大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

北京理工大学计算机学院　孙新

2019年1月

3.2 机器学习的典型任务——回归

- 产生：英国统计学家F.GALTON（法兰西斯·高尔顿）(1822-1911)和其学生K.Pearson（卡尔.皮尔逊）(1856-1936)观察了1078对夫妇，以每对夫妇的平均身高为X，而取他们成年的儿子的身高为Y，得到如下经验方程： Y=33.73+0.516X

- 定义：假定同一个或多个独立变量存在相关关系，寻找相关关系的模型。不同于时间序列法的是：模型的因变量是随机变量，而自变量是可控变量。分为线性回归和非线性回归，通常指连续要素之间的模型关系，是因果关系分析的基础。（回归研究的是数据之间的非确定性关系）

3.2 机器学习的典型任务——回归

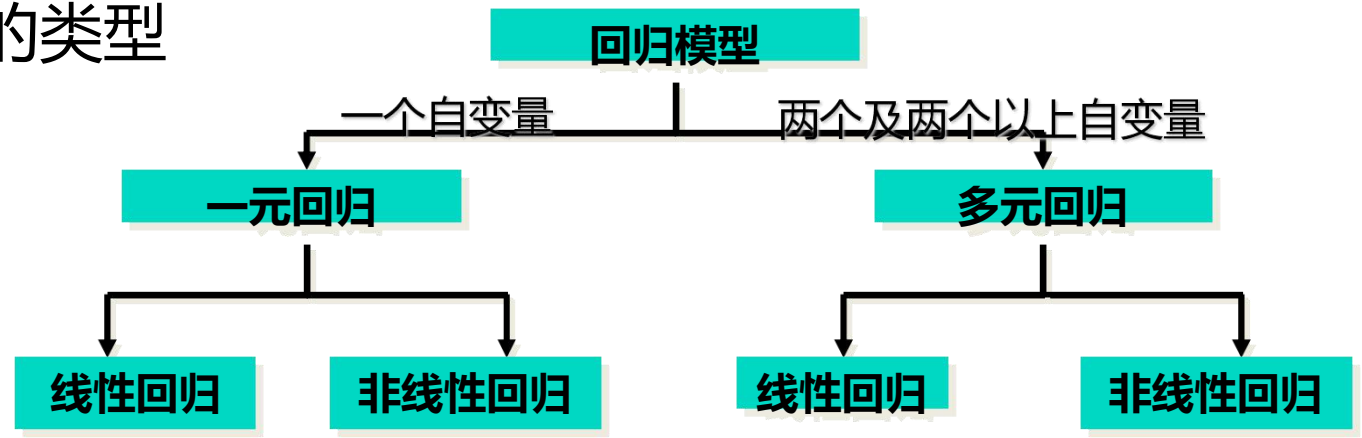
- 回归：

  通过一个给定的输入预测一个连续值而不是离散的输出。

	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}

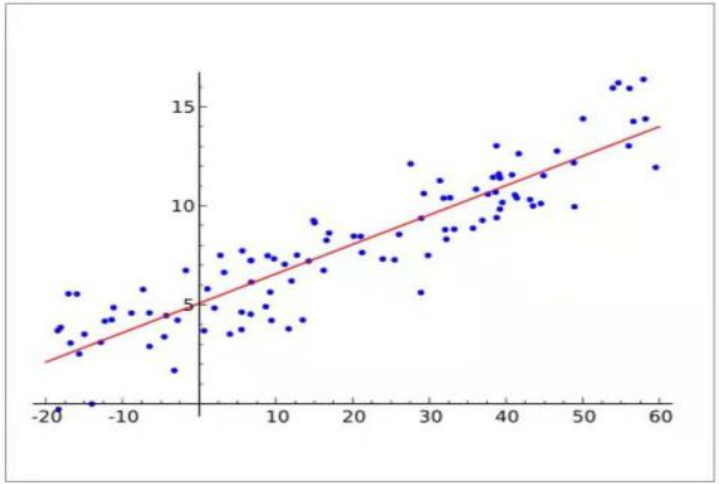
3.2 机器学习的典型任务——回归

- 在大数据分析中，回归分析是一种预测性的建模技术，
- 回归分析就是对具有相关关系的两个或两个以上变量之间数量变化的一般关系进行测定，确定因变量和自变量之间数量变动关系的数学表达式，以便对因变量进行估计或预测的统计分析方法
- 回归模型的类型



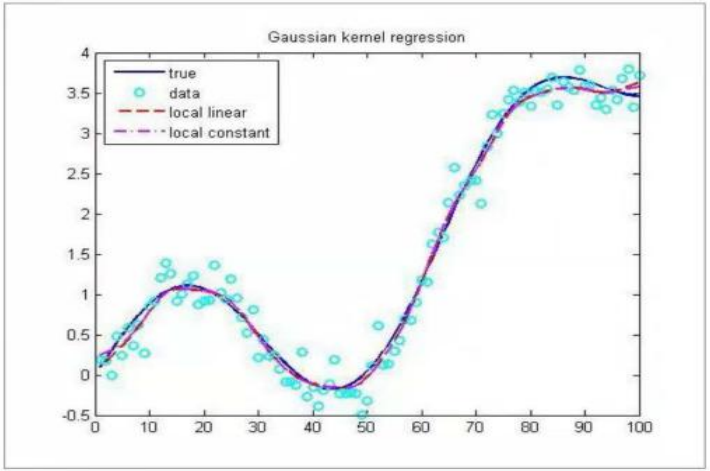
3.2 机器学习的典型任务——回归

线性回归算法寻找属性与预测目标之间的线性关系。线性回归中，采用模型参数的线性组合函数对观测数据进行建模。该模型取决于一个或多个独立变量。



$$y(x) = w \cdot x + b$$

非线性回归中，采用模型参数的非线性组合函数对观测数据进行建模。



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

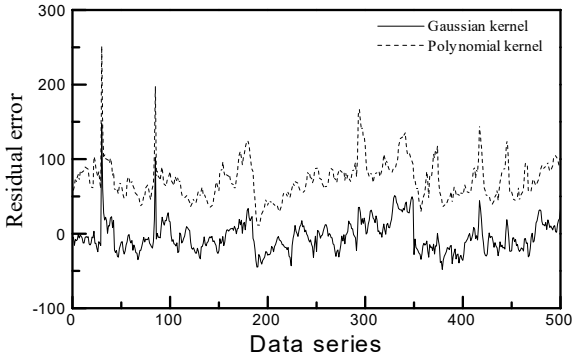
- 回归分析的主要内容
  - ➢ 建立回归方程

    利用回归方程，配合一个表明变量之间数量关系的方程式，而且根据自变量x的变动，来预测因变量y的变动
  - ➢ 测定因变量的估计值与实际值的误差程度

    通过计算估计标准误差指标，可以放映因变量估计值的准确程度，从而将误差控制在一定范围内
  - ➢ **回归分析不仅可以揭示变量 x 对变量 y 的影响大小，还可以由回归方程进行预测和控制**

3.2 机器学习的典型任务——回归

线性回归在建立回归模型之前，可先进行主成分分析，消除属性之间的相关性。最后通过最小二乘法，得到各属性与目标之间的线性系数。



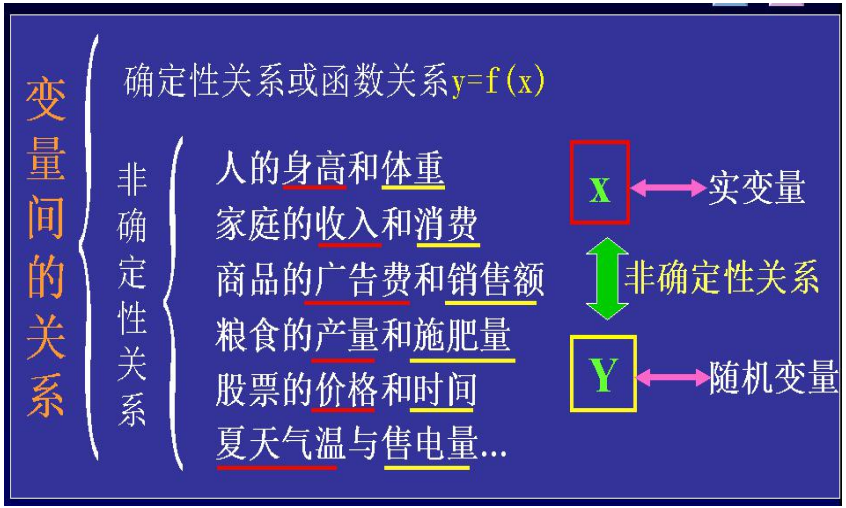
$(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)$ ← 数据(Data)

$y=f(\beta,x)$ → $y=f(x)$ ← 模型(Model)

x_i → $y=f(x)$ → y_i

y 是离散的，如{-1,1}，{0,1,2}为分类问题

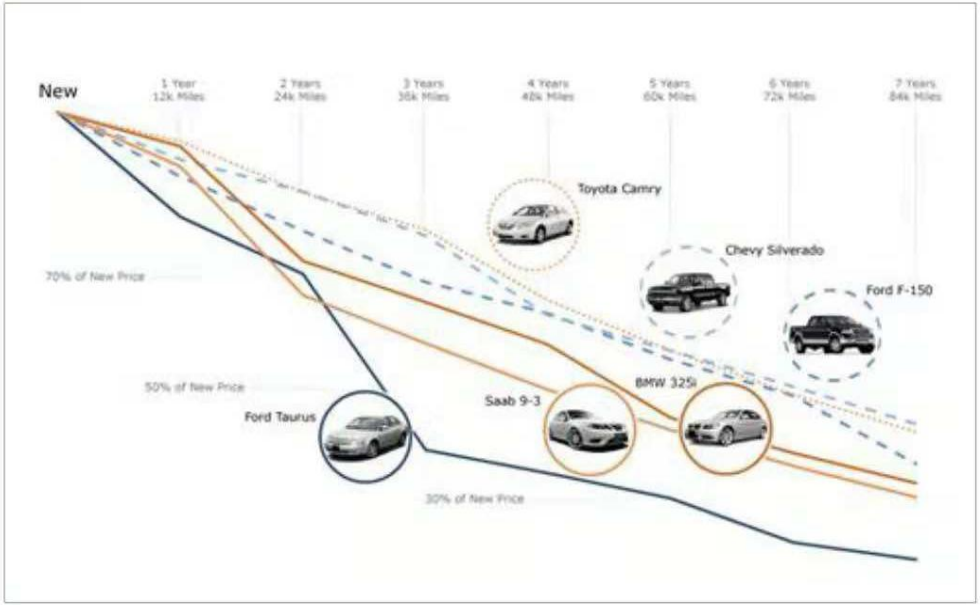
y 是连续值如温度，速度等为回归问题



3.2 机器学习的典型任务——回归

示例：二手车价格

- 构建一个预测二手车价格的系统
- 输入车的各种属性，包括：车辆的类型、名称、型号、注册登记日期、已使用年限、累计行驶里程等信息。
- 输出是二手车量的预测价格

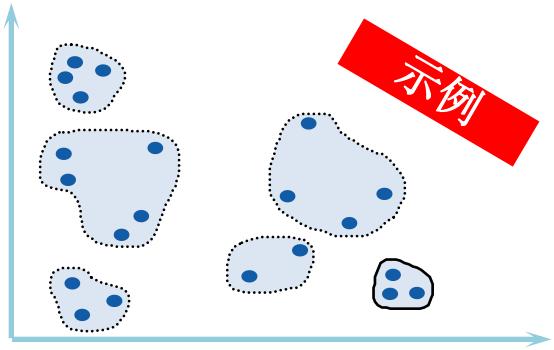


Used car prices

二手车价格

3.2 机器学习的典型任务——聚类

- 聚类分析对具有共同趋势或结构的数据进行分组，将数据项分组成多个簇（类），簇之间的数据差别应尽可能大，簇内的数据差别应尽可能小，即“最小化簇间的相似性，最大化簇内的相似性”。



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。

3.2 机器学习的典型任务——聚类

基于划分的聚类

- 对给定的数据集合，事先指定划分为k个类别。
- **典型算法：k-均值法**和k-中心点算法等。

基于层次的聚类

- 对给定的数据集合进行层次分解，不需要预先给定聚类数，但要给定终止条件，包括凝聚法和分裂法两类。
- **典型算法：** CURE、Chameleon、BIRCH、Agglomerative

基于密度的聚类

- 只要某簇邻近区域的密度超过设定的阈值，则扩大簇的范围，继续聚类。这类算法可以获得任意形状的簇。
- **典型算法：**DBSCAN、OPTICS和DENCLUE等

基于网格的聚类

- 首先将问题空间量化为有限数目的单元，形成一个空间网格结构，随后聚类在这些网格之间进行。
- **典型算法：**STING、WareCluster和CLIQUE等 。

基于模型的聚类

- 为每个簇假定一个模型，寻找数据对模型的最佳拟合。所基于的假设是：数据是根据潜在的概率分布生成的。
- **典型算法：**COBWEB和神经网络算法等。

3.2 机器学习的典型任务——聚类

划分方法
(partitioning)

层次方法
(hierarchical)

基于密度
(Density based)

基于分布
(Distribution based)

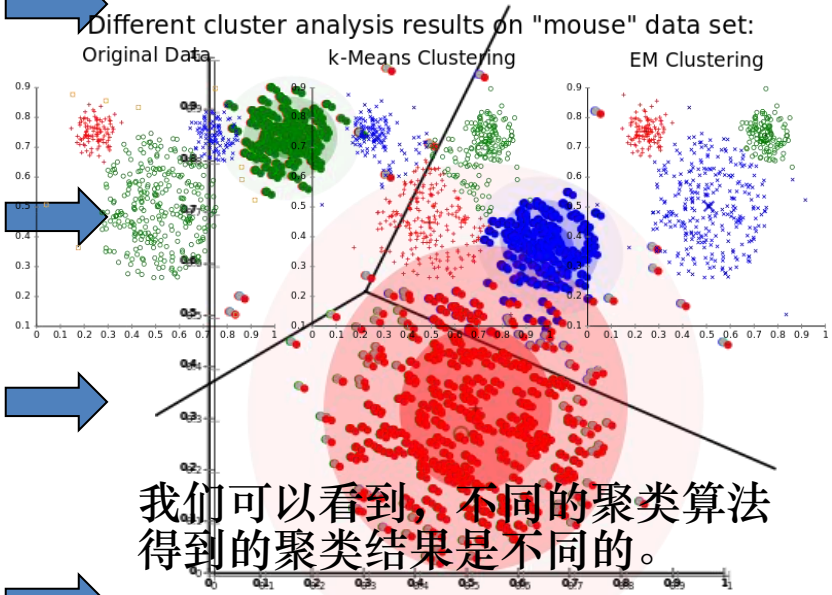
.....

k-Means

Single-linkage

DBSCAN

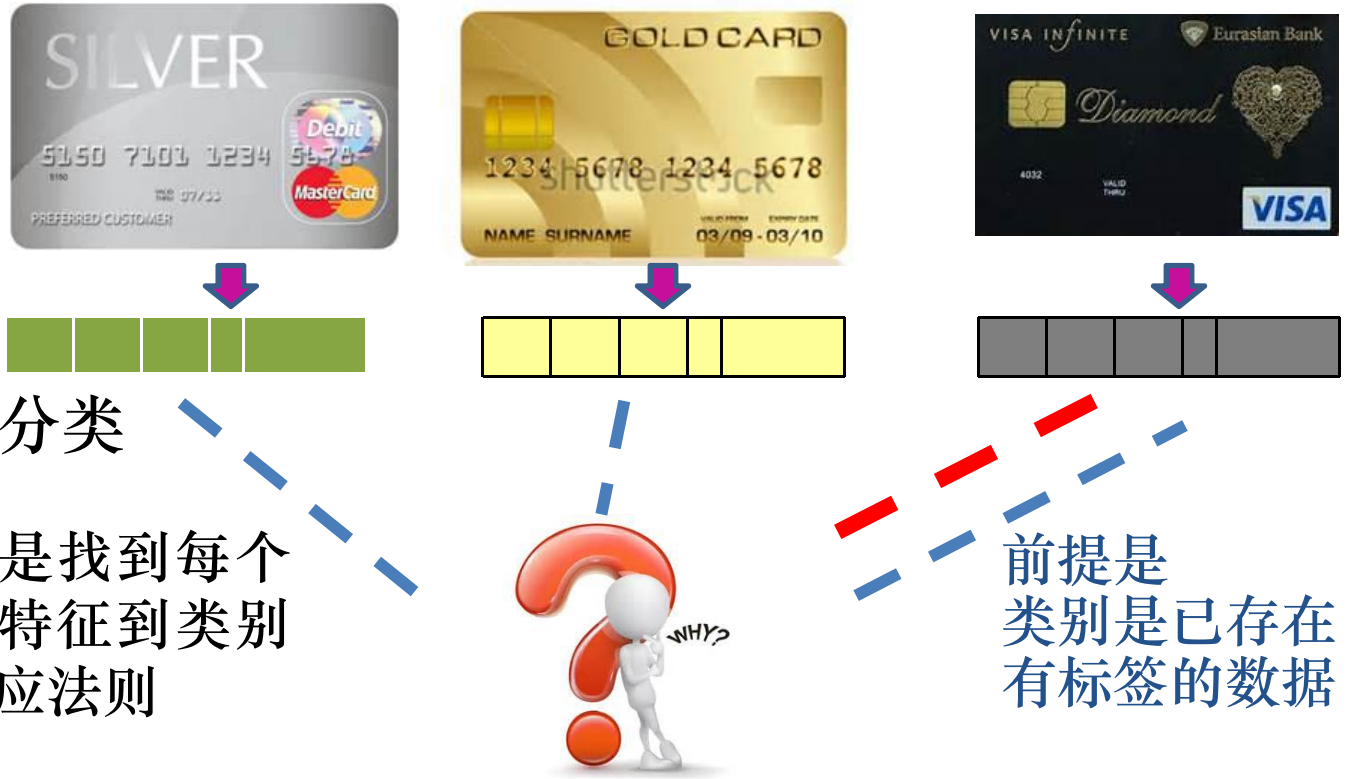
EM
Expectation-maximization algorithm



我们可以看到，不同的聚类算法得到的聚类结果是不同的。

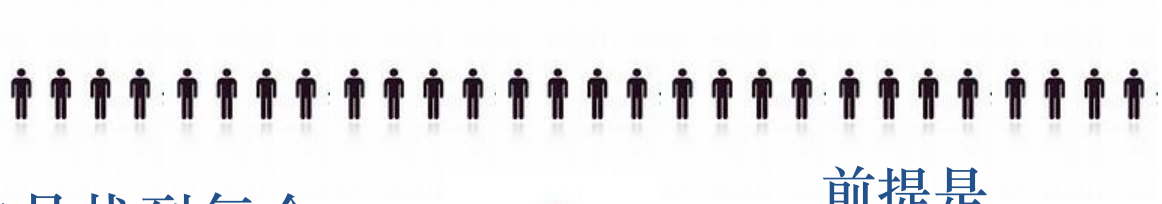
3.2 机器学习的典型任务——聚类

- 举例：分类 - 聚类



3.2 机器学习的典型任务——聚类

- 举例：分类 - 聚类



目的是找到每个样本潜在的类别，并将同类别的样本放在一起

前提是
类别是不存在

无标签的数据

3.2 机器学习的典型任务——聚类

聚类的典型应用

- 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