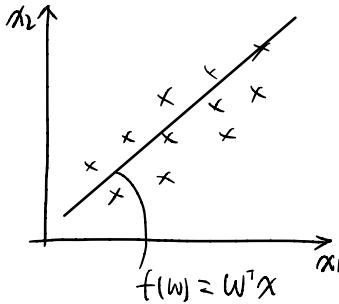


线性回归

一、最小二乘法及其几何意义



$$\begin{aligned} L(w) &= \sum_{i=1}^N \|w^T x_i - y_i\|^2 = \sum_{i=1}^N (w^T x_i - y_i)^2 = \underbrace{(w^T x_1 - y_1, w^T x_2 - y_2, \dots, w^T x_N - y_N)}_{\downarrow} \underbrace{\begin{pmatrix} w^T x_1 - y_1 \\ w^T x_2 - y_2 \\ \vdots \\ w^T x_N - y_N \end{pmatrix}}_{\downarrow} \\ &= (w^T x_1, w^T x_2, \dots, w^T x_N) - (y_1, y_2, \dots, y_N) \\ &= w^T (x_1, x_2, \dots, x_N) - (y_1, y_2, \dots, y_N) \\ &= w^T X - Y \end{aligned}$$

$$\begin{aligned} \nabla_w L(w) &= (w^T X^T - Y^T) (Xw - Y) \\ &= w^T X^T Xw - w^T X^T Y - Y^T Xw + Y^T Y \\ &= w^T X^T Xw - 2w^T X^T Y + Y^T Y \end{aligned}$$

$$\begin{aligned} \hat{w} &= \arg \min_w L(w) \\ \frac{\partial L(w)}{\partial w} &= 2X^T Xw - 2X^T Y = 0 \\ \therefore X^T Xw &= X^T Y \\ \hat{w} &= \underbrace{(X^T X)^{-1}}_{X^T \text{ 为逆}} X^T Y \end{aligned}$$

二. 极大似然估计

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

$$x_i \in \mathbb{R}^p, y_i \in \mathbb{R}, i=1, 2, \dots, N.$$

$$\begin{aligned} X &= (x_1, x_2, \dots, x_N)^T = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix} \\ Y &= \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \end{aligned}$$

$$\text{假设噪声 } \varepsilon \sim N(0, \sigma^2), y = f(w) + \varepsilon, f(w) = w^T x,$$

$$\text{故 } y = w^T x + \varepsilon, y | x; w \sim N(w^T x, \sigma^2) \longrightarrow P(y | x; w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y - w^T x)^2}{2\sigma^2}\right\}$$

MLE:

$$\begin{aligned} L(w) &= \log P(Y | X; w) = \log \prod_{i=1}^N P(y_i | x_i; w) = \sum_{i=1}^N \log P(y_i | x_i; w) \\ &\stackrel{\text{log-likelihood}}{=} \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi}\sigma} + \log \exp\left\{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right\} \\ &= \sum_{i=1}^N \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} (y_i - w^T x_i)^2 \right) \end{aligned}$$

y 是已经出现的值，现在我有了 y 的概率分布，我们希望能确定参数 w ，使得 y 出现的概率最大。因为 y 已呈现实现了，所以使 y 出现概率最大的 w 是我们想要的。

$$\begin{aligned} \hat{w} &= \arg \max_w L(w) \\ &= \arg \max_w \sum_{i=1}^N \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} (y_i - w^T x_i)^2 \right) \\ &= \arg \min_w \sum_{i=1}^N (y_i - w^T x_i)^2 = L(w) \end{aligned}$$

最小二乘估计 \Leftrightarrow 噪声为高斯分布的 MLE.

因为每个 y_i 的分布都是 $N(w^T x_i, \sigma^2)$ ，也就是说 $y_1 \sim N(w^T x_1, \sigma^2)$, $y_2 \sim N(w^T x_2, \sigma^2)$ ，每个 y_i 出现的概率都服从不同的高斯分布，但它们确确实实已经出现了，那就希望它们出现的概率最大。

$$\begin{aligned} \text{即最大化 } & \underbrace{P(y_1 | x_1; w) \cdot P(y_2 | x_2; w) \cdots P(y_N | x_N; w)}_{\prod_{i=1}^N P(y_i | x_i; w)} \end{aligned}$$

三. 正则化.

$$\hat{w} = (X^T X)^{-1} X^T Y$$

若 N < p 或 p 很大，就会导致 $X^T X$ 不可逆，容易造成过拟合。这也是引入正则化的原因。

过拟合 $\rightarrow \left\{ \begin{array}{l} \text{① 加数据} \\ \text{② 特征提取/特征选择, 降维, (PCA)} \\ \text{③ 正则化, 是对 } w \text{ 等参数空间的约束.} \end{array} \right.$

模型：

$$\arg \min_{w} \underbrace{L(w)}_{\text{Loss}} + \lambda P(w)$$

两种常用正则化：

1. L1 正则化, Lasso, $P(w) = \|w\|_1 \rightarrow 1 \text{ 范数.}$
2. L2 正则化, 岷因归, $P(w) = \|w\|_2^2 = w^T w \rightarrow 2 \text{ 范数.}$

L2 正则化, 求 w :

$$\begin{aligned} & L(w) + \lambda P(w) \\ &= \sum_{i=1}^n \|w^T x_i - y_i\|^2 + \lambda w^T w \\ &= (w^T X^T - Y^T)(Xw - Y) + \lambda w^T w \\ &= W^T X^T X W - W^T X^T Y - Y^T X W + Y^T Y + \lambda w^T w \\ &= W^T X^T X W - 2W^T X^T Y + Y^T Y + \lambda w^T w \\ &= W^T (X^T X + \lambda I) W - 2W^T X^T Y + Y^T Y \end{aligned}$$

即 $\hat{w} = \arg \min_w J(w)$

$$\begin{aligned} \frac{\partial J(w)}{\partial w} &= 2(X^T X + \lambda I)W - 2X^T Y = 0 \\ \hat{w} &= (X^T X + \lambda I)^{-1} X^T Y \end{aligned}$$

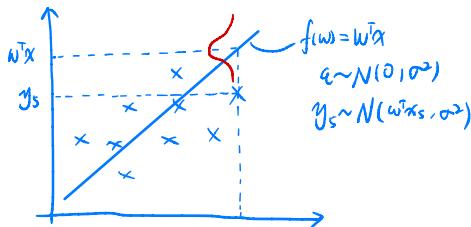


理清：
下惟为对应的参数就有 p, N 个样本都会对应 N 个方程
系数少, 未知参数多, 就会导致有无数解, 矩阵
是奇异阵, 不可逆. 无数解自然就会导致过拟合,
因为会有无数个值适合所有样本.

贝叶斯角度看岭回归

最小二乘估计 \Leftrightarrow MLE (噪声为高斯分布)

最小二乘估计
补充:



贝叶斯角度:

给参数 w 一个先验, 设为 $w \sim N(0, \sigma_0^2)$

$$P(w|y|x) = \frac{P(y|x; w) \cdot P(w)}{P(y|x)}$$

MAP:

$$\begin{aligned}\hat{w} &= \arg \max_w P(w|y|x) \\ &= \arg \max_w P(y|x; w) \cdot P(w) \\ &= \arg \max_w \log [P(y|x; w) \cdot P(w)] \\ &= \arg \max_w \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log \exp \left\{ -\frac{(y - w^T x)^2}{2\sigma^2} - \frac{\|w\|^2}{2\sigma_0^2} \right\} \\ &= \arg \min_w \underbrace{\frac{(y - w^T x)^2}{2\sigma^2} + \frac{\|w\|^2}{2\sigma_0^2}}_{L(w)}\end{aligned}$$

$$\begin{aligned}P(y|x; w) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y - w^T x)^2}{2\sigma^2} \right\} \\ P(w) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left\{ -\frac{\|w\|^2}{2\sigma_0^2} \right\}\end{aligned}$$

即 L2 正则化最小二乘估计 \Leftrightarrow MAP (噪声为高斯分布, 先验为高斯分布)