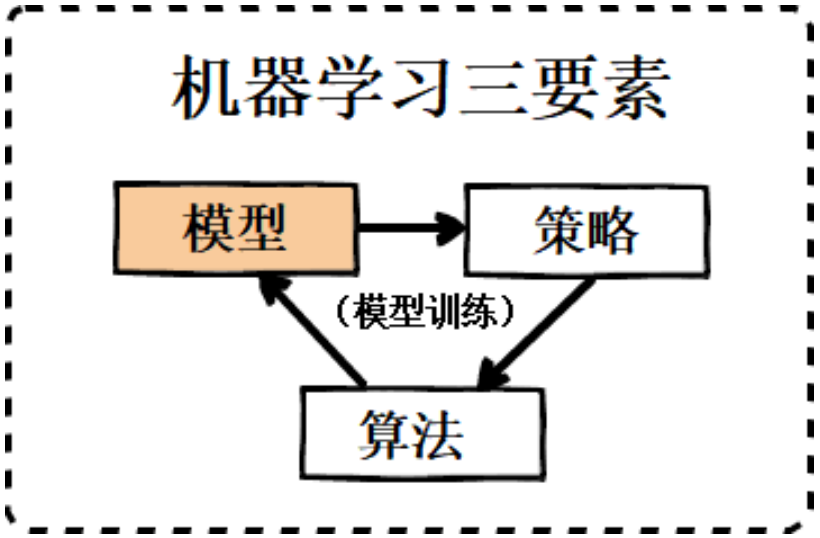
学习算法、线性回归

学习算法三要素

1.模型（Model）：模型指的是数据和预测之间的数学关系。模型中有参数，不同的参数组合构成了不同过得模型，这使得可能的模型数量有多个甚至是无限多。这些不同的模型构成了所谓的“假设空间”。这意味着，假设空间包含了所有可能的模型，每个模型都是对现实问题的一种潜在解释。

2.策略（Strategy）：策略定义了一个“好”的模型的标准。这个标准一般是用损失函数来描述的，损失函数衡量的是模型预测值与实际值之间的差异。一个“好”的模型是指在其损失函数上表现最佳的模型，即错误最小。

3.算法（Algorithm）：算法是指在假设空间中找到最优模型的方法。这涉及到求解一系列数学问题，以确定模型参数的最佳值。



机器学习算法举例

问题阐述：想通过某个房产所在的位置（变量x_1），小区（变量x_2），楼层（变量x_3），是否精装修（变量x_4）来预估这个房产的价值（y）。

模型：基于房产价值与其属性之间可能存在的线性关系的假设，我们将构建一个线性回归模型来预测房产价值，即$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$，其中$w_0$，$w_1$，$w_2$，$w_3$，$w_4$是模型参数。

策略：在这个例子中，一个“好”的模型的标准是预测房价与实际房价的差异越小越好。于是策略就是找到一组参数，它能够最小化预测房价与实际房价之间的差异，这可以通过最小化均方误差来实现。$\sum_{i=1}^{n}(Y^{(i)} - y^{(i)})^2$

算法：为了找到这一组参数，我们可以使用解析方法（通过精确的数学演算找到问题的精确解的方法）或者数值方法（数值方法通过近似和迭代来求解问题，依赖于计算机进行大量计算，以逼近问题的解）来最小化均方误差。

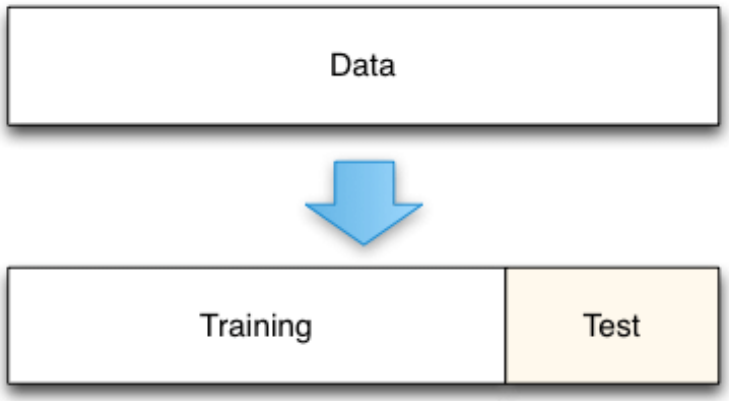
训练集/测试集

- 数据集主要被分为两个部分：训练集、测试集。

1.训练集用于训练模型。在这个阶段，模型学习识别数据中的模式和关系。通过调整模型参数，模型尝试最大限度地减少误差，并在这些已知数据上表现出良好的拟合度。

2.测试集用来评估模型的**泛化能力**。泛化是指模型对新、未见过的数据的处理能力。测试集是独立于训练过程的，因此提供了对模型性能的公正评估。

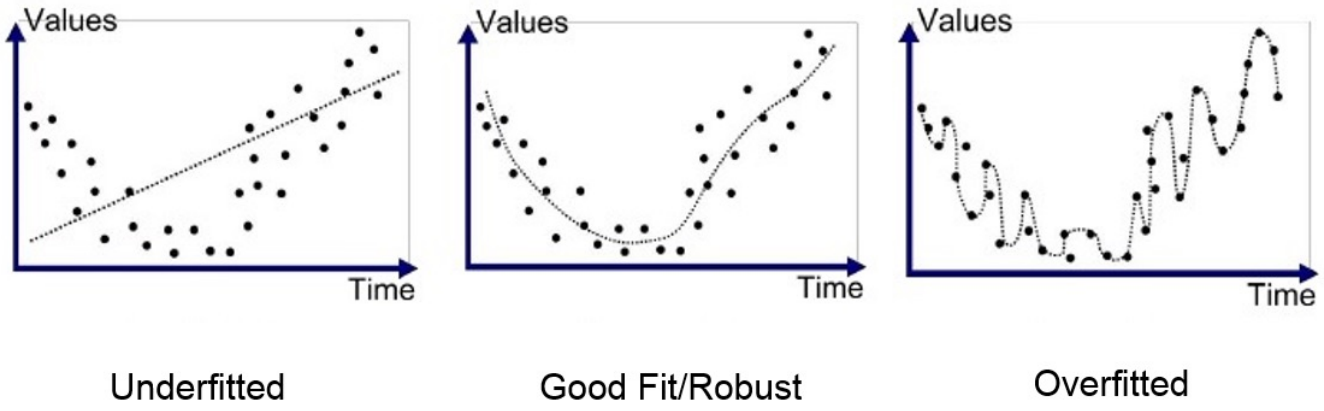
如果一个模型在训练集上表现良好但在测试集上表现不佳，则可能表明模型过拟合了训练数据，未能有效地泛化（到其他数据集上）。



过拟合/欠拟合

我们不仅希望模型能够在训练集上获得较低的误差，还期望模型能够在测试集上也拥有较低的误差（即具有良好的泛化能力）。**测试集上的误差，被称为泛化误差**，是评估模型泛化能力的重要指标。模型泛化能力不足时，通常表现为**过拟合（overfitting）**或**欠拟合（underfitting）**的问题。

- **过拟合**发生在模型对训练数据学习得太过彻底，以至于它开始捕捉到数据中的噪声和偶然的模式，而非真正的趋势。
- **欠拟合**则是发生在模型未能充分学习数据中的趋势和模式，导致在训练集和测试集上都表现不佳。
- 为什么会发生过拟合（**模型太过复杂**），欠拟合（**模型太过简单**）。



特征向量/空间/矩阵

1.特征向量：在机器学习中，每个数据点通常由一系列特征（或属性）表示。这些特征被组合成一个向量，称为特征向量。例如，

2.特征矩阵：当我们将数据集中的所有样本的特征向量汇集起来，形成的矩阵就是特征矩阵。在这个矩阵中，每一行代表一个样本的特征向量，而每一列代表一个特定的特征。

3.特征空间：特征空间是一个抽象的概念，它指的是所有可能的特征向量所构成的空间。在这个空间中，每个维度代表一个特征。这个空间中都有一个唯一的位置，由其特征向量确定。

学习数据的模式和关系之前**通常都需要将数据表示为特征向量和特征矩阵的形式**，这样机器学习算法才可以有效地处理和分析数据

*在使用机器学习算法前，一般会将数据集转化为特征矩阵的形式

特征空间/特征矩阵

以房价预测的例子举例：

1. **特征向量：**每个房产都由一组特征表示：位置（x_1），所在小区（x_2），楼层高度（x_3），以及是否精装修（x_4）每个房产的这四个特征共同构成了一个特征向量（x_1，x_2，x_3，x_4）。
2. **特征矩阵：**收集了多个房产的数据后，可以将每个房产的特征向量作为特征矩阵的一行。在这个矩阵中，每一行代表一个房产，每一列代表一个特定的特征（如位置、小区等）。
3. **特征空间：**特征向量（x_1，x_2，x_3，x_4）所在的空间就是特征空间。在这个空间中，每个维度对应一个特征。在这个例子中，特征空间是一个四维空间，其中每一个维度代表一个房产特征。每个房产样本在特征空间中占据一个唯一的点，这个点由其特征向量确定。

Features（特征）

Samples（样本）

$\boldsymbol{x_1^{(1)}}$	$\boldsymbol{x_2^{(1)}}$	$\boldsymbol{x_3^{(1)}}$	$\boldsymbol{x_4^{(1)}}$
$\boldsymbol{x_1^{(2)}}$	$\boldsymbol{x_2^{(2)}}$	$\boldsymbol{x_3^{(2)}}$	$\boldsymbol{x_4^{(2)}}$
$\boldsymbol{x_1^{(3)}}$	$\boldsymbol{x_2^{(3)}}$	$\boldsymbol{x_3^{(3)}}$	$\boldsymbol{x_4^{(3)}}$
…	…	…	…
$\boldsymbol{x_1^{(n)}}$	$\boldsymbol{x_2^{(n)}}$	$\boldsymbol{x_3^{(n)}}$	$\boldsymbol{x_4^{(n)}}$

线性回归

线性模型

（1）线性模型通过属性的加权线性组合预测变量，即

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

其中 $\boldsymbol{x}$ 是特征向量，w_1，w_2，…，w_{d}等都是特征权重，b 是截距项。

（2）在矩阵符号中，上述模型可表示为

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

其中 $\boldsymbol{w}$是权重向量，x 是特征向量，T表示转置。

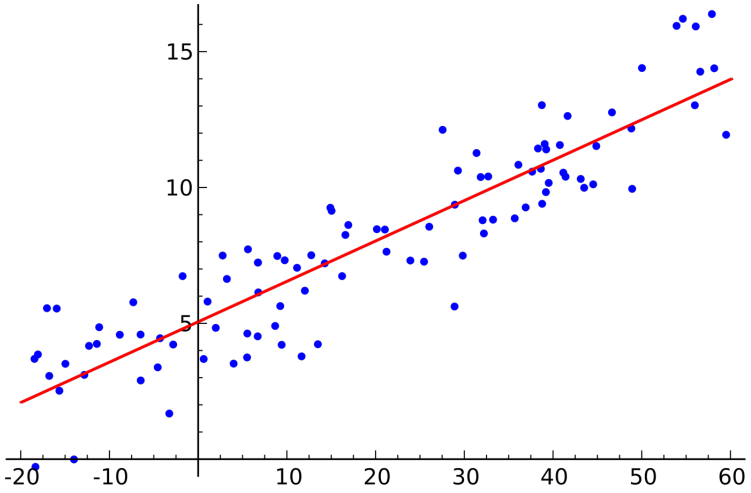
特点

- 线性模型以其简洁的形式而广受欢迎。尽管模型结构基础，但它是多种非线性模型（比如逻辑回归，多项式回归）的理论基础。
- 由于 ω 直观地表达了各属性在预测中的重要性，因此线性模型有很好的可解释性。例：$f_{\text{体感温度}}(x) = 0.1 x_{\text{风速}} + 0.4\, x_{\text{温度}} + 0.5\, x_{\text{湿度}} + \mathrm{b}$

线性回归模型

回归：根据输入特征所取的值，模型产生一个连续的预测值作为输出。

线性回归（Linear Regression）是一种预测技术，它通过建立输入与输出之间的线性关系模型来预测实际数值结果。



蓝点：数据真实值
红线上的点：预测模型给出的数据预测值

通常用于预测模型：

- 通常用于预测如产品销量、股票价格、行业趋势等连续变量
- 通过房产的位置，面积大小，楼层数等预测房产的价格
- ……

线性回归模型的**策略**

- 如何定义一个足够“好”的线性回归模型呢?

  一个好的线性回归模型，它的预测值和真实值一定比较接近。**均方误差（MSE）**可以反应预测值和真实值的差异，故可以用**均方误差**作为衡量线性回归模型“优/劣”的标准。其定义为

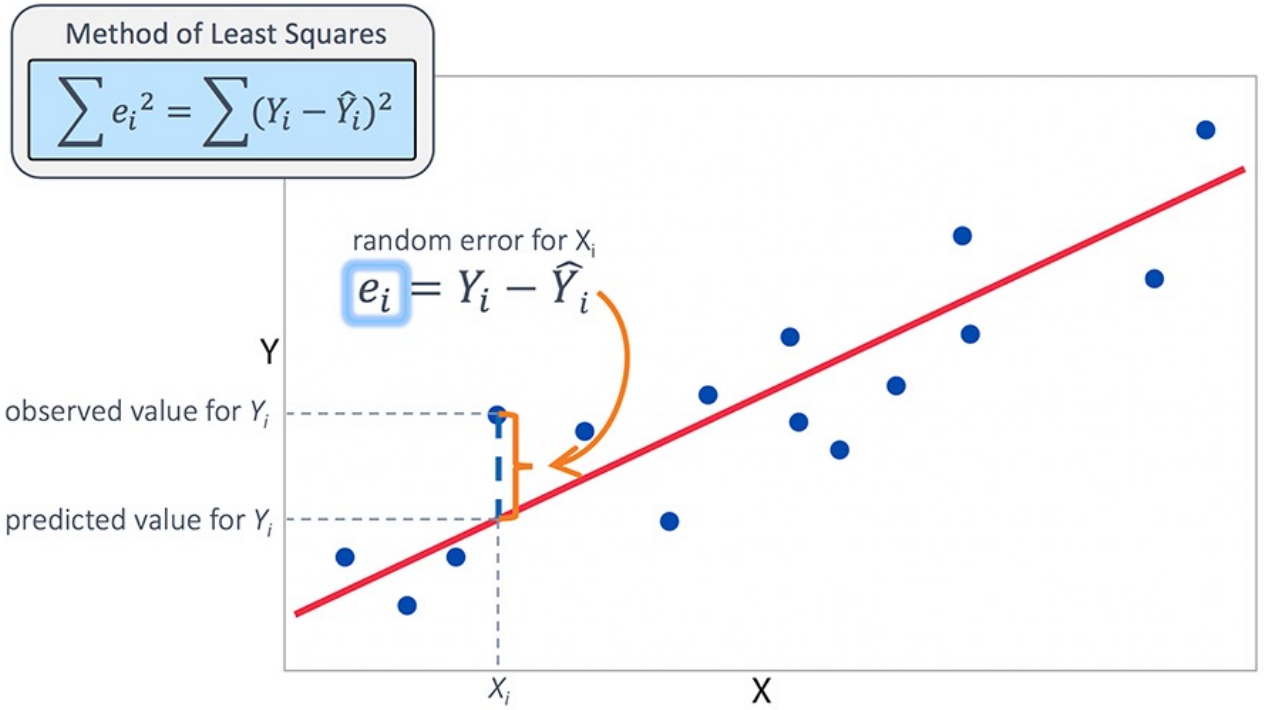
$$\frac{1}{n}\sum_{i=1}^{n}[f(x_i)-y_i]^2$$

  其中，n 是样本数量，$f(x_i)$ 是模型对第 i 个样本的预测值，y_i 是对应的实际值。

- **均方误差**的大小反映了模型预测的准确性，误差越小，模型的预测性能越好。

线性回归模型的**策略**

- “均方误差最小化”有实际的几何意义。在几何上表示为寻找最佳拟合直线，以便使数据集中每个点到该直线上的对应点的欧氏距离的平方和最小。



- 基于**均方误差最小化**的策略来进行模型求解的方法称 为"**最小二乘法**" (Least square methods)

线性回归模型的**算法**

线性回归模型：$Y = \alpha + \beta X + \varepsilon$ （模型的参数是？）

均方误差表达式为：$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

目标是找出使得该式最小的α和β

解析法

通过分别对 α 和 β 求导并设导数为零来求解模型参数，得到正规方程组：

$$\begin{cases} n\,\alpha + \sum_{i=1}^{n} x_i\,\beta = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,\alpha + \sum_{i=1}^{n} x_i^2\,\beta = \sum_{i=1}^{n} x_i y_i \end{cases}$$

线性回归**算法**

解析法

$$\begin{cases} n\,\alpha + \sum_{i=1}^{n} x_i\,\beta = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,\alpha + \sum_{i=1}^{n} x_i^2\,\beta = \sum_{i=1}^{n} x_i y_i \end{cases}$$

根据克莱姆法则，有

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \bar{y} - \bar{x}\hat{\beta}$$

多元线性回归**模型**

解析法

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

如果我们把数据集D写作

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} & 1 \\ x_{21} & x_{22} & \ldots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{md} & 1 \end{pmatrix}$$

把w和b吸收入向量形式$\hat{w} = (w_1, w_2, \ldots, w_d, b)^T$

于是多元线性回归的
一般形式为:

$$f(\boldsymbol{x}) = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} & 1 \\ x_{21} & x_{22} & \ldots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{md} & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ b \end{pmatrix} = \mathrm{X}\hat{w}$$

多元线性回归**策略**

解析法

真值　预测值

$$\hat{\boldsymbol{w}}^* = \arg\min_{\hat{\boldsymbol{w}}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$$

⇩

残差向量

⇩

残差的平方和（均方误差）

⇩

最小化均方误差（最小二乘法）

多元线性回归**算法**

解析法

$$E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$$

对 $E_{\hat{\boldsymbol{w}}}$ 进行求导，并设导数等于零，得

$$\frac{\partial E_{\hat{w}}}{\partial \hat{\mathbf{w}}} = \frac{\partial}{\partial \hat{\mathbf{w}}} (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^{\top} (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = 0$$

展开括号，得到：

$$\frac{\partial E_{\hat{w}}}{\partial \hat{\mathbf{w}}} = \frac{\partial}{\partial \hat{\mathbf{w}}} \left[\mathbf{y}^{\top}\mathbf{y} - \mathbf{y}^{\top}\mathbf{X}\hat{\mathbf{w}} - \hat{\mathbf{w}}^{\top}\mathbf{X}^{\top}\mathbf{y} + \hat{\mathbf{w}}^{\top}\mathbf{X}^{\top}\mathbf{X}\hat{\mathbf{w}} \right]$$

注意，$\mathbf{y}^{\top}\mathbf{y}$ 是一个常数，对 $\hat{\mathbf{w}}$ 的导数为零。现在我们对剩下的三项分别求导：

多元线性回归**算法**

解析法

$$\frac{\partial E_{\hat{w}}}{\partial \hat{\mathbf{w}}} = \frac{\partial}{\partial \hat{\mathbf{w}}} \left[\mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\hat{\mathbf{w}} - \hat{\mathbf{w}}^\top \mathbf{X}^\top \mathbf{y} + \hat{\mathbf{w}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{w}} \right]$$

1、 $\mathbf{y}^\top \mathbf{X}\hat{\mathbf{w}}$和 $\hat{\mathbf{w}}^\top \mathbf{X}^\top \mathbf{y}$ 实际上是等价的，因为它们都将产生相同的标量结果，因此它们的导数也相同，并且是 $\mathbf{X}^\top \mathbf{y}$

2、 $\hat{\mathbf{w}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{w}}$是一个二次项，其导数是 $2\mathbf{X}^\top \mathbf{X}\hat{\mathbf{w}}$

简化得到正规方程（Normal Equation）：

$$\mathbf{X}^\top \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$$

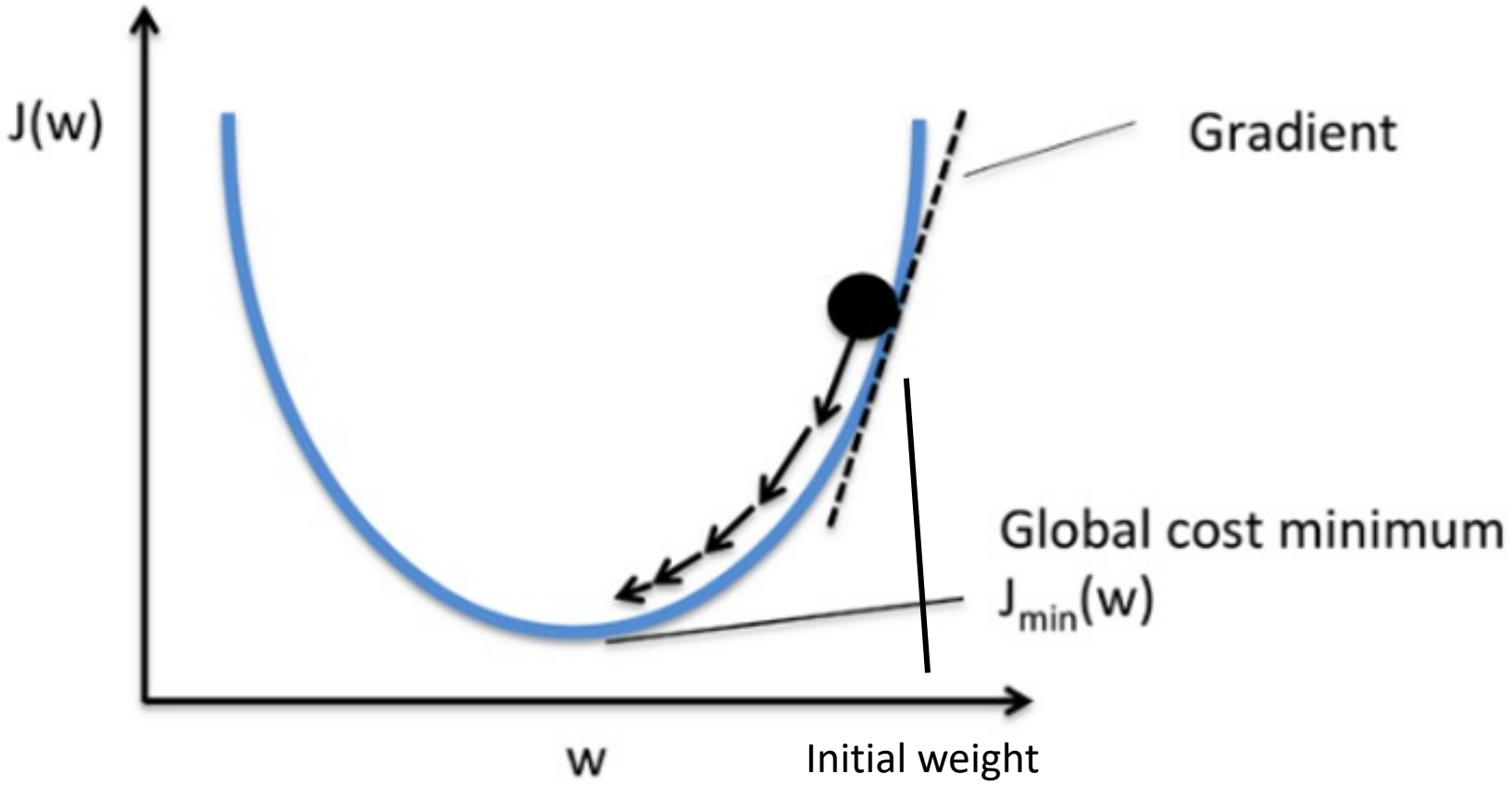
如果矩阵 $\mathbf{X}^\mathrm{T}\mathbf{X}$ 是可逆的，那么参数的最小二乘估计$\hat{\boldsymbol{w}}^*$可以通过下面的方程求得:

$$\hat{\boldsymbol{w}}^* = \left(\mathbf{X}^\mathrm{T}\mathbf{X}\right)^{-1} \mathbf{X}^\mathrm{T} y$$

线性回归**算法**

数值解法

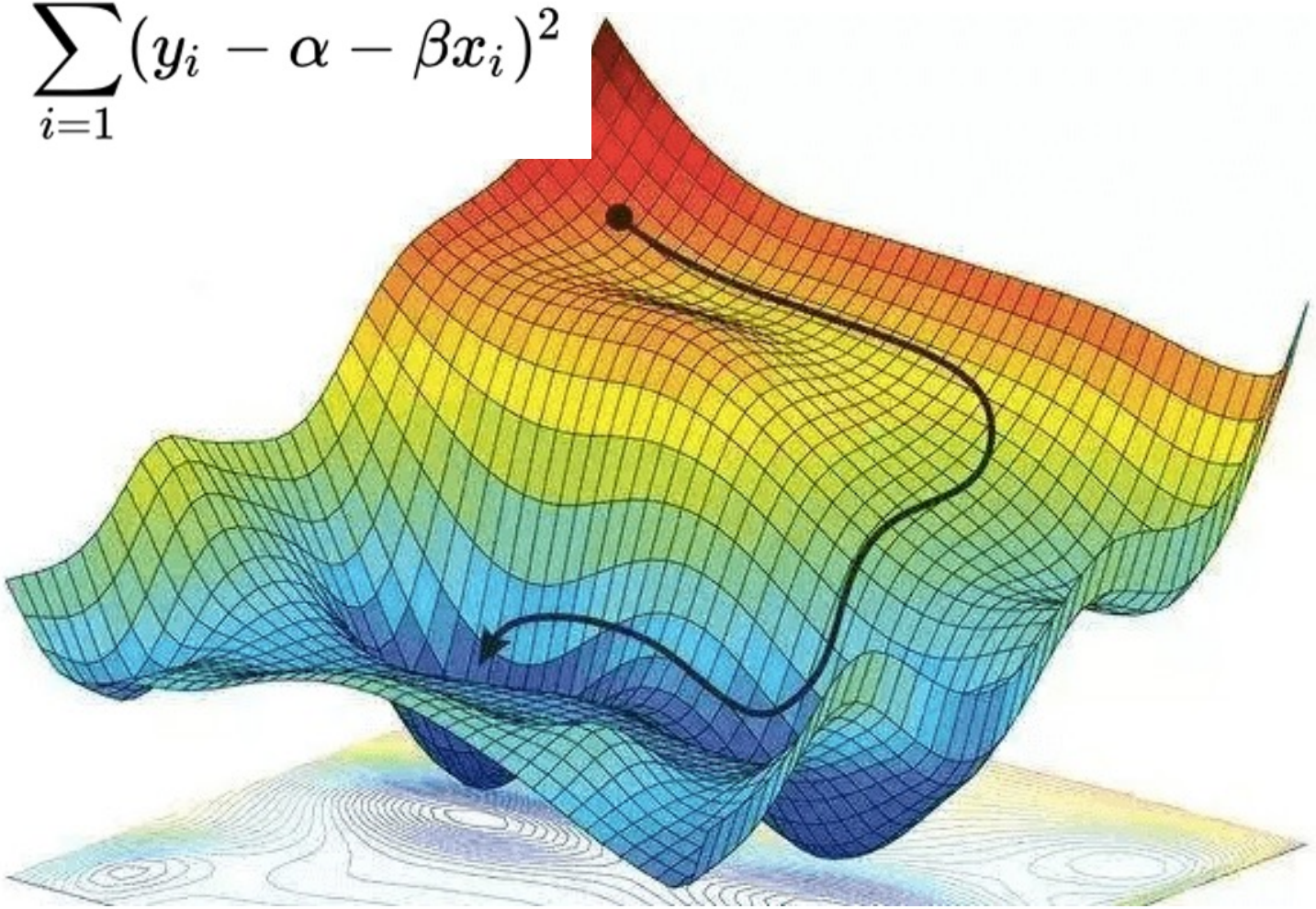
梯度下降



线性回归**算法**

数值解法　梯度下降

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$



线性回归的求解例

有5个数据点(1, 2), (2, 3), (3, 7), (4, 8), (5, 9)，请利用如下公式计算这些点的线性回归方程

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

提示：先计算如下表格

x	y	xy	x^2

线性回归的求解例

通过5个数据点(1, 2), (2, 3), (3, 7), (4, 8), (5, 9)，得到的线性回归方程是

$$y = 0.1 + 1.9\ x$$