Perceptron algorithm

$$f(x) = \text{sign}(wx + b)$$

$$L(w, b) = \max(0, -y(wx + b)) \rightarrow \text{hinge loss}$$

update rule: $w_j := w_j - \alpha \frac{dL}{dw_j}$ (α: learning rate)

$$b := b - \alpha \frac{dL}{db}$$

$$\frac{dL}{dw_j} = \begin{cases} -y x_j & \text{if } y(wx + b) \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{dL}{db} = \begin{cases} -y & \text{if } y(wx + b) \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow$ $w := w + \alpha y x$ and $b := b + \alpha y$

example

Sample	x_1	x_2	y
1	2	3	+1
2	1	1	-1
3	2	1	-1
4	3	3	+1
5	5	5	+1

initialize: $w = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $b = 0$ $\alpha = 1$ ↳ ignore

iteration #1 $x = [2 \quad 3]$ $y = +1$

$$f(x) = \text{sign}\left(\underset{x}{[2 \quad 3]} \underset{w}{\begin{bmatrix} 0 \\ 0 \end{bmatrix}} + \overset{b}{0} \right) = 0 \Rightarrow \text{wrong}$$

↳ prediction

$$w = \underset{w}{\begin{bmatrix} 0 \\ 0 \end{bmatrix}} + \underset{y}{(1)} \underset{x}{\begin{bmatrix} 2 \\ 3 \end{bmatrix}} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$b = 0 + (1) = 1$$

iteration #2 $x = [1 \quad 1]$ $y = -1$

$$f(x) = \text{sign}\left([1 \quad 1] \begin{bmatrix} 2 \\ 3 \end{bmatrix} + 1 \right) = +1 \Rightarrow \text{wrong}$$

$$w = \begin{bmatrix} 2 \\ 3 \end{bmatrix} + (-1) \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$b = 1 + (-1) = 0$$

iteration #3 $x = [2 \quad 1]$ $y = -1$

$$f(x) = \text{sign}\left([2 \quad 1] \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 0 \right) = +1 \Rightarrow \text{wrong}$$

$$w = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + (-1) \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$b = 0 + (-1) = -1$$

iteration #4 $x = [3 \quad 3]$ $y = +1$

$$f(x) = \text{sign}\left([3 \quad 3] \begin{bmatrix} -1 \\ 1 \end{bmatrix} - 1 \right) = -1 \Rightarrow \text{wrong}$$

$$w = \begin{bmatrix} -1 \\ 1 \end{bmatrix} + (1) \begin{bmatrix} 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

$$b = -1 + (1) = 0$$

iteration #5 $x = [5 \quad 5]$ $y = +1$

$$f(x) = \text{sign}\left([5 \quad 5] \begin{bmatrix} 2 \\ 4 \end{bmatrix} + 0 \right) = +1 \Rightarrow \text{correct}$$

no update !

and so on ...

linear sum

- if x is on the decision boundary:

$$wx + b = 0$$

- if x is on the margin :

(x ↓ support vector)

$$wx + b = \pm 1$$

(+ve : upper margin; -ve : lower margin)

- distance from x to the boundary:

$$d = \frac{|wx + b|}{\|w\|}$$

- we want the margin to be 1 unit away from decison boundary $\Rightarrow wx + b = 1$

$$\Rightarrow d = \frac{1}{\|w\|} \quad \text{for one "side"}$$

$$\Rightarrow \frac{2}{\|w\|} = \text{margin distance}$$

- goal: $\max_w \frac{2}{\|w\|}$ or $\min_w \frac{\|w\|^2}{2}$ → makes it convex (global minima); → simplifies derivation

- constraint: $y_i (wx_i + b) \geq 0 \quad \forall i$

- Lagrange multiplier stuff bc constraints: (whatever that means)

$$L(w, b, \alpha) = \frac{\|w\|^2}{2} - \sum_{i=1}^{n} \alpha_i \left(y_i (wx_i + b) - 1 \right)$$

L: minimize; α: maximize

- too many variables ∴ change primal form to dual formation by setting $\frac{dL}{dw}$ & $\frac{dL}{db} = 0$

$$\frac{dL}{dw} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

w to substitute

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

- we have a new w. replace in primal form:

$$\frac{\|w\|^2}{2} = \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

- for the other term in $L(w, b, \alpha)$:

$$\sum_{i=1}^{n} \alpha_i \left(y_i (w x_i + b) - 1 \right) = \sum_{i=1}^{n} \alpha_i y_i (w x_i + b) - \alpha_i$$

has to be 0 from $\frac{dL}{db}$ above

$$= -\sum_{i=1}^{n} \alpha_i$$

- finally:

$$L(\alpha) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{n} \alpha_i$$

goal: $\max_{\alpha} L(\alpha)$

linear regression

$y = X\beta + \varepsilon$ s.t. $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^{n \times 1}$

goal: $\min_{\beta} \| y_i - X_i \beta \|_2^2$

Ridge : $+ \lambda \|\beta\|^2$ Lasso : $+ \lambda |\beta|$ elastic : + both

normal equation $\min_{\beta} J(\beta) = \min_{\beta} \| y_i - X_i \beta \|_2^2$

$$J(\beta) = \min_{\beta} (y - X\beta)^T (y - X\beta)$$

$$J(\beta) = y^T y - 2\beta^T X^T y + \beta^T X^T X \beta$$

take out take out

$$\frac{\partial J}{\partial \beta} = 0 - 2X^T y + 2X^T X \beta = 0$$

$$X^T X \beta = X^T y$$

$$\boxed{\beta = (X^T X)^{-1} X^T y}$$

could be singular (non-invertible)

$\therefore$ perturb $(X^T X + \varepsilon I)^{-1}$

weighted least squares

$$\min_{\beta} J(\beta) = \min_{\beta} \; w \, \| (y - X\beta) \|_2^2$$

$w_i = \frac{1}{\sigma_i^2}$ → variance of observation i

($\uparrow \sigma_i^2 \Rightarrow \downarrow w_i$)

w is a diagonal matrix $W = \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{bmatrix}$

$$\min_{\beta} J(\beta) = \min_{\beta} \; (y - X\beta)^T W (y - X\beta)$$

$$\frac{dJ}{d\beta} = -2X^T W (y - X\beta) = 0$$

$$X^T W y - X^T W X \beta = 0$$

$$X^T W X \beta = X^T W y$$

$$\boxed{\beta = (X^T W X)^{-1} X^T W y}$$

ridge regression

$$\min_{\beta} J(\beta) = \min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|^2$$

$$= (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$$

we know $\frac{\partial J}{\partial \beta}$ of this part

$$\frac{\partial J}{\partial \beta} = -2X^T y + 2X^T X\beta + 2\lambda\beta$$

$$= -X^T y + (X^T X + \lambda I)\beta = 0$$

$$\boxed{\beta = (X^T X + \lambda I)^{-1} X^T y}$$