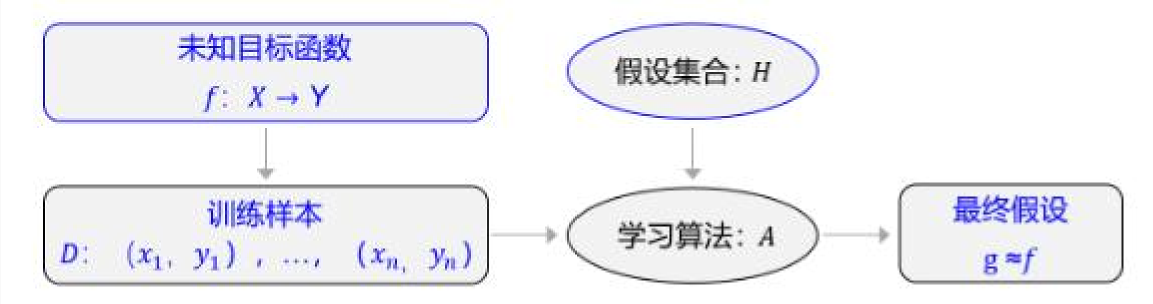
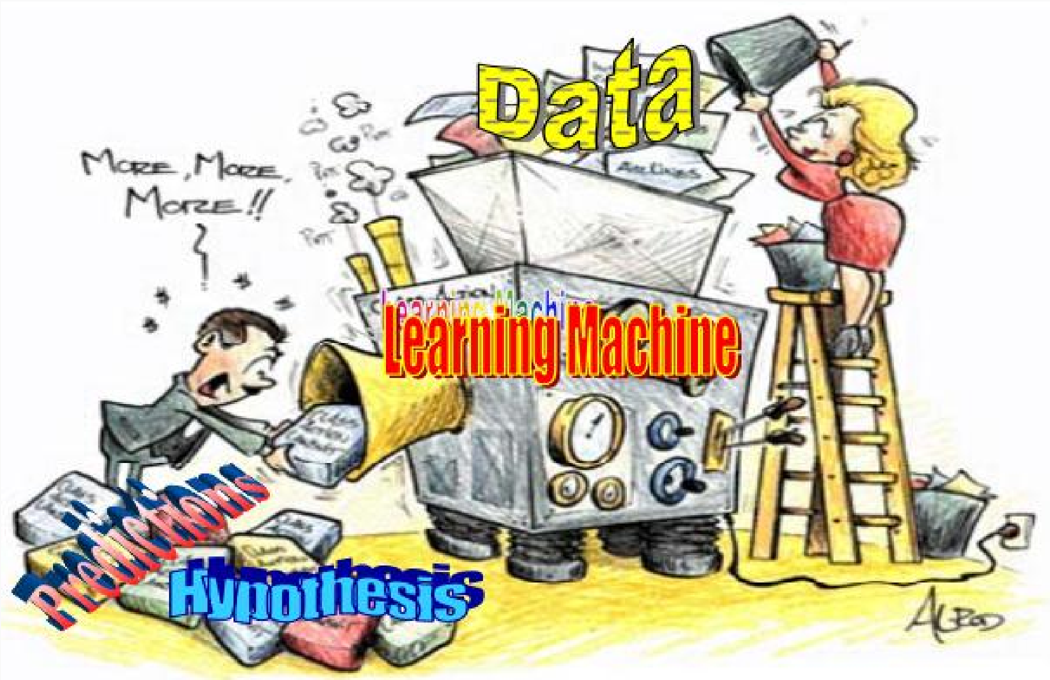
# 第一节 机器学习

## 机器学习发展历史


**人工智能**是基于数据处理来做出决策和预测

**机器学习**利用经验来改善自身的性能



# 第一节 机器学习

## 机器学习基础概念

| 术语 terms | 房价预测案例 |
| --- | --- |
| 数据（raw data） | 房子1，房子2，房子3，房子4 |
| 特征 | 卧室个数，地理位置，朝向，…/不具有可解释性 |
| 模型 | 支持向量机，神经网络，… |
| 参数 | 可学习的，根据目标函数进行优化的 |
| 模型输出 | 预测的房价 |
| 目标 | 真实的房价 |
| 目标函数（损失函数） | |
| 训练集 | 用于训练模型的数据集 |
| 测试集 | 用于评估已训练模型的数据集 |
| 泛化 | 衡量模型对未见过数据样本的分类能力 |

# 第一节 机器学习

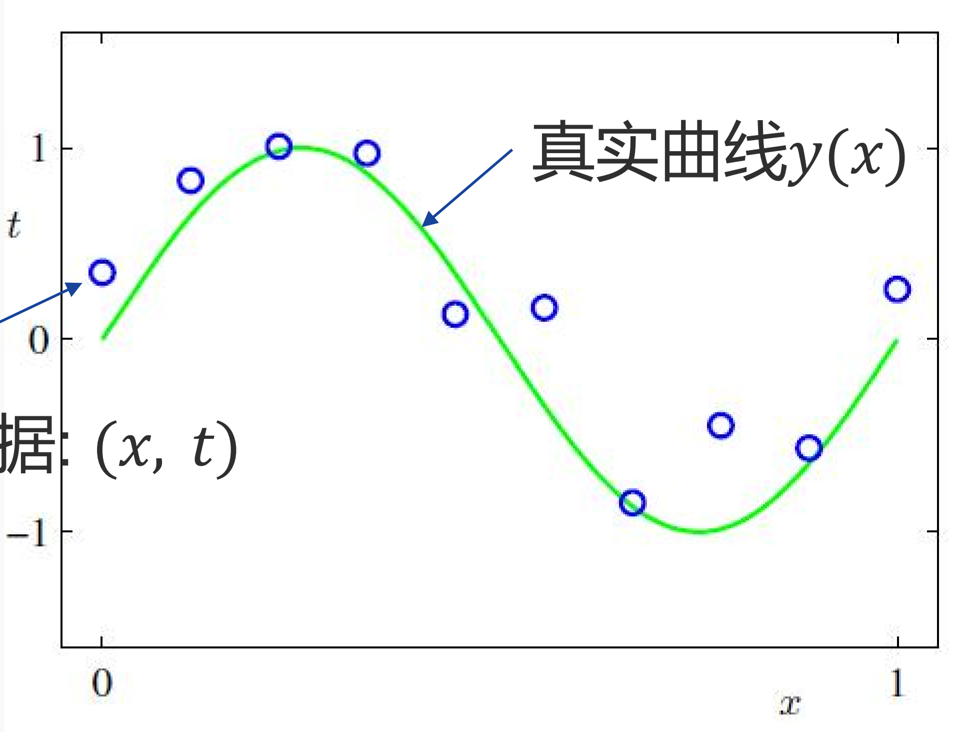
## 机器学习基础概念

| 术语 terms | 案例 |
| --- | --- |
| 超参数 | 模型外部的变量 |
| 误差 | 模型输出与目标的差值 |
| 欠拟合 | 在训练集、测试集上均表现不佳 |
| 过拟合 | 在训练集上表现很好，在测试集上表现不佳 |
| 验证集 | 用于模型自身性能评估及超参数调整 |
| 性能度量 | 准确率，错误率，查全率，查准率 |
| 偏差 | 模型在样本上的输出与真实值之间的误差，衡量模型拟合训练数据的能力 |
| 方差 | 模型每一次输出结果与模型输出期望之间的误差，衡量模型的稳定性 |
| 噪声 | 描述了在当前任务上任何学习算法所能达到的期望泛化误差的下界，即刻画了学习问题本身的难度 |

# 第一节 机器学习

## 机器学习简单例子

真实曲线$y(x)$




观察数据: $(x,\ t)$

学习目标

$$y(x,\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

# 第一节 机器学习

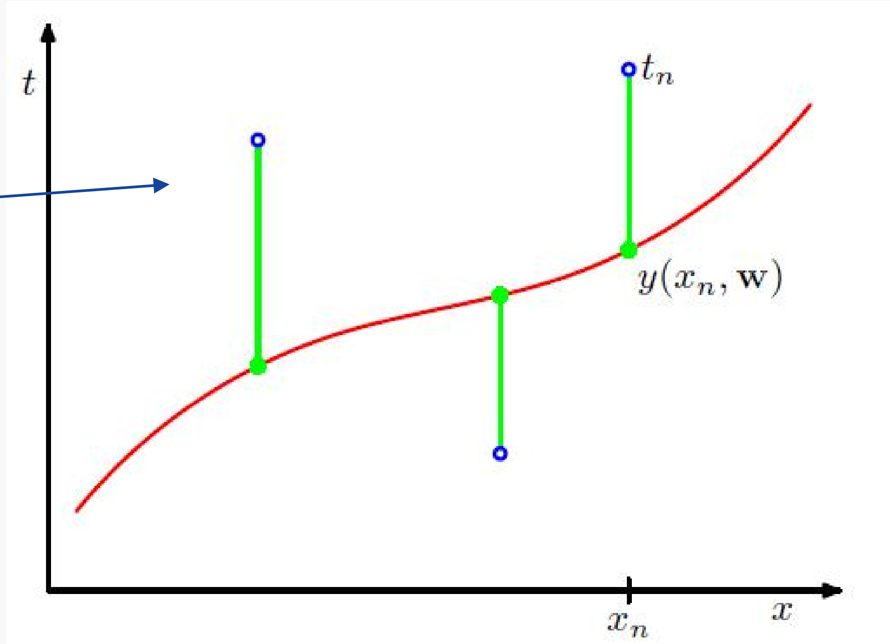
## 机器学习简单例子

$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

- 问题1: 如何学习参数 **w**

拟合偏差

$$y(x_n, \mathbf{w}) - t_n$$




**w** 应该最小化拟合错误:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

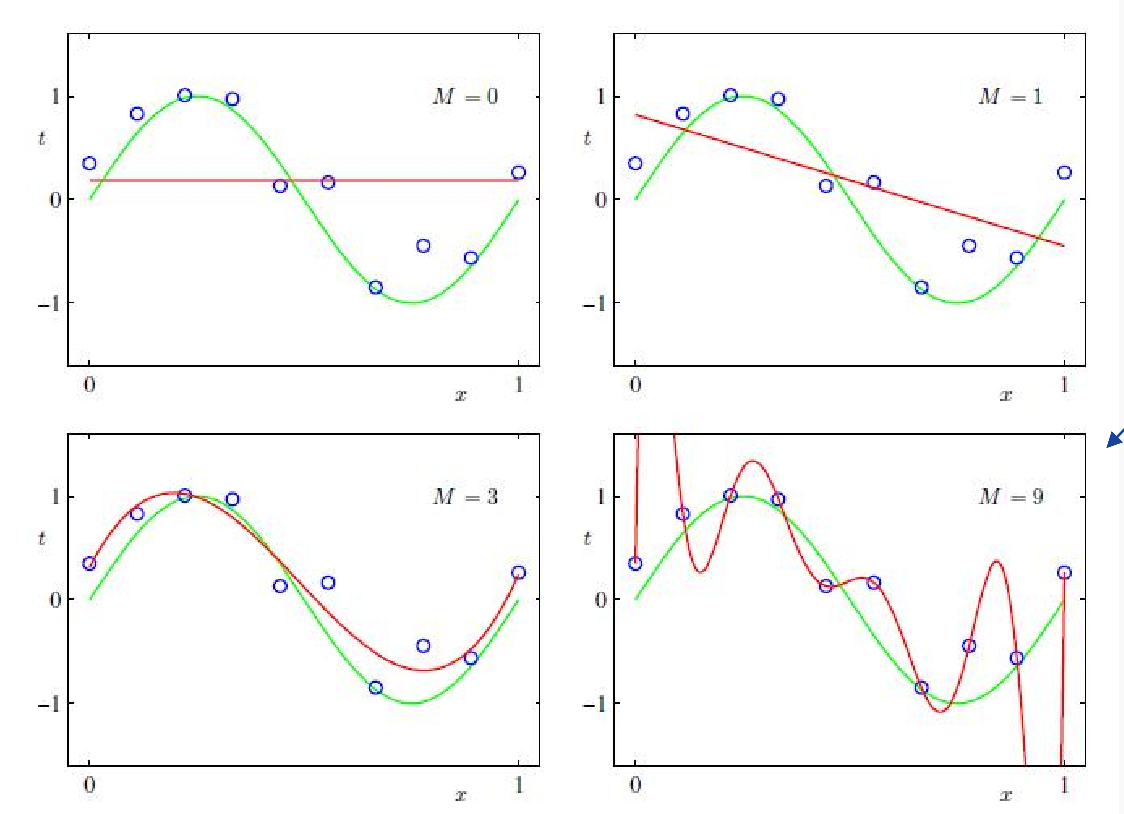
# 第一节 机器学习

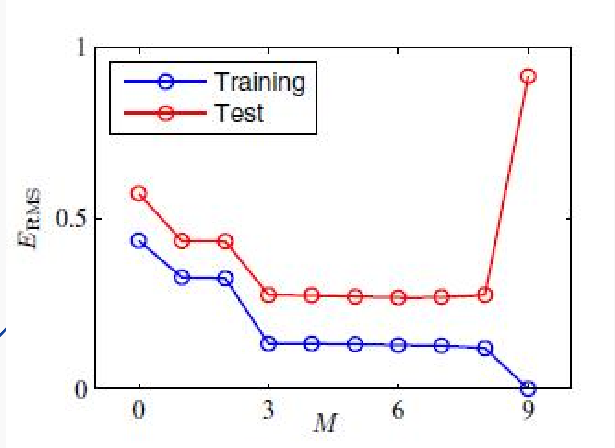
## 机器学习简单例子

$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

- 问题2: 如何决定多项式阶次$M$

最小化误差 = 最好的模型 ？




过拟合（Overfitting）

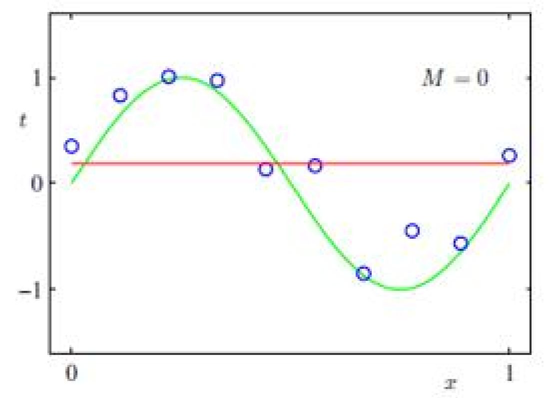
⇒ 泛化能力差

# 第一节 机器学习

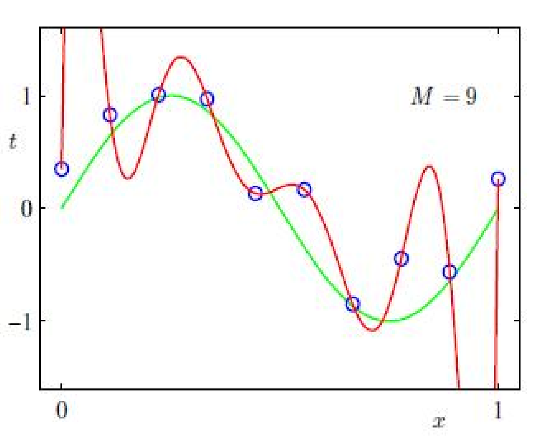
## 机器学习简单例子

$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

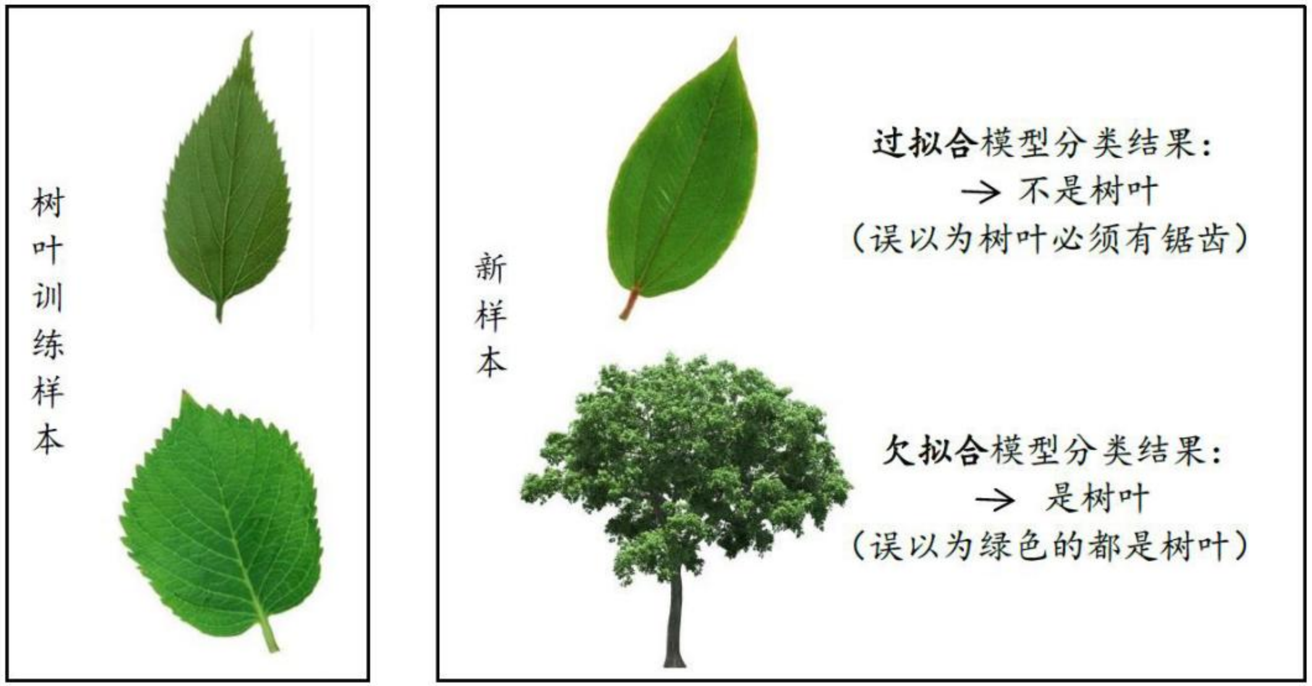
- 问题2: 如何决定多项式阶次$M$




欠拟合




过拟合

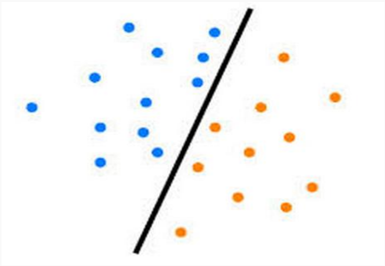



过拟合、欠拟合的直观类比

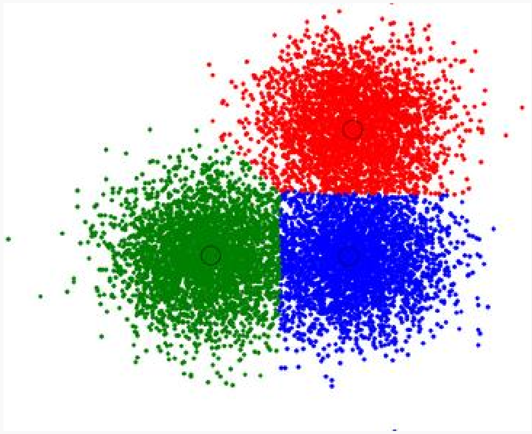
# 第一节 机器学习

## 机器学习主要分类

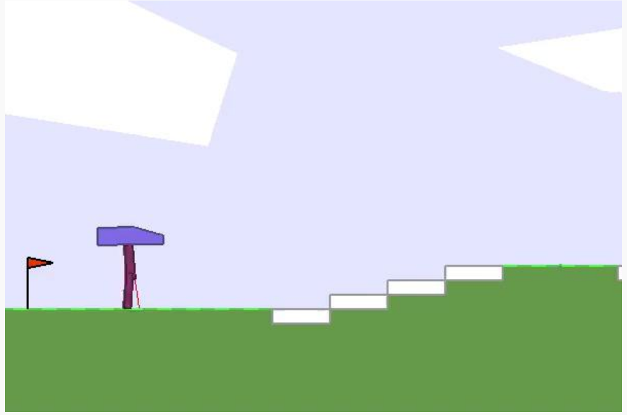
- 监督学习（Supervised learning）
  - 训练数据有目标向量（标签）
  - 分类、回归…



- 非监督学习（Unsupervised learning）
  - 训练数据没有目标向量（标签）
  - 聚类、密度估计、可视化…



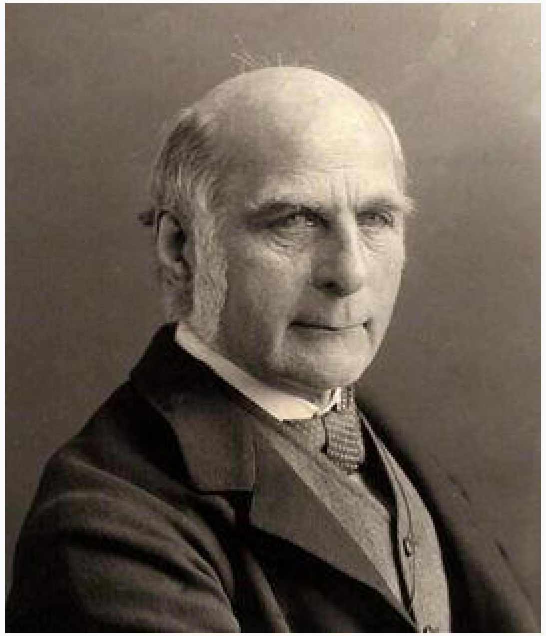
- 强化学习（Reinforcement learning）
  - 和环境存在交互
  - situation, action, reward



# 第一节 机器学习

## 监督学习：回归与分类

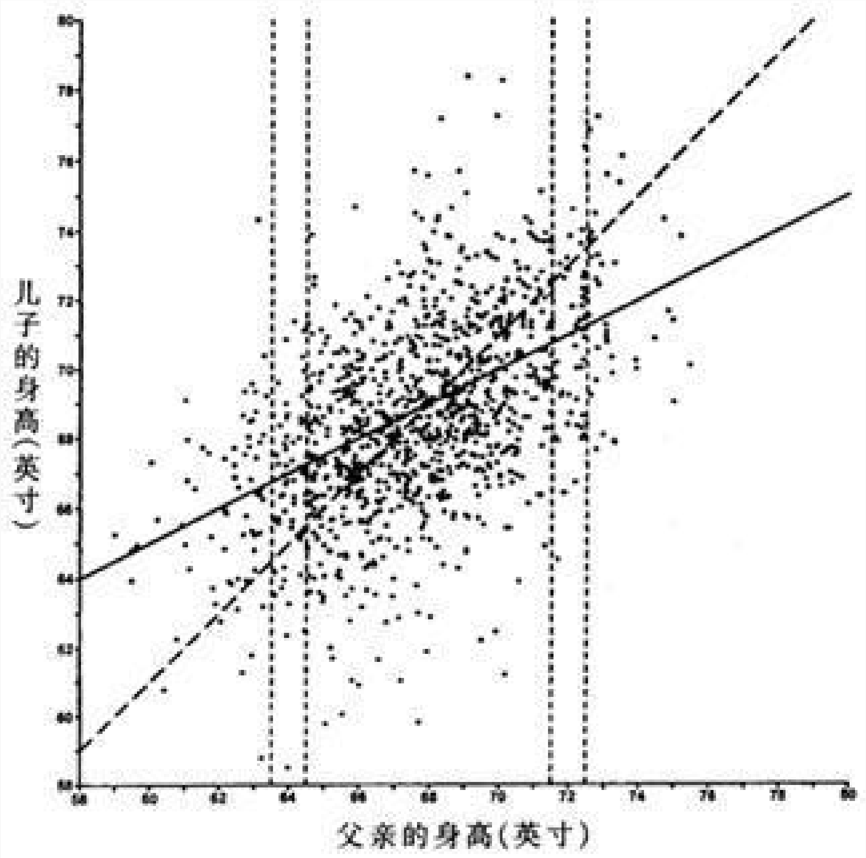
**Francis Galton，英国生物学家，他研究了父母身高与子女身高之间关系后得出，若父母身高高于平均大众身高，则其子女身高倾向于倒退生长，即会比其父母身高矮一些而更接近于大众平均身高。若父母身高小于平均身高，则其子女身高倾向于向上生长，以更接近于大众平均身高。此现象，被Galton称之为回归现象，即regression.**



弗朗西斯·高尔顿（Francis Galton，1822年2月16日－1911年1月17日）

# 第一节 机器学习

## “ 监督学习：回归与分类




1855年， 高尔顿、卡尔•皮尔逊
《遗传的身高向平均数方向的回归》

**Y= 0.8567+0.516*X**

**从左图中可以看到，后代的身高倾向于“回归”到一个平均值。**

# 第一节 机器学习

## 监督学习：回归与分类

**例：房价估计**

训练样本

| Size (feet²) | Number of bedrooms | Number of floors | Age of home (years) | Price ($1000) |
|---|---|---|---|---|
| 2104 | 5 | 1 | 45 | 460 |
| 1416 | 3 | 2 | 40 | 232 |
| 1534 | 3 | 2 | 30 | 315 |
| 852 | 2 | 1 | 36 | 178 |
| … | … | … | … | … |

测试样本

| Size (feet²) | Number of bedrooms | Number of floors | Age of home (years) | Price ($1000) |
|---|---|---|---|---|
| 1500 | 3 | 2 | 30 | ? |

# 第一节 机器学习

## “ 监督学习：回归与分类

**例：房价估计**

训练样本

| Size (feet²) $x_1$ | Number of bedrooms $x_2$ | Number of floors $x_3$ | Age of home (years) $x_4$ | Price ($1000) $y$ |
|---|---|---|---|---|
| 2104 | 5 | 1 | 45 | 460 |
| 1416 | 3 | 2 | 40 | 232 |
| 1534 | 3 | 2 | 30 | 315 |
| 852 | 2 | 1 | 36 | 178 |
| … | … | … | … | … |

$N$

**监督学习**

**回归问题**

$N$：**训练样本个数**

x：**输入变量/“特征”**

y：**输出变量/目标变量**

# 第一节 机器学习

## 监督学习：回归与分类

- 给定数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$

其中 $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$ ， $y_i \in \mathbb{R}$

- 线性回归（linear regression）：学得一个线性模型以尽可能准确地预测实值输出标记

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

$\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$ 是描述特征，其中 $x_i$ 是 $\boldsymbol{x}$ 的第 $i$ 类特征取值

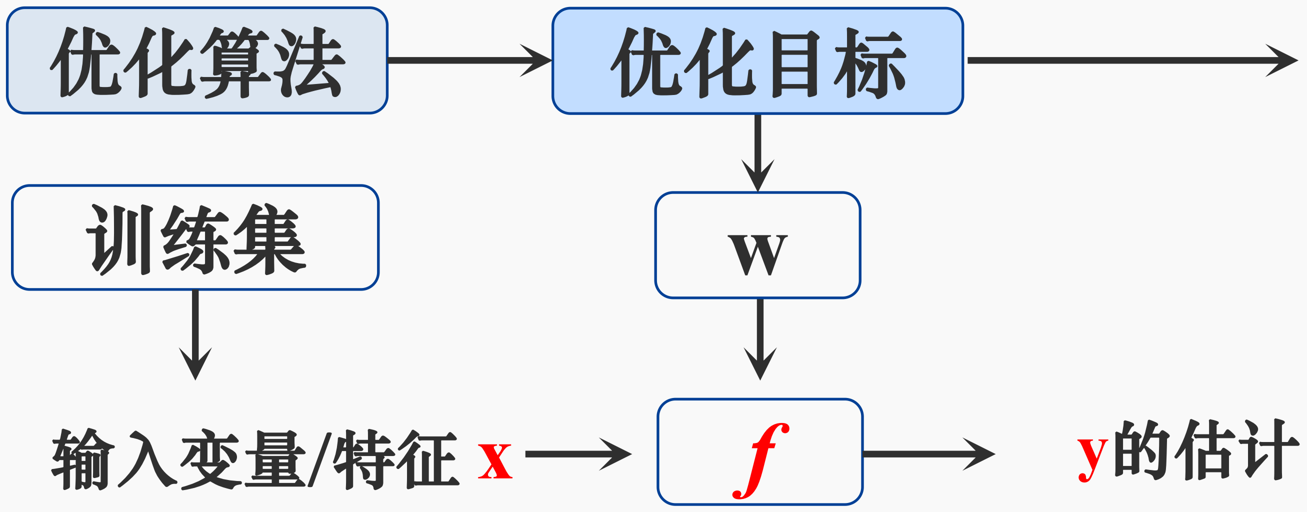
- 向量形式

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b$$

$$\boldsymbol{w} = (w_1; w_2; \ldots; w_d)$$

# 第一节 机器学习

## 监督学习：回归与分类




$$
\begin{aligned}
f(\mathbf{x}) &= w_0 + w_1 x_1 + \ldots + w_m x_m \\
&= w_0 + \sum_{i=1}^{m} w_i x_i \\
&= \sum_{i=0}^{m} w_i x_i \qquad \boxed{x_0 = 1}
\end{aligned}
$$

- 最小化均方误差

$$
\begin{aligned}
(w^*, b^*) &= \underset{(w,b)}{\arg\min} \sum_{i=1}^{m} \left(f\left(x_i\right) - y_i\right)^2 \\
&= \underset{(w,b)}{\arg\min} \sum_{i=1}^{m} \left(y_i - w x_i - b\right)^2
\end{aligned}
$$

- 梯度下降优化算法

$$
w_j^t = w_j^{t-1} - \alpha \frac{\partial}{\partial w_j} J(\mathbf{w})
$$

# 第一节 机器学习

## 监督学习：回归与分类

- 线性回归： $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$
- 二分类： $y \in \{0, 1\}$
- 如何建立分类与线性回归的联系？

- 最理想的函数——单位阶跃函数

$$y = \begin{cases} 0, & z < 0; \\ 0.5, & z = 0; \\ 1, & z > 0, \end{cases}$$

**缺点：不连续**

- 单调可微、任意阶可导——逻辑函数（logistic function）

$$y = \frac{1}{1 + e^{-z}}$$

$$y = \frac{1}{1 + e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)}}$$