Linear Regression

Note Title 3/9/2010

Training Data with N observations

$\{(\underline{x}_i, y_i),\ i = 1,2,\ldots N\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

$\underline{x}_i$: data point; y_i: targeted variable associated with $\underline{x}_i$

$\underline{x}_i = \begin{pmatrix} x_i(1) \\ x_i(2) \\ \vdots \\ x_i(d) \end{pmatrix}$ ← features/attributes

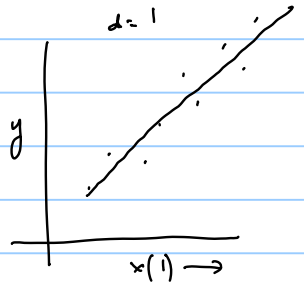
$\hat{y} = f(\underline{x})$

Learning problem: Given training data, come up with a "good f"

Linear Regression: $f(\underline{x})$ is a linear function of $\underline{x}$

$$\underline{x} = \begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(d) \end{pmatrix}, \quad f(\underline{x}) = w_0 + w_1 x(1) + w_2 x(2) + \ldots + w_d x(d)$$

$f(\underline{x}_i) \approx y_i, \quad i = 1,2,\ldots N$



$$\min_{w_0, w_1, \ldots, w_d} \sum_{i=1}^{N} \left[\left((w_0 + w_1 x_i(1) + w_2 x_i(2) + \ldots + w_d x_i(d)) - y_i \right) \right]^2$$

Special case when $d = 1$

$f(x) = w_0 + w_1 x$

$$\min_{w_0, w_1} \underbrace{\sum_{i=1}^{N} (w_0 + w_1 x_i - y_i)^2}_{E(w)}$$ — Least Squares formulation

$$\frac{\partial E}{\partial w_0} = \sum_{i=1}^{N} 2(w_0 + w_1 x_i - y_i) = 0 \Rightarrow N w_0 + w_1 \left(\sum_{i=1}^{N} x_i\right) = \sum_{i=1}^{N} y_i \quad (1)$$

$$\frac{\partial E}{\partial w_1} = \sum_{i=1}^{N} 2(w_0 + w_1 x_i - y_i) x_i = 0 \Rightarrow \left(\sum x_i\right) w_0 + w_1 \left(\sum_{i=1}^{N} x_i^2\right) = \sum_{i=1}^{N} x_i y_i \quad (2)$$

$$\begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum x_i y_i \end{bmatrix}$$ — Normal Equations

$\bar{x} = \frac{1}{N} \sum x_i$, $\bar{y} = \frac{1}{N} \sum y_i$

$$\begin{bmatrix} N & N\bar{x} \\ N\bar{x} & \sum x_i^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} N\bar{y} \\ \sum x_i y_i \end{bmatrix}$$

$$\begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N} \sum x_i^2 \end{bmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \frac{1}{N} \sum x_i y_i \end{pmatrix}$$

$$Aw = b$$
$$w = A^{-1}b$$

$$\det(A) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sigma^2 = \sigma_{xx}$$

$$w = \frac{1}{\sigma^2}\begin{bmatrix} \frac{1}{N}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\begin{bmatrix} \bar{y} \\ \frac{1}{N}\sum x_i y_i \end{bmatrix}$$

$$w_0 = \frac{1}{\sigma^2}\left[\frac{1}{N}\bar{y}\sum x_i^2 - \frac{1}{N}\bar{x}\sum x_i y_i\right]$$

$$w_1 = \frac{1}{\sigma^2}\left[-\bar{x}\bar{y} + \frac{1}{N}\sum x_i y_i\right] = \sigma_{xy}$$

$$\boxed{w_1 = \frac{\sigma_{xy}}{\sigma_{xx}}}$$

$f(x_2)$ $f(x_1)$

$$X = \begin{matrix} \uparrow \\ N \\ \downarrow \end{matrix}\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & \vdots \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \qquad w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \quad Xw = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ w_0 + w_1 x_3 \\ \vdots \\ w_0 + w_1 x_N \end{bmatrix}$$

Least Squares Objective: $\min_w \|Xw - y\|_2^2$ where $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{bmatrix}\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$(X^T X)w = X^T y$$ — Same as Normal Equation

Training points x_i are d-dimensional, $x_i \in \mathbb{R}^d$, $x_i = \begin{bmatrix} x_i(1) \\ x_i(2) \\ \vdots \\ x_i(d) \end{bmatrix}$

$$X = \begin{bmatrix} 1 & x_1(1) & x_1(2) & \dots & x_1(d) \\ 1 & x_2(1) & x_2(2) & \dots & x_2(d) \\ \vdots & & & & \vdots \\ 1 & x_n(1) & x_n(2) & \dots & x_n(d) \end{bmatrix} \qquad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$Xw = \begin{bmatrix} w_0 + w_1 x_1(1) + w_2 x_1(2) + \dots + w_d x_1(d) \\ w_0 + w_1 x_2(1) + w_2 x_2(2) + \dots + w_d x_2(d) \\ \vdots \\ w_0 + w_1 x_n(1) + w_2 x_n(2) + \dots + w_d x_n(d) \end{bmatrix}$$

Least Squares Objective: $\min_w \underbrace{\|Xw - y\|_2^2}_{g(w)}$

$$g(w) = \|Xw - y\|_2^2 = (Xw - y)^T (Xw - y)$$
$$= w^T X^T X w - 2 w^T X^T y + y^T y$$

$$\nabla_w g = \frac{\partial g}{\partial w} = \begin{bmatrix} \frac{\partial g}{\partial w_0} \\ \frac{\partial g}{\partial w_1} \\ \vdots \\ \frac{\partial g}{\partial w_d} \end{bmatrix} = 2X^T X w - 2X^T y = 0$$

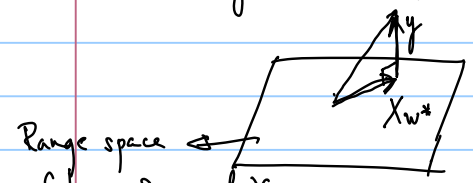
$$\Rightarrow \boxed{X^T X w = X^T y}$$ — Normal Equations

BLUE

Best Linear Unbiased Estimation

Geometric View

$Xw \approx y$ $\quad \min_w \|Xw - y\|_2^2$



$Xw^* - y \perp Xw$ for all w

$$(Xw)^T (Xw^* - y) = 0$$
$$\Rightarrow w^T X^T (Xw^* - y) = 0 \quad \forall w$$
$$\Rightarrow X^T X w^* - X^T y = 0$$
$$\Rightarrow \boxed{X^T X w^* = X^T y}$$

Instead of solving $\min_w \|Xw - y\|_2^2$

Use Regularization $\min_w \left(\|Xw - y\|_2^2 + \lambda \|w\|_2^2 \right), \lambda \geq 0$ ← regularization parameter

Ridge Regression

$\min_w \|Xw - y\|_2^2 + \lambda \|w\|_1$ — Lasso

Solution: $(X^T X + \lambda I) w = X^T y$ — Ridge Regression

Ridge Regression Identical to solving $\min_w \|Xw - y\|_2^2$ st $\|w\|_2 \leq t$

Lasso: $\min_w \|Xw - y\|_2^2$ st $\|w\|_1 \leq t$