Linear and Non-linear Regression Notes

Linear Regression:

- 1D Case:
- We want to find $y = f(x) + \varepsilon$ where:

a) $f(x) = \underbrace{w}_{\text{Weight}} x + \underbrace{b}_{\text{bias}}$

"w" and "b" are the parameters of "f".

b) ε is the error term (noise)

- We want to estimate "w" and "b" s.t. $f(x)$ fits the training data as well as possible.

The training data is a set of input/output pairs, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

Note: x_i's can be a scalar or vector.

- One way to do this is to minimize the vertical dist btwn the actual value and the predicted value. We can do this using Least Square's Method.

Let $e_i = y_i - f(x_i)$
$\quad\;\; = y_i - (wx_i + b)$

Note: x_i and y_i are from training data.

The loss function, $L(w,b)$, is equal to $\sum_{i=1}^{n} (e_i)^2$

$$= \sum_{i=1}^{n} (y_i - wx_i - b)^2$$

- We need to square the error because of negative values.

- Finding the line that minimizes the squared error is equivalent to solving for "w" and "b" that minimizes L(w,b). This can be done by setting the derivatives of L w.r.t "w" and "b" to 0 and then solving.

For b:

$$\frac{\partial L}{\partial b} = -2\sum_{i=1}^{n}(y_i - wx_i - b) = 0$$

$$0 = \sum_{i=1}^{n} y_i - w\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b$$

$$= \sum_{i=1}^{n} y_i - w\sum_{i=1}^{n} x_i - bn$$

$$bn = \sum_{i=1}^{n} y_i - w\sum_{i=1}^{n} x_i$$

$$b^* = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{w\sum_{i=1}^{n} x_i}{n}$$

$$= \hat{y} - w\hat{x}$$

For ω:

First, we can rewrite L by substituting $\hat{y} - \omega \hat{x}$ for b.

$$L = \sum_{i=1}^{n} (y_i - \omega x_i - (\hat{y} - \omega \hat{x}))^2$$

$$= \sum_{i=1}^{n} ((y_i - \hat{y}) - \omega(x_i - \hat{x}))^2$$

$$\frac{\partial L}{\partial \omega} = -2 \sum_{i=1}^{n} ((y_i - \hat{y}) - \omega(x_i - \hat{x}))(x_i - x) = 0$$

$$0 = \sum_{i=1}^{n} (y_i - \hat{y})(x_i - \hat{x}) - \omega(x_i - \hat{x})^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y})(x_i - \hat{x}) - \sum_{i=1}^{n} \omega(x_i - x)^2$$

$$\omega^* = \frac{\sum_{i=1}^{n} (y_i - \hat{y})(x_i - \hat{x})}{\sum_{i=1}^{n} (x_i - \hat{x})^2}$$

- Multi-Dimensional Inputs:
- Now, let $x \in \mathbb{R}^D$. I.e. $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$ ← 1 data point with D features.

- $$f(x) = w^T x + b$$
$$= \sum_{i=1}^{n} w_i x_i + b$$

We can add b to w and 1 to x to "absorb" b.

$$w = \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix}$$

Now, $f(x) = w^T x$
$$= \sum_{i=1}^{n} w_i x_i$$

- $$L(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$
$$= \|y - Xw\|^2$$ where $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ x_N \end{bmatrix}$

$$X = \begin{bmatrix} — x_1^T — \\ — x_2^T — \\ \vdots \\ — x_N^T — \end{bmatrix}$$

$$\overbrace{L(w) = (y - Xw)^T (y - Xw)}^{\|y - Xw\|^2}$$
$$= (y^T - w^T X^T)(y - Xw)$$
$$= y^T y - y^T X w - w^T X^T y + w^T X^T X w$$
$$= w^T X^T X w - 2 y^T X w + y^T y$$

- Now, to find w^*:

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

$$0 = (X^T X) w - X^T y$$
$$(X^T X) w = X^T y$$
$$w^* = \underbrace{(X^T X)^{-1} X^T}_{\text{Pseudo Inverse}} y$$

Note: $(X^T X)$ isn't always invertible.

Non-linear Regression:

- In basis function regression, we introduce a basis function denoted by $b_k(x)$.

- 2 common basis functions are the polynomials and radia basis functions (RBF).

- For polynomial, we have $b_k(x) = x^k$.

$$f(x) = \sum_{i=1}^{N} w_i b_i(x)$$ ← General basis function representation.

$$= \sum_{i=1}^{N} w_i x^i$$

- For RBF, we have $b_K(x) = \exp\left(-\frac{(x-\mu_K)^2}{2\sigma_K^2}\right)$

μ_K is the center of the basis function.

σ_K^2 is the width of the basis function.

- Examples of polynomial basis function:

1. Let $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ be data points.

$$
\begin{aligned}
f(x) &= \sum_{i=0}^{k=2} w_i b_i(x) \\
&= \sum_{i=0}^{k=2} w_i x^i \\
&= w_0 x^0 + w_1 x + w_2 x^2 \\
&= w_0 + w_1 x + w_2 x^2 \\
&= \underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ & \vdots & \\ 1 & x_N & x_N^2 \end{bmatrix}}_{\text{Basis function matrix, } B} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}
\end{aligned}
$$

$B_{i,j} = b_j(x_i)$

$B \in R^{N,K}$

Each row in B corresponds to 1 data point.

2. Let $\{([x_{11} \cdots x_{1D}], y_1), ([x_{21} \cdots x_{2D}], y_2), \ldots ([x_{N1} \cdots x_{ND}], y_N)\}$ be the data points.

$$f(x) = \sum_{i=0}^{k=2} w_i b_i(x)$$

$$= \sum_{i=0}^{k=2} w_i x^i$$

$$= \begin{bmatrix} 1 & [x_{11}, x_{12}, \ldots, x_{1D}]^1 & [x_{11}, x_{12}, \ldots, x_{1D}]^2 \\ & \vdots & \\ 1 & [x_{N1}, x_{N2}, \ldots, x_{ND}]^1 & [x_{N1}, x_{N2}, \ldots, x_{ND}]^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

Basis function matrix

– $L(w) = \sum_i (y_i - f(x_i))^2$

$$= \sum_i \left(y_i - \sum_j w_j b_j(x)\right)^2$$

$$= \|y - Bw\|^2$$
$$= (y - Bw)^T (y - Bw)$$
$$= (y^T - w^T B^T)(y - Bw)$$
$$= w^T B^T B w - 2y^T B w + y^T y$$

$$\frac{\partial L}{\partial w} = 2(B^T B)w - 2B^T y + 0 = 0$$

$$w^* = (B^T B)^{-1} B^T y$$

- Regularized Least Squares:
- $L(w) = \underbrace{\|y - Bw\|^2}_{\text{Data Term}} + \underbrace{\lambda\|w\|^2}_{\text{Smoothness Term}}$

$$= (y - Bw)^T (y - Bw) + \lambda w^T w$$
$$= w^T B^T B w - \qquad + \lambda w^T w + y^T y$$

$$\frac{\partial L}{\partial w} = 2B^T B w - 2B^T y + 2\lambda w = 0$$

$$0 = B^T B w - B^T y + \lambda w$$
$$= (B^T B + \lambda I) w - B^T y$$
$$(B^T B + \lambda I) w = B^T y$$
$$w^* = \underbrace{(B^T B + \lambda I)^{-1}}_{\text{Always invertible}} B^T y$$