10. Tasks in Machine Learning

Contents:

Regression

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}

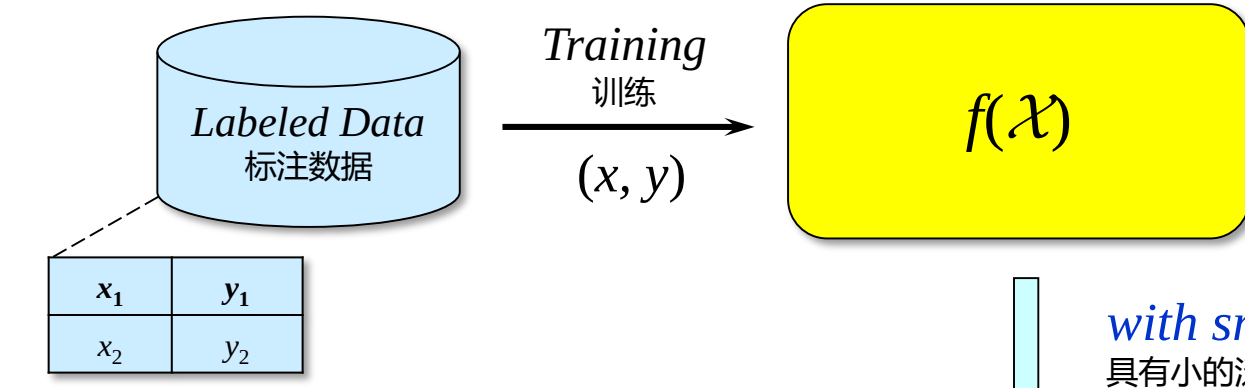
10.2. Regression

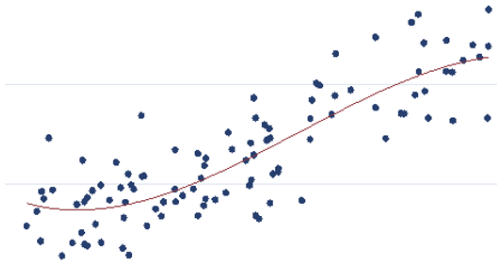
Contents:

Regression: Training 回归：训练

Known Output Value
已知输出值

Learning Algorithm
学习算法





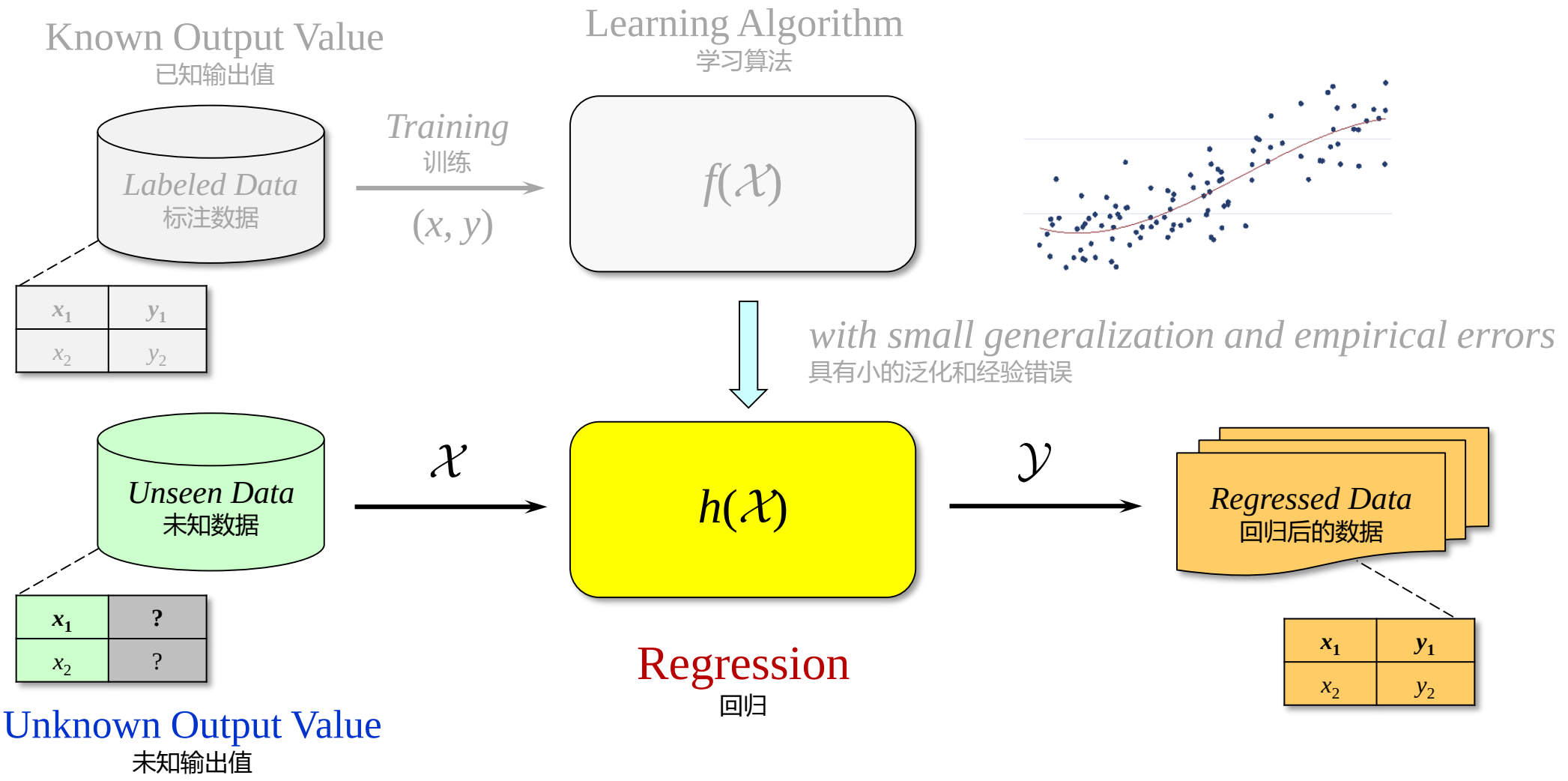
Labeling function
标注函数

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：实测



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：

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

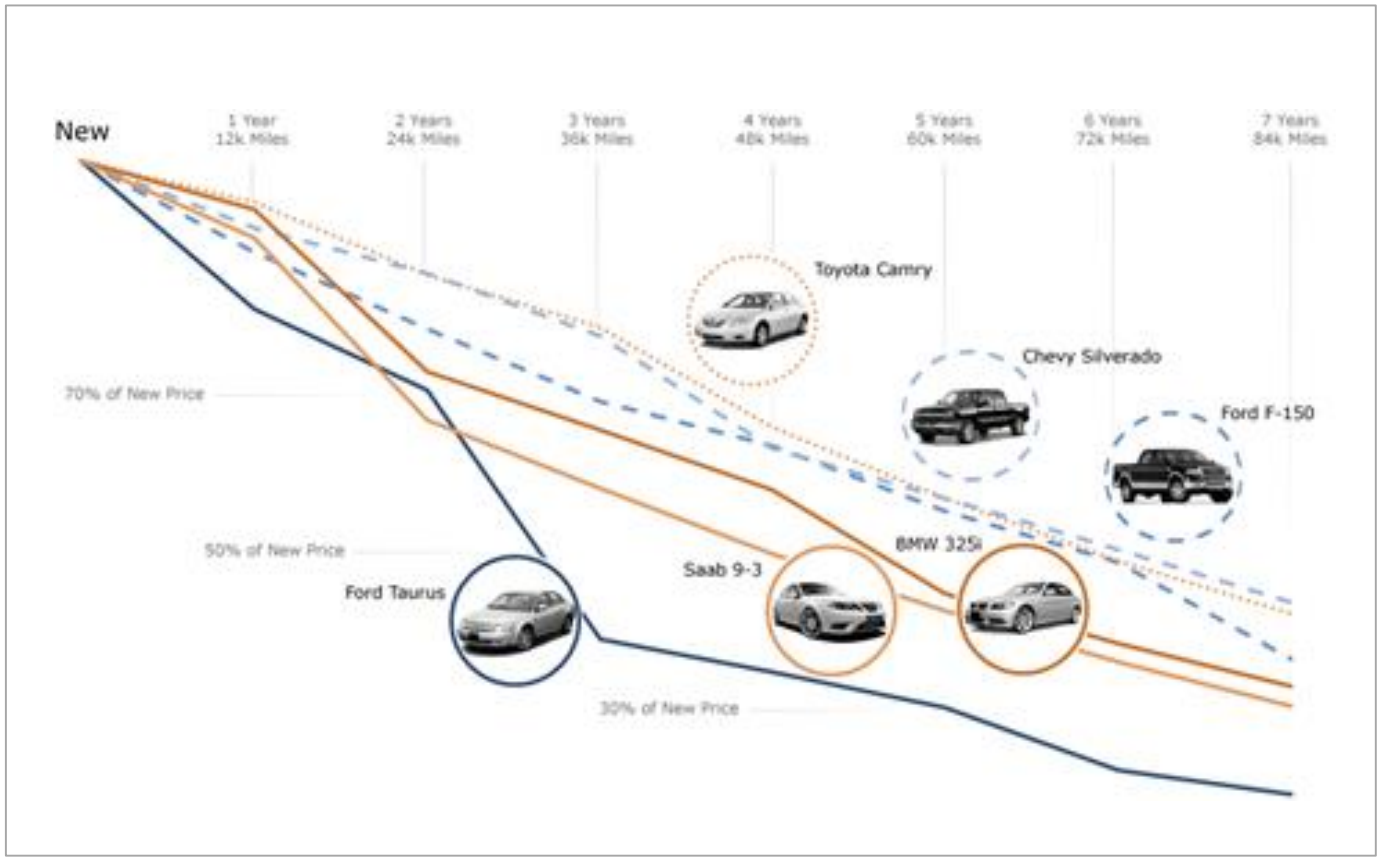
  Note, in which: 注意，其中

  $\mathcal{Y}$ 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:

Linear Regression 线性回归

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

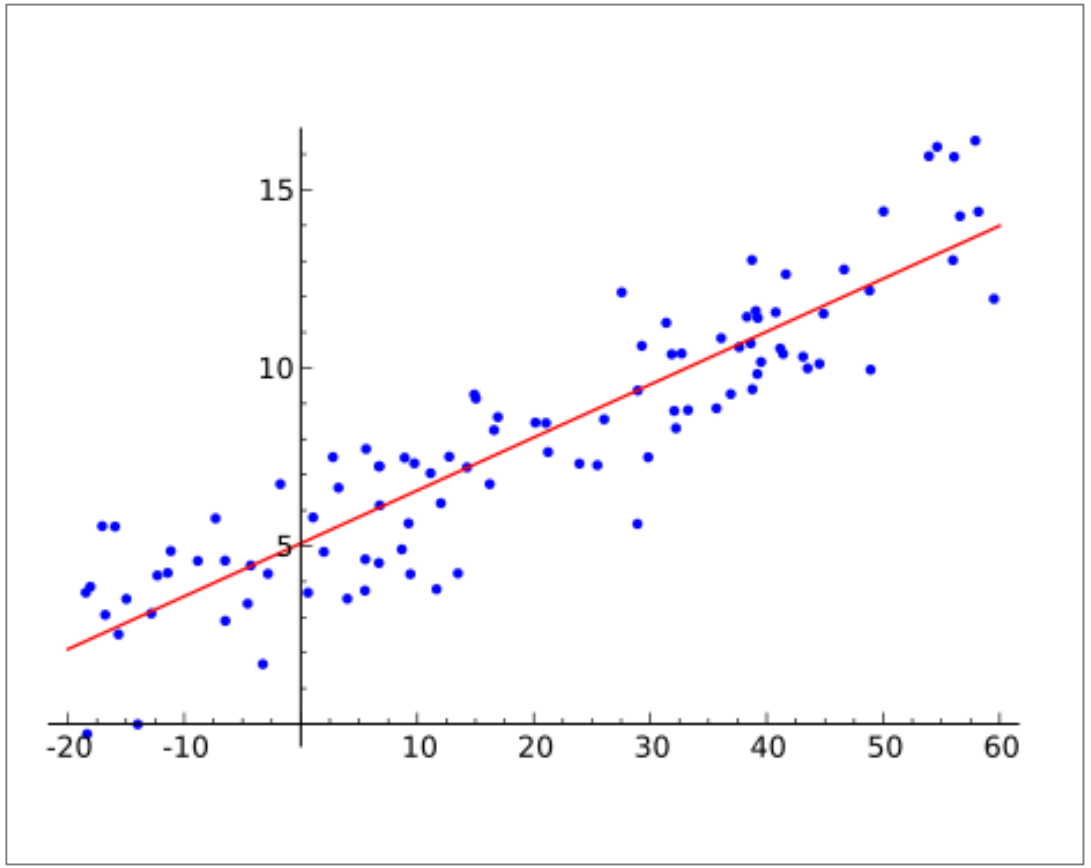
  线性回归中，采用具有如下特征的函数对观测数据进行建模：

  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:

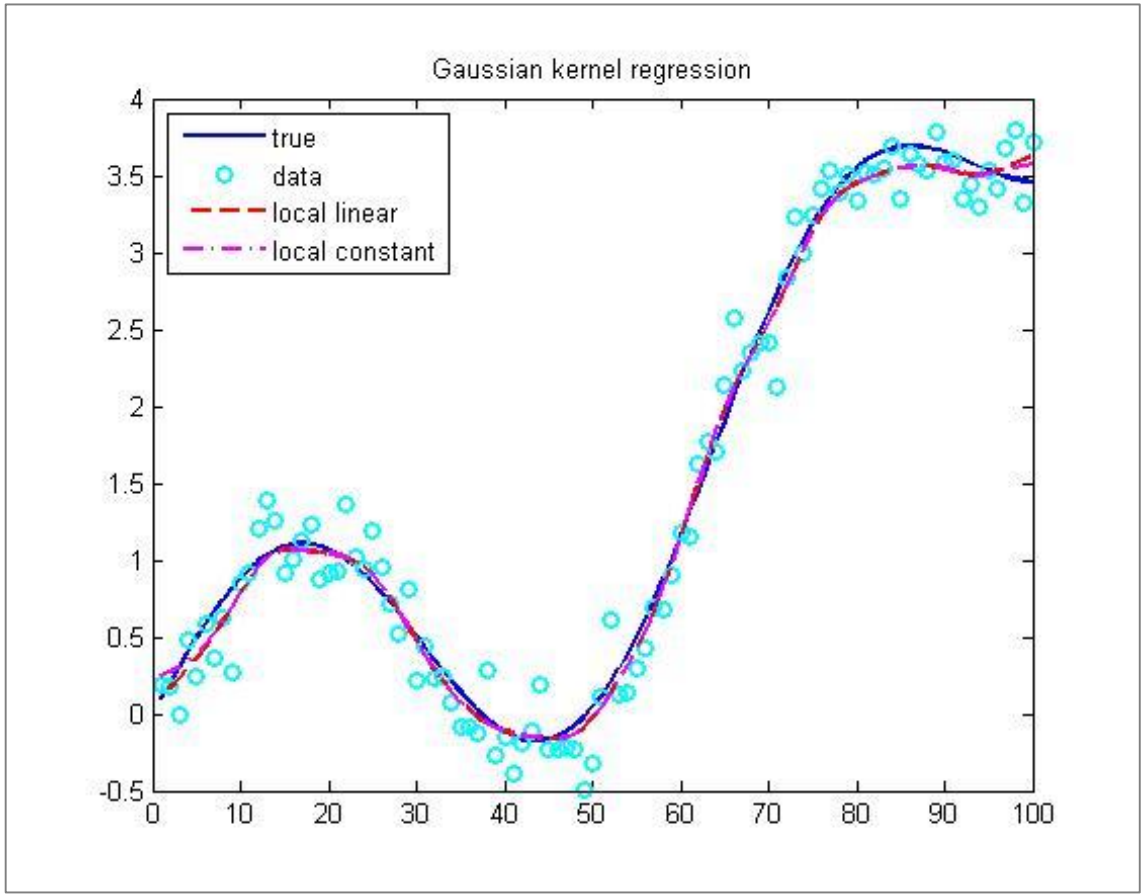
  非线性回归中，采用具有如下特征的函数对观测数据进行建模：

  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。



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

10.2. Regression

Contents:

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