中国科学院大学
University of Chinese Academy of Sciences

机器学习基础

AI DISCOVERY

什么是机器学习

AI DISCOVERY

机器学习是从人工智能中产生的一个重要学科分支，是实现智能化的关键。

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

Machine learning is the study of **algorithms** and mathematical **models** that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model of **sample data**, known as “training data”, in order to make predictions or decisions without being explicitly programmed to perform the task.

Wikipedia

机器学习方法

AI DISCOVERY

有监督学习（supervised learning）：从给定的**有标注的训练数据集**中学习出一个函数（模型参数），当新的数据到来时可以根据这个函数预测结果。 常见任务包括**分类**与**回归**。

分类：输出是类别标签

Classification: Y is discrete

Y：年轻人(1)，老年人(-1)

X： x_1 黑头发的比例， 值域（0，1）；

x_2 行走速度， 值域（0，100）米/每分钟.

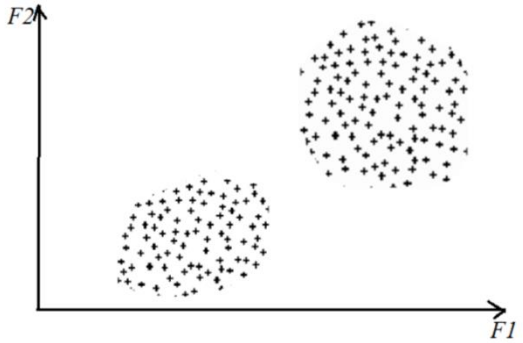
Training Data:

Y=1：(1，99)、(0.9，80)、(0.80，100) …

Y=-1：(0.2，30)、(0.5，50)、(0.4，30) …

Test:

X=(0.85, 98), Y=?



回归：输出是实数

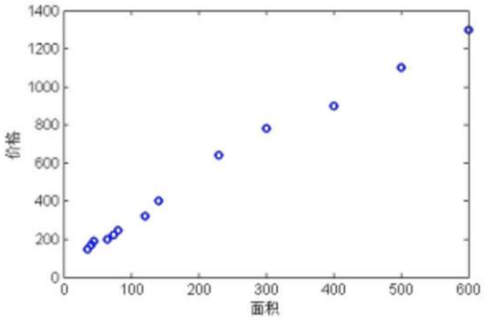
Regression: Y is continue

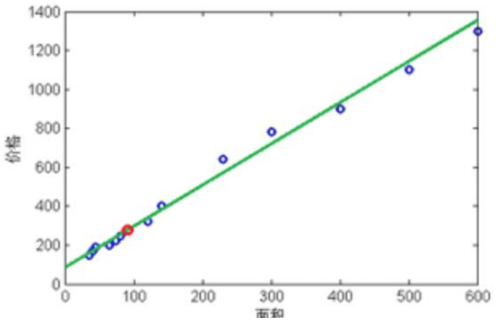
Y：房屋价钱（万元），值域 Y>=0.

X: x1=房屋面积 m^2.

Training Data:

35	150
40	170
45	190
65	200
74	224
80	245
120	320
140	400
230	640
300	780
400	900
500	1100
600	1300





$y=ax+b$

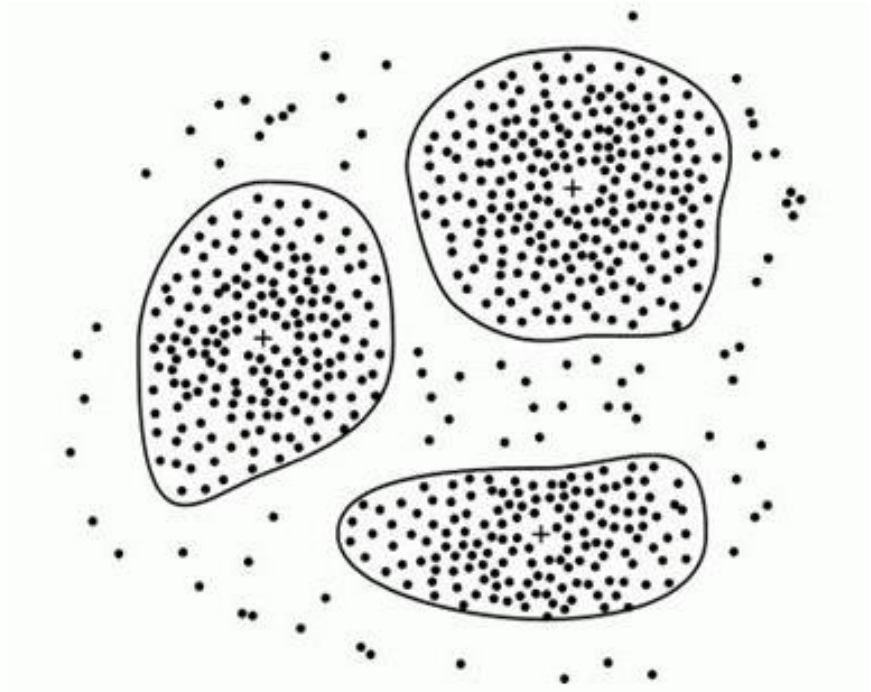
Test: X=90 Y= ?

AI DISCOVERY

机器学习方法

AI DISCOVERY

无监督学习（unsupervised learning）：没有标注的训练数据集，需要根据样本间的统计规律对样本集进行分析，常见任务如**聚类**等。

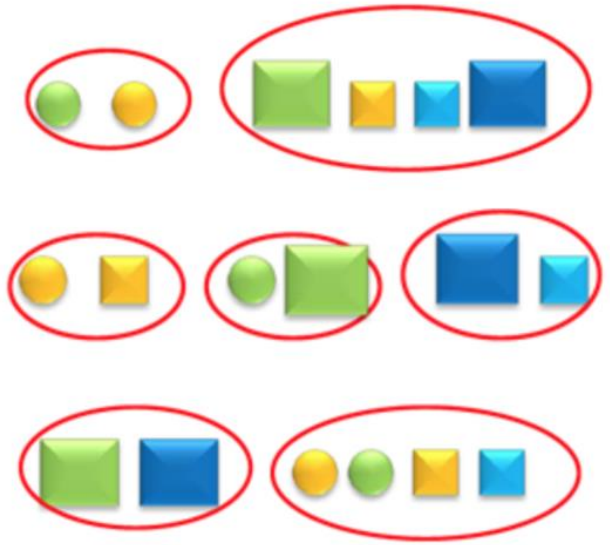


Clustering:

X: (颜色，形状，大小)

Data:

For all the data, Y=?

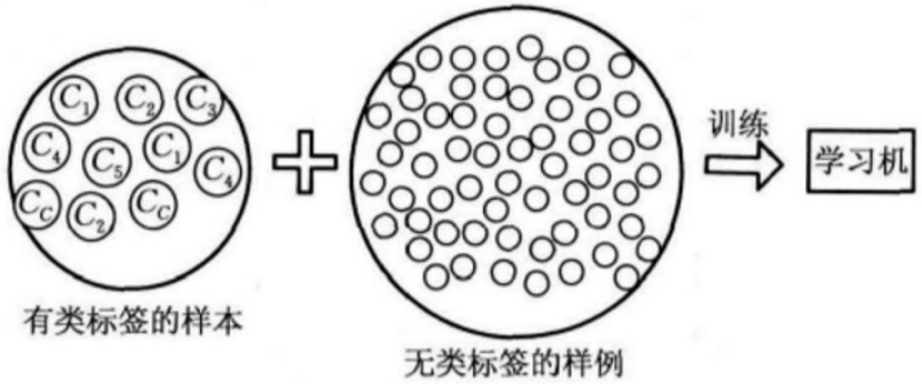


AI DISCOVERY

机器学习方法

AI DISCOVERY

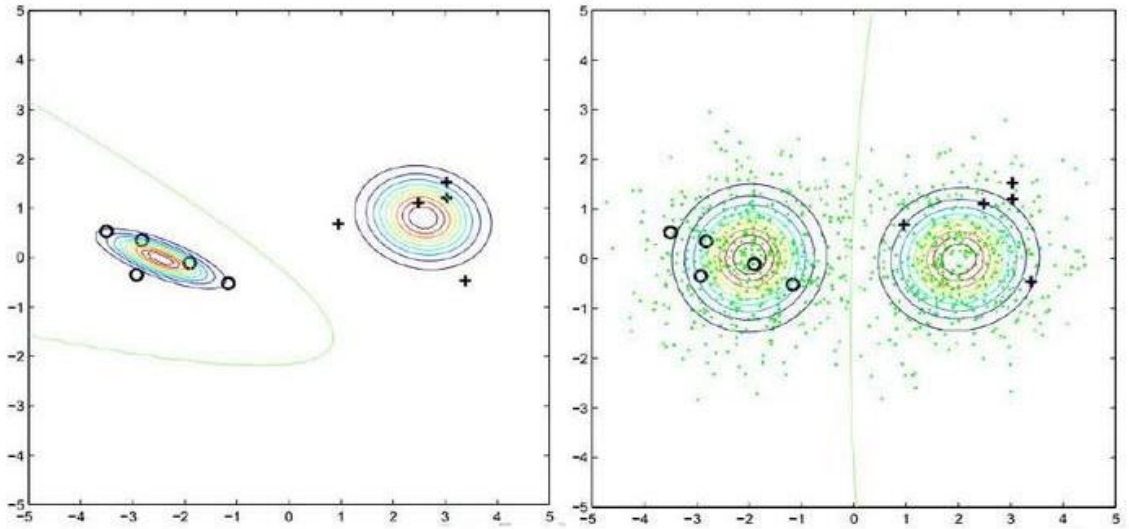
半监督学习（Semi-supervised learning）：结合**（少量的）标注训练数据**和**（大量的）未标注数据**来进行数据的分类学习。



两个基本假设：

- **聚类假设：**处在相同聚类中的样本示例有较大的可能拥有相同的标记。

- **流形假设：**处于一个很小的局部区域内的样本示例具有相似的性质，因此，其标记也应该相似。



3:0！AlphaGo 完胜柯洁

柯洁：中国围棋职业九段棋手，世界排名第一

AlphaGo：Google DeepMind 开发的机器学习围棋程序

AlphaGo使用**蒙特卡罗树**搜索与两个**深度神经网络**相结合的方法，其中一个是以估值网络来评估大量的选点，而以走棋网络来选择落子。

无人驾驶车队亮相2018春晚

百度发布“**Apollo（阿波罗）**”软件平台，向汽车行业及自动驾驶领域提供一套完整的平台。

无人驾驶主要包括三个环节：感知、**决策**、和控制

核心技术：异步多传感器同步+深度**数据融合**

机器学习面临的难题与挑战

- **数据稀疏性：** 训练一个模型，需要大量（标注）数据，但是数据往往比较稀疏。
- **高数量和高质量标注数据需求：** 获取标定数据需要耗费大量人力和财力。而且，人会出错，有主观性。
- **冷启动问题：** 对于一个新产品，在初期，要面临数据不足的冷启动问题。
- **泛化能力问题：** 训练数据不能全面、均衡的代表真实数据。

机器学习面临的难题与挑战

模型

策略

算法

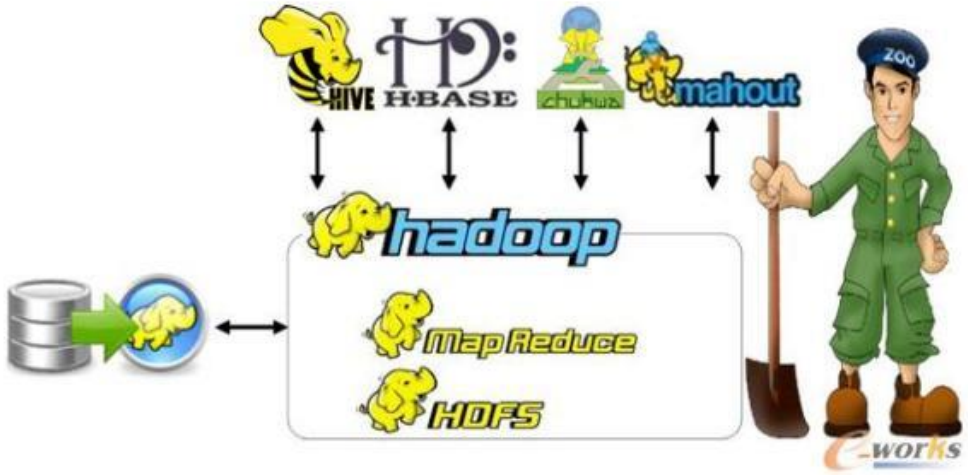
- **模型抽象困难：** 总结归纳实际问题中的数学表示非常困难。
- **模型评估困难：** 在很多实际问题中，很难形式化的、定量的评估一个模型结果的好坏。
- **寻找最优解困难：** 要解决的实际问题非常复杂，将其形式化后的目标函数也非常复杂，往往在目前还不存在一个有效的算法能找到目标函数的最优值。

机器学习面临的难题与挑战

AI DISCOVERY

◆**Scalability**是互联网的核心问题之一。搜索引擎索引的重要网页超过 100 亿: 如果1台机器每秒处理1000 网页，需要至少100天。

◆**速度**是互联网核心的用户体验。线下模型训练可以花费很长时间：比如，Google 某个模型更新一次需要几千台机器，大约训练半年时间。但是，线上使用模型的时候要求一定要 “快，实时 (real-time)”



◆**online learning**：互联网每时每刻都在产生大量新数据，要求模型随之不停更新，所以online learning是机器学习的一个重要研究方向。

AI DISCOVERY

数据集拆分

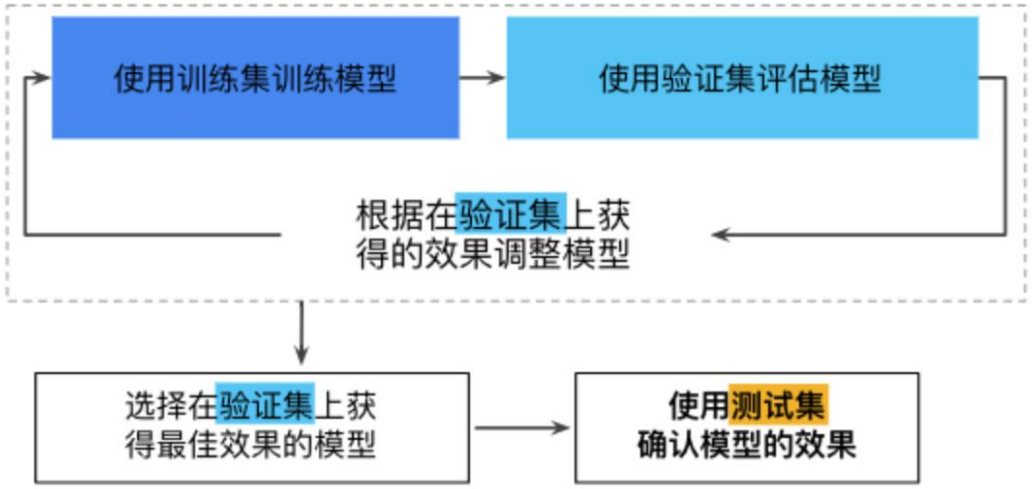
AI DISCOVERY

口 机器学习中将数据划分为3份

① **训练数据集（train dataset）**：用来构建机器学习模型

② **验证数据集（validation dataset）**：辅助构建模型，用于在构建过程中评估模型，提供无偏估计，进而调整模型参数

③ **测试数据集（test dataset）**：用来评估



口 常用拆分方法

✓ **留出法**（Hold-Out）：直接将数据集划分为互斥的集合，如通常选择 70% 数据作为训练集，30% 作为测试集。需要注意的是保持划分后集合数据分布的一致性，避免划分过程中引入额外的偏差而对最终结果产生影响。

✓ **K-折交叉验证法**：将数据集划分为 k 个大小相似的互斥子集，并且尽量保证每个子集数据分布的一致性。这样，就可以获取 k 组训练 - 测试集，从而进行 k 次训练和测试，k通常取值为10。

AI DISCOVERY

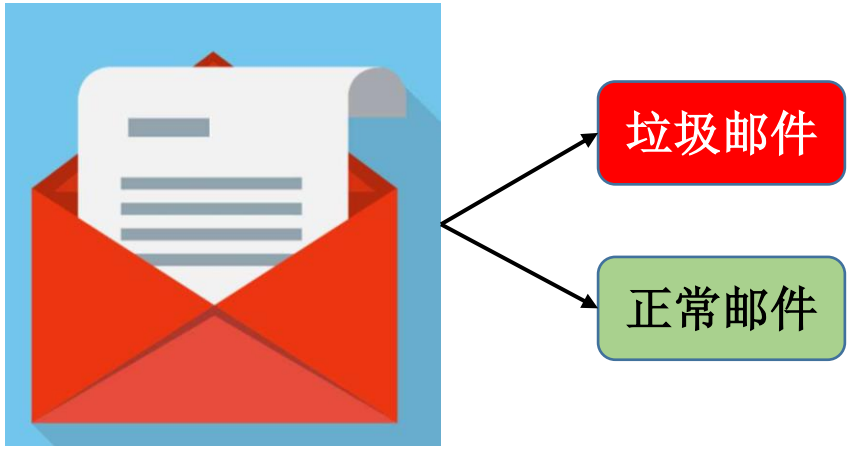
分类问题

AI DISCOVERY

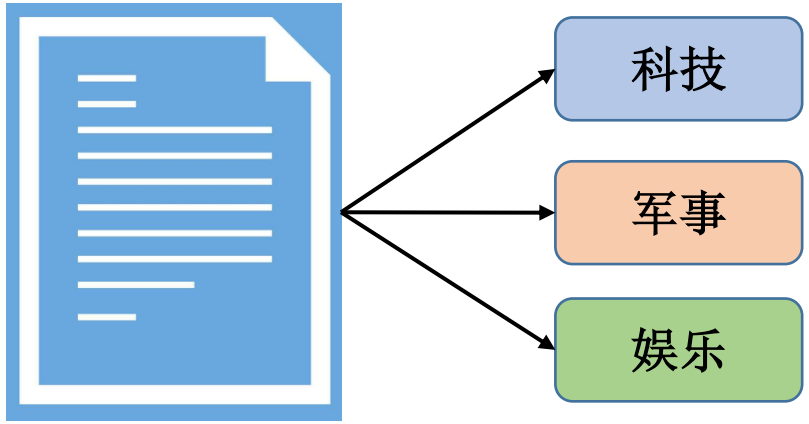
分类问题是监督学习的一个核心问题，它从数据中学习一个分类决策函数或分类模型(分类器（classifier）)，对新的输入进行输出预测，输出变量取有限个离散值。

分类在我们日常生活中很常见

✓ 二分类问题



✓ 多分类问题



核心算法

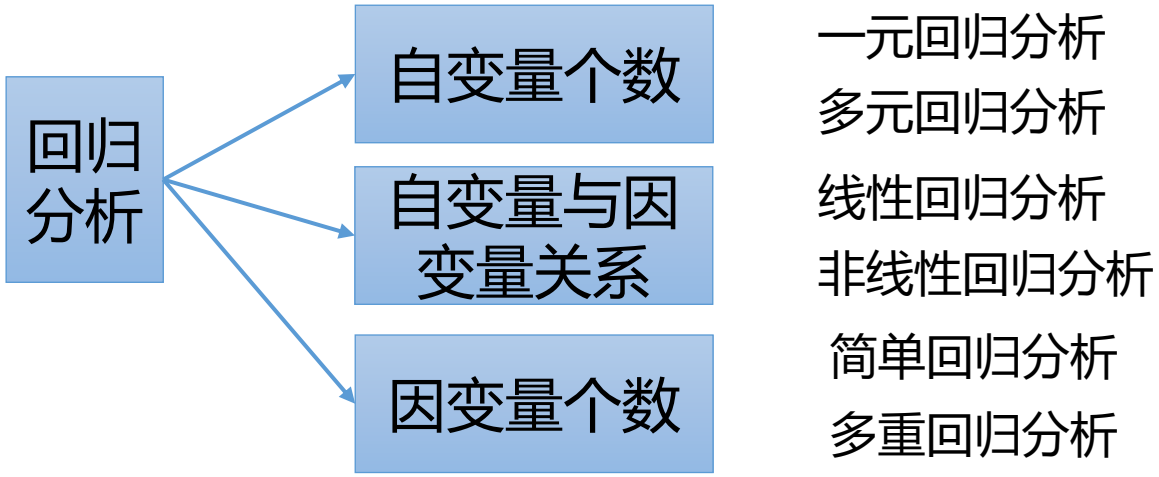
✓ 决策树、贝叶斯、SVM、逻辑回归

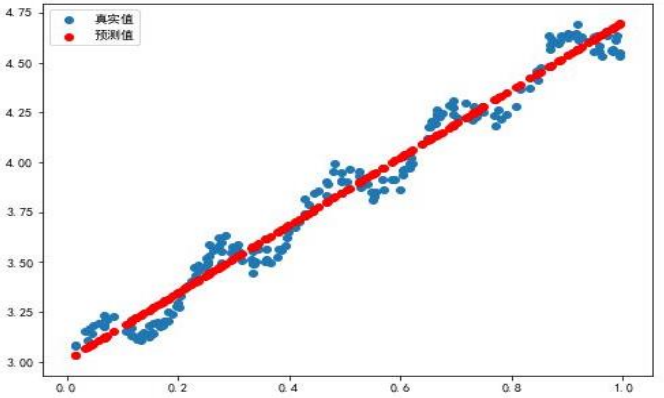
AI DISCOVERY

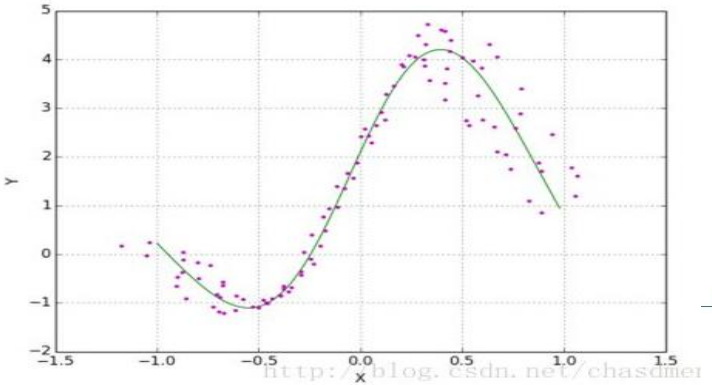
回归问题

AI DISCOVERY

回归分析用于预测输入变量（自变量）和输出变量（因变量）之间的关系，特别是当输入变量的值发生变化时，输出变量值随之发生变化。



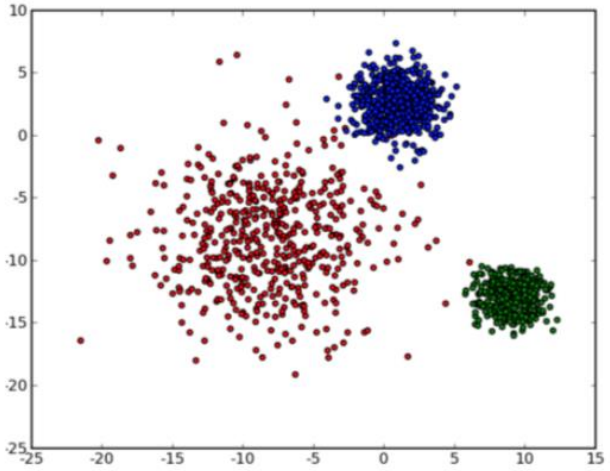


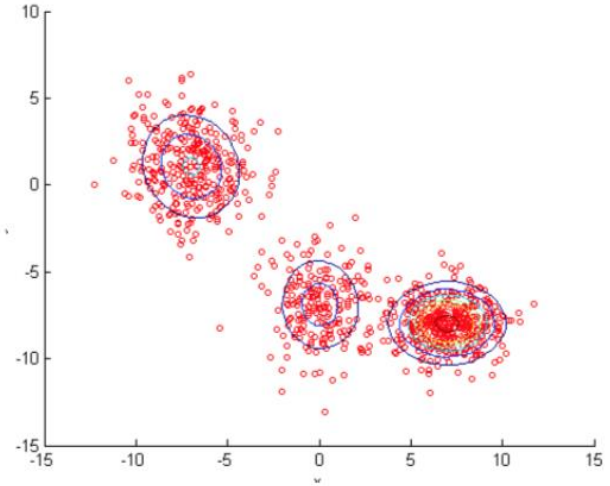


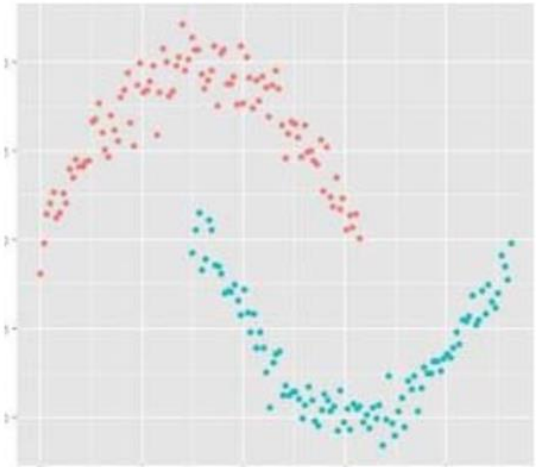
为什么叫回归？：达尔文表兄弟Francis Galton发明的。

聚类问题

聚类问题是无监督学习的问题，算法的思想就是“物以类聚，人以群分”。聚类算法感知样本间的相似度，进行类别归纳，对新的输入进行输出预测，输出变量取有限个离散值。







- ✓ 可以作为一个单独过程，用于寻找数据内在的分布结构
- ✓ 可以作为分类、稀疏表示等其他学习任务的前驱过程