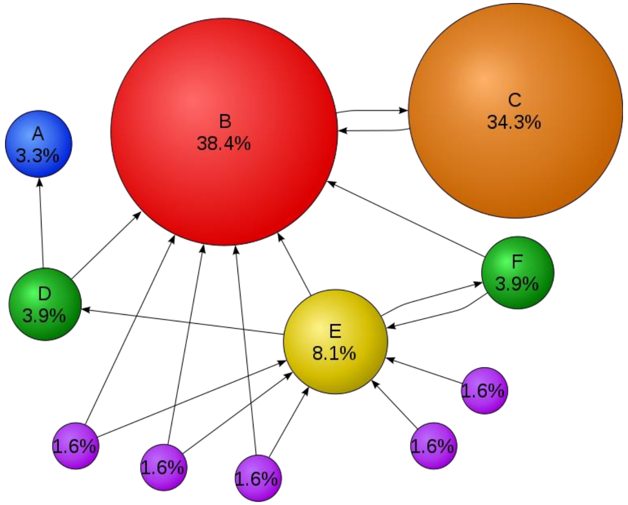
  谷歌用于在其搜索引擎中对网站进行排名的一种算法，以谷歌创始人之一拉里·佩奇的名字命名。

- □ PageRank works by counting the number and quality of links to a page to determine how important the website is.

  PageRank通过计算网页的链接数量和质量来决定该网站的重要性。

- □ The underlying assumption is that more important websites are likely to receive more links from other websites.

  其基本假设是：越重要的网站，就会被越多其它网站所链接。



Thank you for your attention!

AI

10. Tasks in Machine Learning

Contents:

- □ 10.1. Classification
- □ 10.2. Regression
- □ 10.3. Clustering
- □ 10.4. Ranking
- □ 10.5. Dimensionality Reduction

Dimensionality Reduction

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

What is Dimensionality Reduction 什么是降维

□ A longer description 较长描述

To transform an initial very high-dimensional representation of data into a lower-dimensional representation of these data while preserving some properties of the initial representation.

将初始的极高维数据表示转换为这些数据的低维表示，而保留原始表示的某些性质。

□ A shorter description 较短描述

To simplify inputs by mapping high-dimensional space into a lower dimensional representation.

通过将高维空间映射到低维空间表示来简化输入。

□ A very short description 极简描述

To map inputs into a lower dimensional space.

将输入映射到低维空间。

10.5. Dimensionality Reduction

Contents:

□ 10.5.1. Why Dimensionality Reduction

□ 10.5.2. Linear and Nonlinear

□ 10.5.3. Applications

Why Dimensionality Reduction 为什么降维

□ Curse of dimensionality 维度灾难

- This phenomena arises when analyzing data in high-dimensional spaces.
  当在高维空间对数据进行分析时，该现象就会发生。

□ Data sparsity or irrelevant 数据稀疏或无关

- When the dimensionality increases, the volume of the space increases so fast that the available data become sparse.
  随着维度的增加，空间的体积增长非常迅速，使得可用的数据变得稀疏。
- Some features may be irrelevant.
  某些特征可能是无关的。

□ Visualization 可视化

- The data with two or three dimensions is easy to represent.
  二维或三维数据易于表示。

10.5. Dimensionality Reduction

Contents:

- □ 10.5.1. Why Dimensionality Reduction
- □ 10.5.2. Linear and Nonlinear
- □ 10.5.3. Applications

Linear and Nonlinear 线性与非线性

- Linear Dimensionality Reduction 线性降维
  - performs a linear mapping high-dimensional input data to a lower dimensional space.

    采用某种线性方式将高维输入数据映射到低维空间。
- Nonlinear Dimensionality Reduction 非线性降维
  - performs a nonlinear mapping high-dimensional input data to a lower dimensional space.

    采用某种非线性方式将高维输入数据映射到低维空间。

Typical Methods of Linear Dimensionality Reduction 线性降维的典型方法

- □ Principal Component Analysis (PCA) 主成分分析 (PCA)
- □ Linear Discriminate Analysis (LDA) 线性判别分析 (LDA)
- □ Multilinear subspace learning 多线性子空间学习
  - ■ Multilinear Principal Component Analysis (MPCA) 多线性主成分分析
  - ■ Multilinear Linear Descriminant Analysis (MLDA) 多线性线性判别分析

Example: Principal Component Analysis (PCA) 主成分分析 (PCA)

- □ PCA is a statistical procedure.

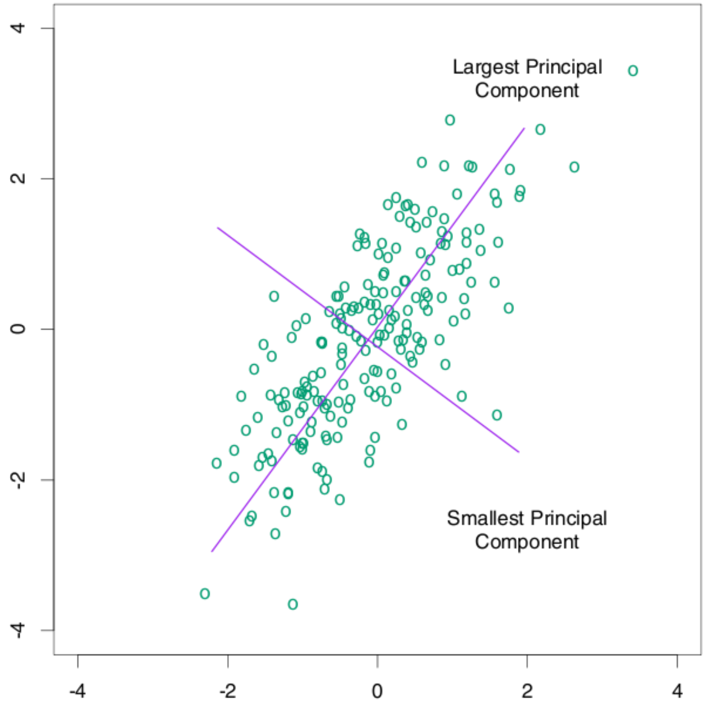
  PCA是一种统计过程。

- □ It uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

  它采用正交变换方法将一组可能相关变量的观测值转换成一组称为主成分的线性不相关变量值。

- □ The number of principal components is less than the number of original variables.

  主成分的数量小于原始变量的数量。



Approaches of Nonlinear Dimensionality Reduction 非线性降维的方法

Multi-dimensional Scaling 多元尺度分析

- Classical multidimensional scaling 经典多元尺度分析
- Metric multidimensional scaling 度量多元尺度分析
- Non-metric multidimensional scaling 非度量多元尺度分析
- Generalized multidimensional scaling 广义多元尺度分析

Kernel approaches 核方法

- Kernel Principal Component Analysis 核主成分分析
- Kernel Fisher Discriminant Analysis (KFD) 核费希尔判别分析

Manifold learning approaches 流形学习方法

- Isometric feature mapping (Isomap) 等距特征映射
- Locally-linear embedding (LLE) 局部线性嵌入

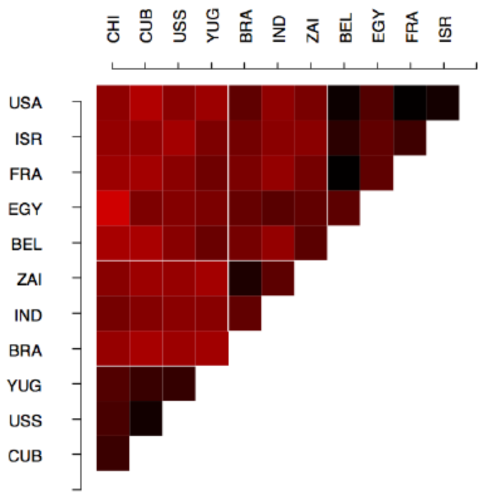
Example: Multi-dimensional Scaling (MDS) 多元尺度分析(MDS)

- □ MDS is a set of related statistical techniques often used in data visualisation for exploring similarities or dissimilarities in data.

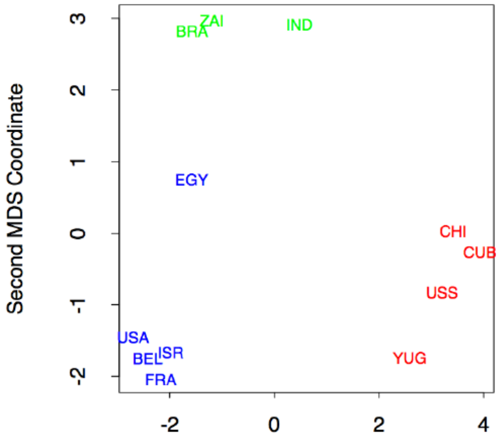
  MDS是用于数据可视化的一些相关统计技术，用来考察数据中的相似性和非相似性。

- □ An MDS algorithm takes a matrix of pair-wise distances between all points, then computes a position for each point in a low-dimensional space, suitable for 2D or 3D visualisation.

  MDS算法构建一个所有点之间的成对距离矩阵，然后在低维空间计算每个点的位置，便于二维或三维可视化。



Recorded Dissimilarity Matrix
录得的非相似矩阵



First MDS Coordinate
第一个MDS坐标

10.5. Dimensionality Reduction

Contents:

- □ 10.5.1. Why Dimensionality Reduction
- □ 10.5.2. Linear and Nonlinear
- □ 10.5.3. Applications

Typical Applications of Dimensionality Reduction 降维的典型应用

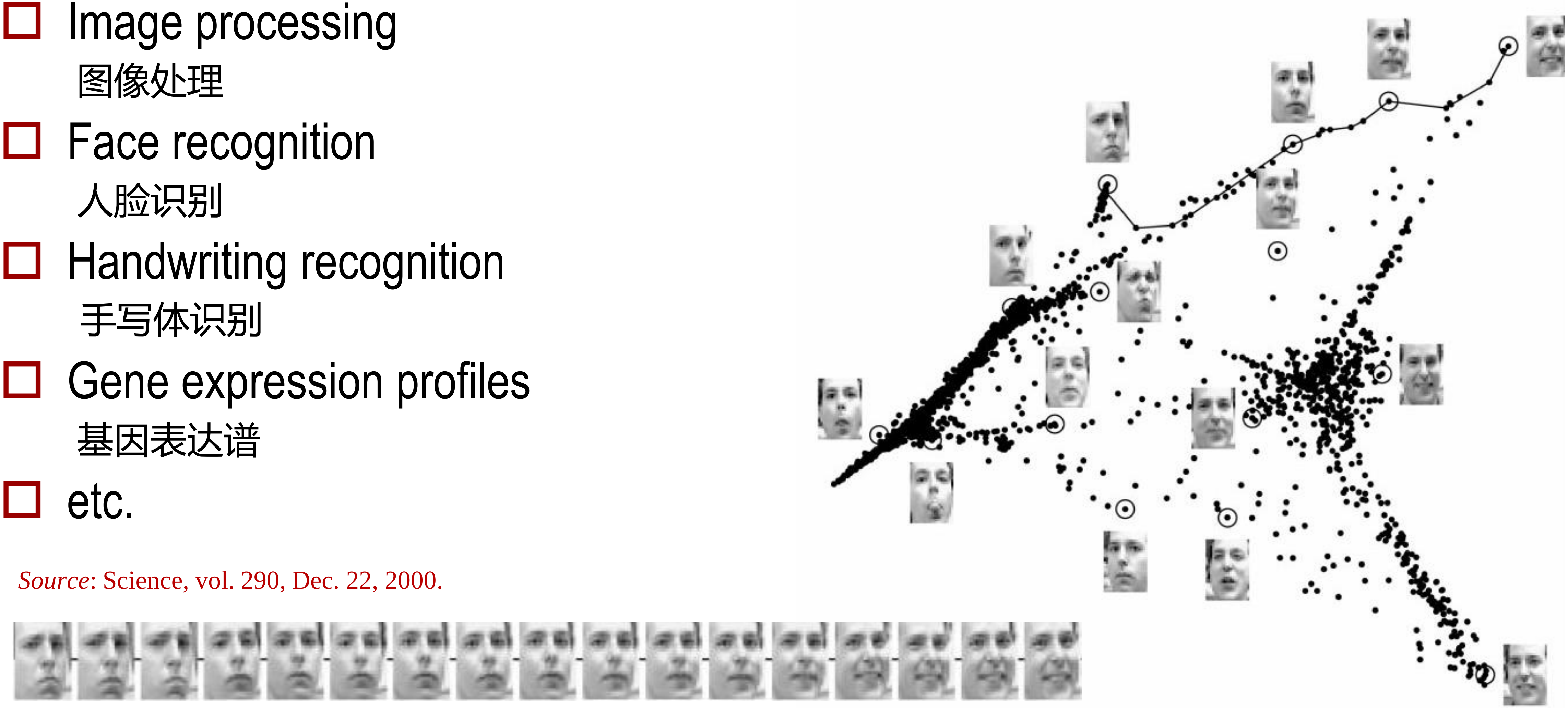
□ Image processing
图像处理

□ Face recognition
人脸识别

□ Handwriting recognition
手写体识别

□ Gene expression profiles
基因表达谱

□ etc.



Source: Science, vol. 290, Dec. 22, 2000.

Thank you for your attention!

AI