Linear Regression

1. $\boxed{\text{Clean / Augment data}}$

Receive $((x_i, y_i))_{i=1}^{n}$

label

$y_i \in \mathbb{R}$

$x_i \in \mathbb{R}^d$

2. $\boxed{\text{Model}}$

linear model $x \to w^T x$

parameters/weights

3. $\boxed{\text{loss}}$

truth

$(\hat{y}, y) \to \frac{1}{2}(y - \hat{y})^2$ since abs() non-differentiable.

prediction

$\boxed{\text{Empirical risk / training error}}$

$\hat{y}$ prediction

$$\hat{R}(w) := \frac{1}{n} \sum_{i=1}^{n} \ell(w^T x, y_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2}(w^T x_i - y_i)^2$$

$$= \frac{1}{2n} \|Xw - y\|^2$$

$$X = \begin{bmatrix} - x_1^T - \\ | \\ - x_n^T - \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\|v\|_2 = \sqrt{\langle v, v\rangle} \qquad \|x\|_p = \left(|x_1|^p + \cdots |x_n|^p\right)^{1/p}$$

$\boxed{\text{predictor}}$:

$$x \to w^T x + b$$

"feature engineering"

$$x \to \begin{bmatrix} x \\ 1 \end{bmatrix} \in \mathbb{R}^{d+1}$$

$$= \underbrace{\begin{bmatrix} w \\ | \\ b \end{bmatrix}}_{\mathbb{R}^{d+1}}^T \underbrace{\begin{bmatrix} x \\ 1 \end{bmatrix}}_{\mathbb{R}^{d+1}}$$

$$X := \begin{bmatrix} \leftarrow & x_1^T & \rightarrow & 1 \\ & \vdots & & \vdots \\ \leftarrow & x_n^T & \rightarrow & 1 \end{bmatrix} \qquad W := \begin{bmatrix} w_1 \\ \vdots \\ w_d \\ b \end{bmatrix}$$

4. Gradient Descent

① $w_0 = 0$

② $w_{i+1} := w_i - \eta \nabla \hat{R}(w_i) = w_i - \frac{2}{n} X^T (X w_i - y)$

$\boxed{\text{Regression Towards the mean}}$

$$\underset{w \in \mathbb{R}^1}{\arg\min} \|Xw - y\|^2 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$\boxed{\text{Other Loss}}$: $L(\hat{y}, y) = \ln(1 + \exp(-\hat{y}y))$

$$\hat{R}(w_t) \xrightarrow[t\to\infty]{} \min_{v \in \mathbb{R}^d} \hat{R}(v)$$

$\hat{R}(w_t)$ finally converge.

$\|\cdot\|$ norm

$\boxed{\text{Normal equations and SVD}}$

We want to find $\vec{w}$ s.t.

$$2n\,\hat{R}(w) = \|X\hat{w} - y\|^2 = \min_{w \in \mathbb{R}^d} \|Xw - y\|^2$$

$$\nabla w \tfrac{1}{2}\|Xw - y\|^2 = X^T(Xw - y) = 0$$

$$X^T X w = X^T y$$ (normal equations)

$$\Rightarrow \quad w = (X^T X)^{-1} X^T y$$ (X full rank)

$\hat{R}(\hat{w}) = \min_{w} \hat{R}(w) \Leftarrow\Rightarrow$ w satisfies normal equations

$(\Leftarrow)$ Suppose $\hat{w}$ satisfies $X^T X \hat{w} = X^T y$

$\forall\ w \in \mathbb{R}^d$, s.t. $\hat{w} \in \mathbb{R}^d$, $\hat{R}(w) \geq \hat{R}(\hat{w})$

$$\underbrace{\frac{1}{2}\|Xw - y\|^2}_{\hat{R}(w)} = \frac{1}{2}\|Xw - X\hat{w} + X\hat{w} - y\|^2$$

$$= \underbrace{\frac{1}{2}\|Xw - X\hat{w}\|^2}_{\geq 0} + \underbrace{\langle Xw - X\hat{w},\ X\hat{w} - y\rangle}_{=0} + \underbrace{\frac{1}{2}\|X\hat{w} - y\|^2}_{\hat{R}(\hat{w})}$$

$$= (Xw - X\hat{w})^T (X\hat{w} - y)$$

$$= (w - \hat{w})^T \underbrace{(X^T X\hat{w} - X^T y)}_{0} = 0$$

$$\Rightarrow \hat{R}(w) \geq \hat{R}(\hat{w})$$

other direction

$(\Rightarrow)$

($\Rightarrow$) Suppose $\hat{R}(\hat{w}) = \min_w \hat{R}(w)$

so $\nabla_w \hat{R}(\hat{w}) = 0$

so $\nabla_w \frac{1}{2}\|X\hat{w} - y\|^2 = 0$

$$\Rightarrow \nabla_w \frac{1}{2}\langle X\hat{w} - y, X\hat{w} - y\rangle = 0$$

$$\Rightarrow \cancel{\tfrac{1}{2}}\cancel{2}\, X^T \cdot (X\hat{w} - y) = 0$$

$$\Rightarrow X^T X \hat{w} = X^T y$$

So $\hat{R}(\hat{w}) = \min_w \hat{R}(w) \Leftrightarrow w$ satisfies normal equation

$\boxed{\text{Singular Value Decomposition}}$

$M \in R^{n \times d}$, $((s_i, u_i, v_i))_{i=1}^{r}$ is an SVD of M if

1. M has rank r
2. $s_1 \geq s_2 \ldots \geq s_r > 0$. (singular values)
3. $(u_i)_{i=1}^{r}$ are orthonormal and span col space of M.
4. $(v_i)_{i=1}^{r}$ are orthonormal and span row space of M.
5. $M = \sum_i^r s_i u_i v_i^T$

$\boxed{\text{Properties of SVD}}$:

① $(s_1, \dots, s_r)$ is unique.

② SVD always exists, and real valued.

③ For $k < r$, low rank approximation

$$\sum_{i=1}^{k} s_i u_i v_i^T \approx M$$

$\boxed{\text{Pseudoinverse}}$

Given $M = \sum_i s_i u_i v_i^T$ $\quad M^+ := \sum_{i=1}^{r} \frac{1}{s_i} v_i u_i^T$

$$MM^+ = \sum_i^r s_i u_i v_i^T \sum_{i=1}^{r} \frac{1}{s_i} v_i u_i^T$$

$$= \sum_{i,j}^{r} \frac{s_i}{s_j} u_i \underbrace{v_i^T v_j}_{\text{orthonormal}} u_j^T$$

so $v_i^T v_j = \begin{cases} & \text{if } i \neq j \\ & \text{o.w.} \end{cases}$

$$\Rightarrow MM^+ = \sum_i \cancel{\frac{s_i}{s_i}} u_i u_i^T = \sum_i^r u_i u_i^T$$

$$MM^+ = \sum_i^r v_i v_i^T$$

$\boxed{M^+ \text{ unique}}$

$$\underbrace{MM^+}M = M$$

$$\underbrace{M^+M}M^+ = M^+$$

$\Rightarrow$ $M^{-1} = M^+$ if M^{-1} exists.

$M = 0 \Rightarrow r = 0$ and $M^+ = 0$

$\boxed{\text{OLS solution via SVD}}$

Given $\hat{R}(w) = \|Xw - y\|^2/2n$

OLS: $\hat{w}_{ols} = X^+ y$

$\hat{R}(\hat{w}_{ols}) = \min_w \hat{R}(w)$

$\Rightarrow X^T X \hat{w}_{ols} = X^T y$

$X^T X X^+ y = X^T y$

$\Rightarrow \hat{R}(\hat{w}_{ols}) = \hat{R}(w')$