Linear Regression

1 Linear Regression

1.1 Least Square

The dataset is $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. The model is $f(w) = w^T x$. The loss function is

$$
\begin{aligned}
L(w) &= ||w^T x_i - y_i||^2 \\
&= (w^T x_1 - y_1, \cdots, w^T x_n - y_n) \cdot (w^T x_1 - y_1, \cdots, w^T x_n - y_n)^T \\
&= (w^T X^T - Y^T) \cdot (Xw - Y) \\
&= w^T X^T X w - w^T X^T Y - Y^T X w + Y^T Y
\end{aligned} \tag{1}
$$

Therefore,

$$
\begin{aligned}
\hat{w} &= \arg\min_w L(w) = 0 \\
\because \quad \frac{\partial a^T x}{\partial x} &= \frac{\partial x^T a}{\partial x} = a \\
\therefore \quad \frac{\partial L}{\partial w} &= 2X^T X w - 2X^T Y = 0 \\
\hat{w} &= \{X^T X\}^{-1} X^T Y
\end{aligned} \tag{2}
$$

1.2 MLE with Gaussian Noise

Given $y = w^T x + \eta,\ \eta \backsim \mathcal{N}(0, \sigma^2)$, so $y \backsim \mathcal{N}(w^T x, \sigma^2)$. Therefore,

$$
\begin{aligned}
L(w) &= \log p(Y|X, w) \\
&= \log \prod_{i=1}^{N} p(y_i|x_i, w) \\
&= \sum_{i=1}^{N} \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}) \right) \\
\therefore \quad \arg\max_w L(w) &= \arg\min_w \sum_{i=1}^{N} (y_i - w^T x_i)^2
\end{aligned} \tag{3}
$$

From Bayesian, having $w \backsim \mathcal{N}(0, \sigma^2)$, therefore,

$$
\begin{aligned}
\hat{w} &= \arg\max_w \log p(w|Y) \\
&= \arg\max_w \log(p(Y|w)p(w)) \\
&= \arg\max_w (\log p(Y|w) + \log p(w)) \\
\because \quad p(Y|w) &= \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}) \\
\because \quad p(w) &= \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{||w||^2}{2\sigma^2}),\ w \backsim \mathcal{N}(0, \sigma^2) \\
\therefore \quad \hat{w} &= \arg\max_w \left(-\frac{(Y - w^T x)^2}{2\sigma^2} - \frac{||w||^2}{2\sigma^2} \right) \\
&= \underbrace{\arg\min_w \left((Y - w^T x)^2 + ||w||^2 \right)}_{\text{ridge regression}}
\end{aligned} \tag{4}
$$