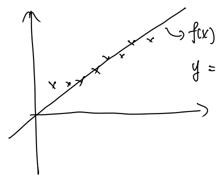
线性回归回顾

Data: $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

$$X = (x_1, x_2, \cdots, x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}_{N \times p}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N \times 1}$$



$f(x) = w^T x = x^T w$

$y = f(x) + \varepsilon$ 其中 ε 代表 noise

x, y, ε 都是 r.v

$x \in \mathbb{R}^p$ $y \in \mathbb{R}$ $\varepsilon \sim N(0, \sigma^2)$

LASSO, Ridge

LSE ⟶ Regularized LSE

⇑ ⇕

MLE ↗ 容易过拟合 (noise is Gaussian)

MAP - 最大后验估计 (noise is Gaussian)

→ 频率派方法 (w is unknow constant)
↓ 优化问题
↓ 点估计

→ Bayesian Method
↓ 贝叶斯派
↓ w是随机变量 (r.v)
↓ 不是点估计，而是求 P(w|Data)

$W_{MLE} = \arg\max_w P(Data|w)$

$P(w|Data) \propto P(Data|w) P(w)$ (似然，先验)

$W_{MAP} = \arg\max_w P(w|Data)$

$= \arg\max_w P(Data|w) P(w)$ → $\begin{cases} N(\mu_0, \Sigma) & \to \text{Ridge} \\ \text{Laplace} & \to \text{LASSO} \end{cases}$

Bayesian Linear Regression

Data: $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

$$X = (x_1, x_2, \cdots, x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}_{N \times p}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N \times 1}$$

$y = f(x) + \varepsilon$ 其中 ε 代表 noise

x, y, ε 都是 r.v

$x \in \mathbb{R}^p$ $y \in \mathbb{R}$ $\varepsilon \sim N(0, \sigma^2)$

↓

$P(y|w, x) = N(w^T x, \sigma^2)$

Bayes Method

因为贝叶斯方法假设 w 是一个分布而不是一个常数

- Inference: posterior (w)
- Prediction: $x^* \to y^*$

Inference: $P(w|Data) = P(w|X, Y) = \dfrac{P(w, Y|X)}{P(Y|X)} = \dfrac{P(Y|w, X) P(w)}{\int P(Y|w, X) P(w) dw}$ (likelihood，先验)

$$P(Y|w, X) = \prod_{i=1}^N P(y_i|w, x_i)$$

$$= \prod_{i=1}^N N(y_i | w^T x_i, \sigma^2)$$

$$P(w) = N(0, \Sigma_p)$$

共轭：Gaussian 分布是自共轭的

$$P(w|Data) \propto P(Y|w, X) P(w)$$

($P(w|Data)$: Gaussian；$P(Y|w,X)$: Gaussian；$P(w)$: Gaussian)

$$\propto \prod_{i=1}^N N(y_i | w^T x_i, \sigma^2) \cdot N(0, \Sigma_p)$$

Gaussian.
↓
假设 $N(\mu_w, \Sigma_w)$
一定可以通过计算得
到 $\mu_w = \cdots$
$\Sigma_w = \cdots$

↓

推导过程如下

likelihood: $P(Y|X, w) = \prod_{i=1}^N \frac{1}{(2\pi)^{\frac{1}{2}} \sigma} \exp\left\{-\frac{1}{2\sigma^2} (y_i - w^T x_i)^2\right\}$

高斯分布

$$= \frac{1}{2\pi^{\frac{N}{2}} \sigma^N} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - w^T x_i)^2\right\}$$

$$= \frac{1}{2\pi^{\frac{N}{2}} \sigma^N} \exp\left\{-\frac{1}{2} (Y - Xw)^T \sigma^{-2} I (Y - Xw)\right\}$$

$$= N(Xw, \sigma^{-2} I)$$

$$\sum_{i=1}^N (y_i - w^T x_i)^2 = (y_1 - w^T x_1, \; y_2 - w^T x_2 \cdots y_N - w^T x_N) \begin{pmatrix} y_1 - w^T x_1 \\ y_2 - w^T x_2 \\ \vdots \\ y_N - w^T x_N \end{pmatrix}$$

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

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

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

$$P(w|Data) \propto N(Xw, \sigma^{-2} I) N(0, \Sigma_p)$$

$$\propto \exp\left\{-\frac{1}{2}(Y - Xw)^T \sigma^{-2} I (Y - Xw)\right\} \cdot \exp\left\{-\frac{1}{2} w^T \Sigma_p^{-1} w\right\}$$

与 w 相关的项全都在 exp 里面的

$$p(w|Data) \propto N(Xw, \sigma^{-2}I)\,N(0, \Sigma_p)$$

$$\propto \exp\left\{-\frac{1}{2}(Y-Xw)^T\sigma^{-2}I(Y-Xw)\right\}\cdot\exp\left\{-\frac{1}{2}w^T\Sigma_p^{-1}w\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}(Y^T - w^TX^T)(Y-Xw) - \frac{1}{2}w^T\Sigma_p^{-1}w\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}(Y^TY - Y^TXw - w^TX^TY + w^TX^TXw) - \frac{1}{2}w^T\Sigma_p^{-1}w\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}(Y^TY - 2Y^TXw + w^TX^TXw) - \frac{1}{2}w^T\Sigma_p^{-1}w\right\}$$

$N(\mu_w, \Sigma_w)$

$\mu_w = ?$

$\Sigma_w = ?$

再次，使用配方法.

以一维 $P(x) = N(\mu, \Sigma)$，其指数部分如下.

$$\exp\left(-\frac{1}{2}(x-u)^T\Sigma^{-1}(x-u)\right)$$

$$= -\frac{1}{2}\left((x^T\Sigma^{-1} - u^T\Sigma^{-1})(x-u)\right)$$

$$= -\frac{1}{2}\left(\underbrace{x^T\Sigma^{-1}x}_{\text{二次项}} - \underbrace{2u^T\Sigma^{-1}x}_{\text{一次项}} + \text{常数}\right)$$

再回到原来的式子.

$$\exp\left\{-\frac{1}{2\sigma^2}(Y^TY - 2Y^TXw + w^TX^TXw) - \frac{1}{2}w^T\Sigma_p^{-1}w\right\}$$

二次项 (因为是 w 的积分多项式，所以 w): $-\frac{1}{2\sigma^2}w^TX^TXw - \frac{1}{2}w^T\Sigma_p^{-1}w$

$$= -\frac{1}{2}\left(w^T\underbrace{(\sigma^{-2}X^TX + \Sigma_p^{-1})}_{\Sigma_w^{-1} = A\ (\text{精度矩阵})}w\right)$$

一次项：$-\frac{1}{2\sigma^2}\cdot(-2)Y^TXw = \sigma^{-2}Y^TXw$

$\mu_w^T\Sigma_w^{-1} = \mu_w^TA$ 要求 μ 先验量，A 是精度矩阵(对称矩阵) $\Rightarrow A^T = A$

$$A\mu_w = \sigma^{-2}X^TY$$

$$\mu_w = \sigma^{-2}A^{-1}X^TY$$

$\therefore P(w|Data) = N(\mu_w, \Sigma_w)$ $\quad \mu_w = \sigma^{-2}A^{-1}X^TY$

$\Sigma_w = A^{-1}$ $\quad (A = \sigma^{-2}X^TX + \Sigma_p^{-1})$

Prediction.

Inference: $P(w|Data)$ $\quad w|Data \sim N(\mu_w, \Sigma_w)$

$\mu_w = \sigma^{-2}A^{-1}X^TY$

$\Sigma_w = A^{-1}$

$A = \sigma^{-2}X^TX + \Sigma^{-1}$

Prediction: Given x^* 求 y^* $\quad$ Model $\begin{cases} f(x) = w^Tx = x^Tw \\ y = f(x) + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \end{cases}$

① $f(x^*) = x^{*T}w$ (noise free)

$\rightarrow P(w|Data) = N(\mu_w, \Sigma_w)$

$w \sim N(\mu_w, \Sigma_w)$ 分线 μ_w, Σ_w

$x^{*T}w \sim N(x^{*T}\mu_w, x^{*T}\Sigma_wx^*)$

$P(f(x^*)|Data, x^*) = N(x^{*T}\mu_w, x^{*T}\Sigma_wx^*)$

② y^* $\quad y^* = f(x^*) + \varepsilon$ 其中 $\varepsilon \sim N(0, \sigma^2)$ (noise)

$\therefore P(y^*|x^*, Data) = N(x^{*T}\mu_w, x^{*T}\Sigma_wx^* + \sigma^2)$

Summary

Data: $\{(x_i, y_i)\}_{i=1}^N \quad x_i \in \mathbb{R}^p \quad y_i \in \mathbb{R}$

Model: $\begin{cases} f(x) = w^T x = x^T w \\ y = f(x) + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \end{cases}$

Bayesian Method:

参数 w 不是采样的参数，w 是一个概率分布.

① Inference: $P(w \mid Data) \rightarrow$ posterior

$P(w \mid Data) \propto$ likelihood $\times$ prior

likelihood $N(0, \sigma)$ prior 共轭 $N(\Delta, D)$

$N(\Delta, \sigma)$

$\mu_w = ?$

$\Sigma_w = ?$

② prediction: given x^*, y^*?

$$p(y^* \mid Data, x^*) = \int \underbrace{P(y^* \mid w, Data, x^*)}_{w \text{和} Data \text{无关} = P(y^* \mid w, x^*)} \underbrace{P(w \mid Data, x^*)}_{P(w \mid Data) \text{ posterior}} dw$$

x^* 与之无关.