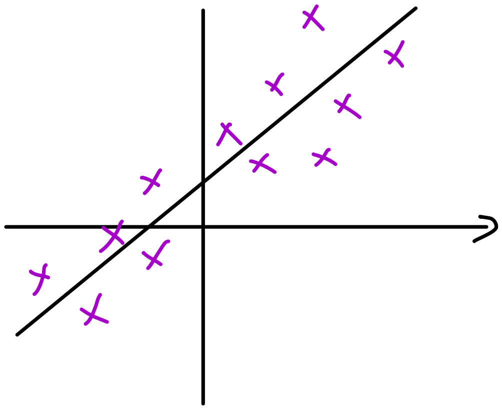
线性回归 (Linear Regression)

1. Introduction

· 2-d 线性拟合



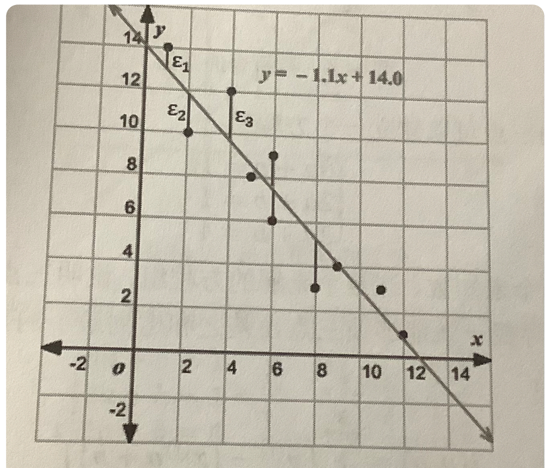


图 14.3　最小二乘法的几何意义

· p维线性拟合

前提:

样本 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$

其中, x_i 是一个p维向量. 表达第i个样本被观察的p个特征.

$$x_i=\begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}_{p\times 1}$$

N个样本集合. 写作 $X=[x_1 \quad x_2 \quad \cdots \quad x_N]^T_{p\times p}$

y_i 表示第i个样本的取值

$$Y=[y_1 \quad y_2 \quad \cdots \quad y_n]^T$$

假设:

回归到一条直线上. 即 $y=w^Tx+b$

其中 $w=\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix}$.

为了简化. 将偏置b简化成 w_0x_{i0} 的形式. 则 w 和 x_i 向量定义成:

$$w=\begin{bmatrix} w_0 \\ w_1 \\ \vdots \end{bmatrix} \qquad x_i=\begin{bmatrix} 1 \\ x_{i1} \\ \vdots \end{bmatrix}$$

$$w = \begin{bmatrix} w_2 \\ \vdots \\ w_p \end{bmatrix}, \quad \begin{bmatrix} x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$$

即原有的映射关系可以写成 $\boxed{y = w^T x}$

2. 最小二乘估计

- 如何估计上述的 w？

最小二乘估计给出以下公式

$$L(w) = \sum_{i=1}^{N} | \underbrace{w^T x_i}_{\text{估计值}} - \underbrace{y_i}_{\text{真实值}} |^2$$

展开上式：

$$L(w) = (w^T x_1 - y_1, w^T x_2 - y_2, \cdots w^T x_N - y_N) \cdot (w^T x_1 - y_1, w^T x_2 - y_2, \cdots w^T x_N - y_N)^T$$

$$= (w^T x_1, w^T x_2 \cdots w^T x_N) - (y_1, y_2, \cdots y_N)$$

$$= w^T (x_1, x_2, \cdots x_N) - (y_1, y_2, \cdots, y_N)$$

$$= (w^T X^T - Y^T)(w^T X^T - Y^T)^T$$

$$= (w^T X^T - Y^T)(Xw - Y)$$

$$= w^T X^T X w - w^T X^T Y - Y^T X w + Y^T Y$$

$X = (x_1, x_2, \cdots, x_N)^T$

$Y = (y_1, y_2, \cdots y_N)^T$

$(w^T X^T Y)^T$

$(k)^T = k$. 实数转置为本身

$Y^T X W \rightarrow 1 \times p \times p \times p \times p \times 1$

$\because (X^T X)^T = X^T X \rightarrow$ 对称矩阵

对于对称矩阵 $\frac{\partial w^T A w}{\partial w} = 2Ax$

$\frac{\partial x^T A}{\partial x} = A$

$$\boxed{L(w) = w^T X^T X w - 2 w^T X^T Y + Y^T Y}$$

$$\frac{\partial}{\partial w} L(w) = 2 X^T X w - 2 X^T Y = 0$$

$$\Rightarrow \boxed{w = (X^T X)^{-1} \cdot X^T Y}$$

- 从高斯噪声角度看最小二乘法
  - 设噪声 ε 服从高斯分布 $\epsilon \sim N(0, \sigma^2)$

    则真实拟合值为 $y = \underbrace{w^T x}_{\text{看成常数}} + \varepsilon$

$$\Rightarrow y | x; w \sim N(w^T x, \sigma^2)$$

MLE: $L(w) = \log P(y | x; w)$

$$= \log \prod_{i=1}^{N} P(y | x; w)$$

$$P(y | x; w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y - w^T x)^2}{2\sigma^2}\right]$$

$$= \sum_{i=1}^{N} \log P(y \mid x_i; w)$$

$$= \sum_{i=1}^{N} \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y - w^T x_i)^2}{2\sigma^2} \right] \right)$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} + \log \exp\left[-\frac{(y - w^T x_i)^2}{2\sigma^2} \right]$$

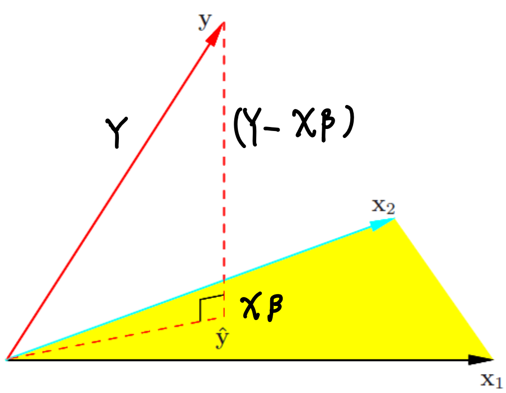
$$= \sum_{i=1}^{N} \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} (y - w^T x_i)^2 \right]$$

$$\Rightarrow \hat{w} = \arg\max_{w} \mathcal{L}(w)$$

$$= \arg\max_{w} \sum_{i=1}^{N} \left(-\frac{1}{2\sigma^2} (y - w^T x_i)^2 \right]$$

$$= \arg\min_{w} \sum_{i=1}^{N} (y - w^T x_i)^2$$ → 同之前的表达式

· 最小二乘法几何意义



$$X^T(Y - X\beta) = 0$$

$$X^TY - X^TX\beta = 0$$

$$\Rightarrow \beta = (X^TX)^{-1} X^TY$$

$X = span(X_1, \cdots, X_n)$ $\Rightarrow$ 写成 $f(w) = X\beta$

3. 正则化 (Regulation)

· Background

$$X = \begin{pmatrix} x_{11} & x_{21} \end{pmatrix}$$

Loss Function: $\mathcal{L}(w) = \sum_{i=1}^{N} | w^T x_i - y_i |^2$

$$\Rightarrow \hat{w} = (X^TX)^{-1} X^TY$$

$X \rightarrow n \times p$ 矩阵. 当 $n >> p$ 时.
(n个样本, $x_i \in \mathbb{R}^p$)

当样本数量不能远大于特征维度

数学解:

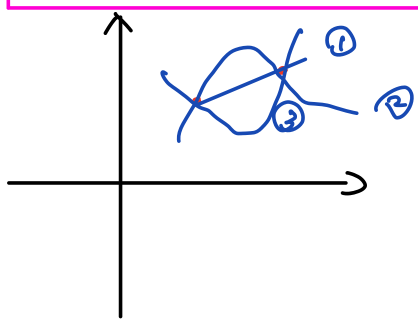
X^TX 将不再是满秩矩阵 $\Longleftrightarrow$ X^TX 存在线性相关值 $\Longleftrightarrow$ X^TX 不可逆

(因变量过多了，出现多解)

e.x. $x_1+x_2+x_3+x_4+x_5=1$

$x_2+x_3+x_4=10$

几何解释: $\boxed{\text{过拟合 (overfitting)}}$



→ 有无多解

- 解决方法

① 增加数据

② 降维/特征提取 (PCA)

③ 正则化

- 正则化的框架

$$\arg\min_{w} \left[L(w) + \lambda P(w) \right]$$

L_1: Lasso Regression $\quad P(w) = \|w\|$

L_2: Ridge Regression $\quad P(w) = \|w\|_2^2 = w^T w$

$$L_2: \quad J(w) = \sum_{i=1}^{N} \|w^T x_i - y_i\|^2 + \lambda w^T w$$

$$= (w^T x_1, w^T x_2 \cdots w^T x_n) - (y_1, y_2, \cdots, y_n)$$

$$= (w^T X^T - Y^T)(w^T X^T - Y^T)^T + \lambda w^T w$$

$$= (w^T X^T - Y^T)(Xw - Y) + \lambda w^T w$$

$$= w^T X^T X w - w^T X^T Y - Y^T X w + Y^T Y + \lambda w^T w$$

$$= w^T X^T X w - 2 w^T X^T Y + \lambda w^T w + Y^T Y$$

$$= \underline{w^T X^T X w} - 2 w^T X^T Y + Y^T Y + \underline{\lambda w^T w}$$

$$= w^T (X^T X + \lambda I) w - 2 w^T X^T Y + Y^T Y$$

$$\hat{w} = \arg\min_{w} J(w)$$

$$\frac{\partial J(w)}{\partial w} = 2(X^T X + \lambda I) w - 2 X^T Y = 0$$

$$\Rightarrow \boxed{w = (X^T X + \lambda I)^{-1} X^T Y}$$

避免了 $X^T X$ 不可逆

• 从贝叶斯角度看

取先验分布 $w \sim N(0, \sigma^2)$. 于是

$$p(w|y) = \frac{p(w) \cdot p(y|w)}{p(y)} \Rightarrow p(w|y) \propto p(w) \cdot p(y|w).$$

(最大后验概率)

$$p(w) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{w^2}{2\sigma^2}\right)$$

$$p(w|y) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(y - w^T x)^2}{2\sigma_0^2}\right)$$

$$\text{MAP: } \hat{w} = \arg\max_{w} \log p(w|y)$$

$$= \arg\max_{w} \log p(w) \cdot p(y|w)$$

$$= \arg\max_{w} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{1}{\sqrt{2\pi}\sigma_0}\right) + \log \exp\left[-\frac{w^2}{2\sigma^2} - \frac{(y - w^T x)^2}{2\sigma_0^2}\right]$$

$$= \arg\min_{w} \frac{(y - w^T x)^2}{2\sigma_0^2} + \frac{w^2}{2\sigma^2}$$

$$\boxed{= \arg\min_{w} (y - w^T x)^2 + \frac{2\sigma_0^2}{2\sigma^2} w^2}$$

Lasso!

(LSE) 最小二乘估计 $\Longleftrightarrow$ 极大似然估计 (noise 为 Gaussian Distribution)

Regularized LSE $\Longleftrightarrow$ MAP (先验分布为 Gaussian Distribution)