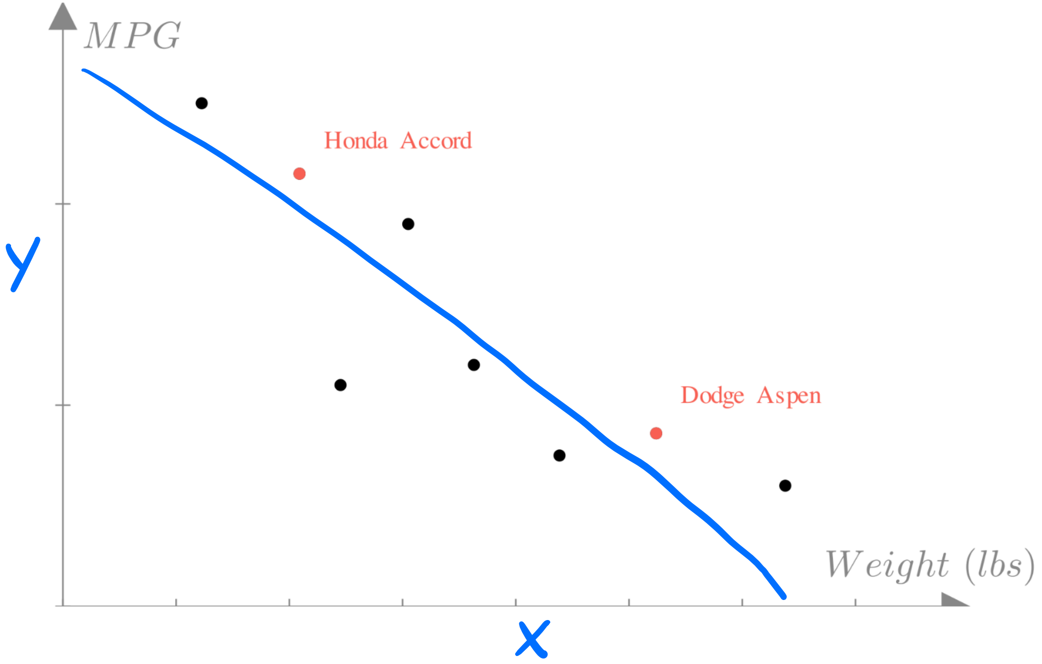
$$\text{Input: } \mathbf{x}_i = \begin{bmatrix} \text{Weight} \\ \text{Horsepower} \\ \text{Displacement} \\ \text{0-60mph} \end{bmatrix}, \quad \text{Output: } y_i = MPG$$

Predicted MPG =

$$= (\text{weight})w_1 + (\text{horsepower})w_2 + (\text{displacement})w_3 + (\text{0-60mph})w_4 + b$$

w

$$f(\mathbf{x}) = \begin{bmatrix} \text{Weight} \\ \text{Horsepower} \\ \text{Displacement} \\ \text{0-60mph} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} + b$$

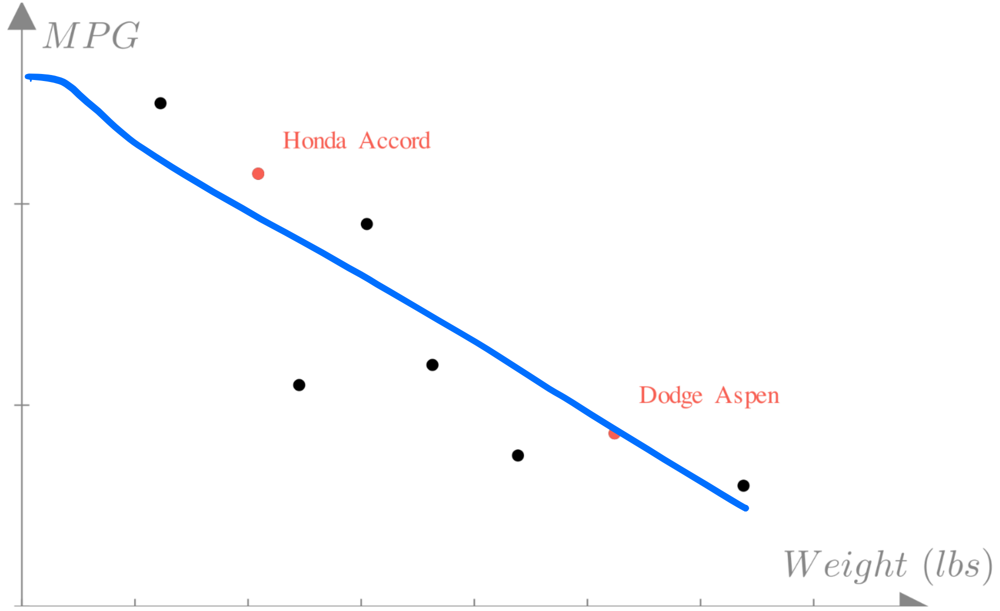


$$\text{Input: } \mathbf{x}_i = \begin{bmatrix} \text{Weight} \\ \text{Horsepower} \\ \text{Displacement} \\ \text{0-60mph} \end{bmatrix}, \quad \text{Output: } y_i = MPG$$

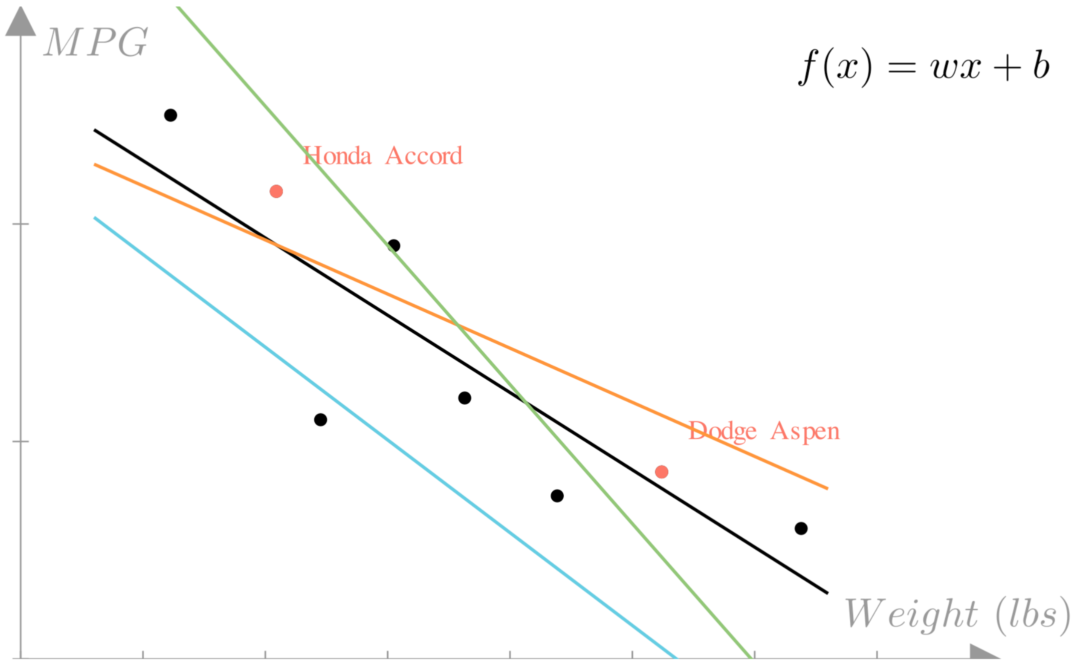
$$\text{Dataset: } \mathcal{D} = \{(\mathbf{x}_i,\ y_i) \text{ for } i \in 1 \ldots N\}$$

obs

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



$$\text{Input: } \mathbf{x}_i = \begin{bmatrix} \text{Weight} \\ \text{Horsepower} \\ \text{Displacement} \\ \text{0-60mph} \end{bmatrix}, \quad \text{Output: } y_i = MPG$$



$$\text{Dataset: } \mathcal{D} = \{(\mathbf{x}_i,\ y_i) \text{ for } i \in 1 \ldots N\}$$

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

$$MSE = \underbrace{\frac{1}{N}\sum_{i=1}^{N}}(f(\mathbf{x}_i) - y_i)^2 = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2}$$

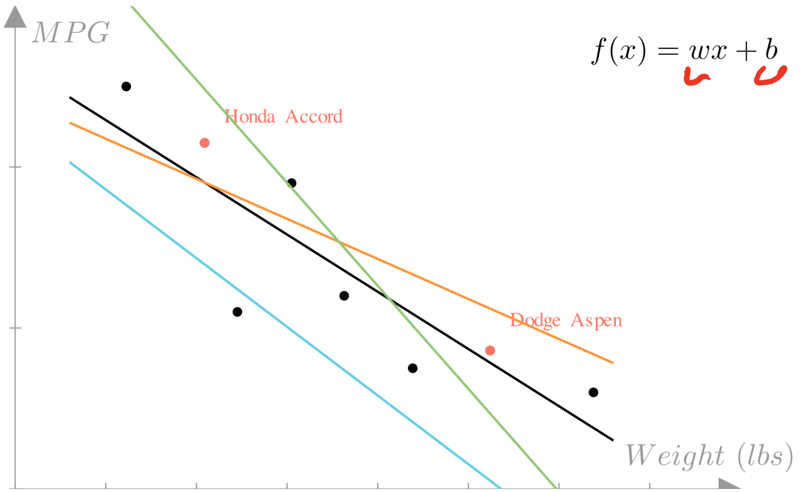
Dataset: $\mathcal{D} = \{(\mathbf{x}_i,\ y_i) \text{ for } i \in 1 \dots N\}$

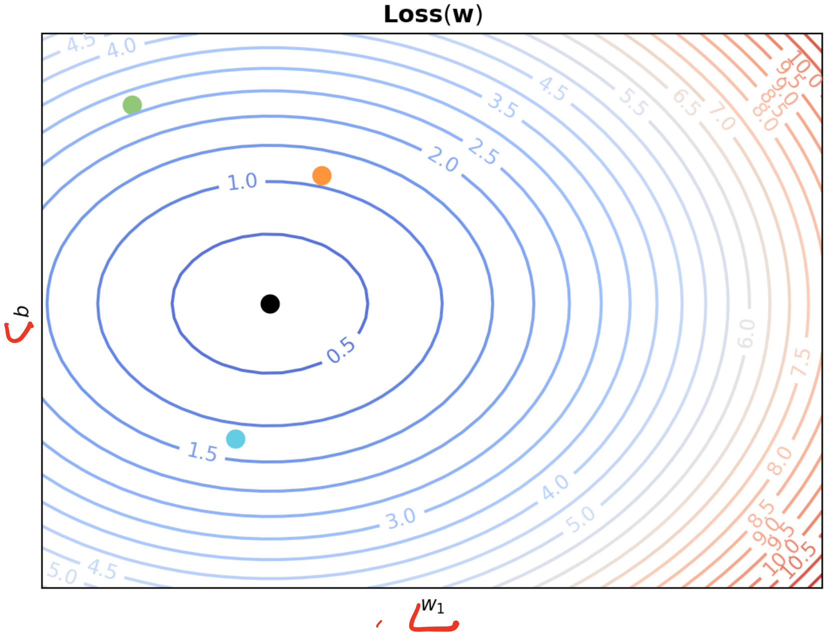
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i) - y_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmin}} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2$$

Prediction func. $f(x)$

Loss func Loss(w)



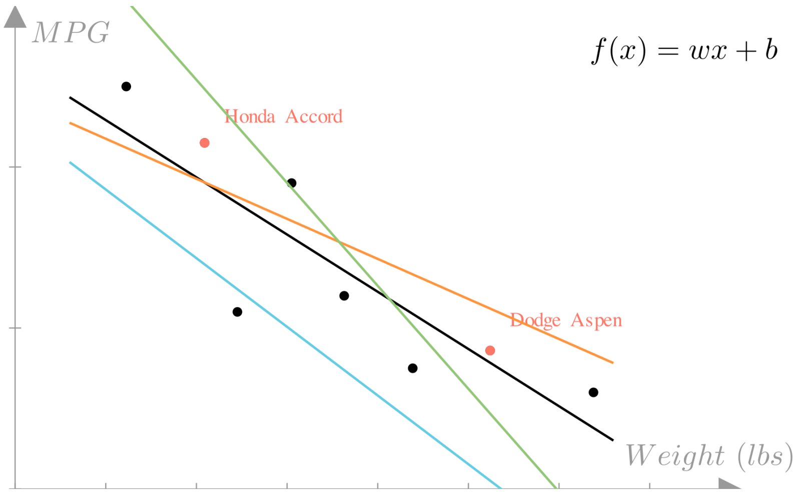


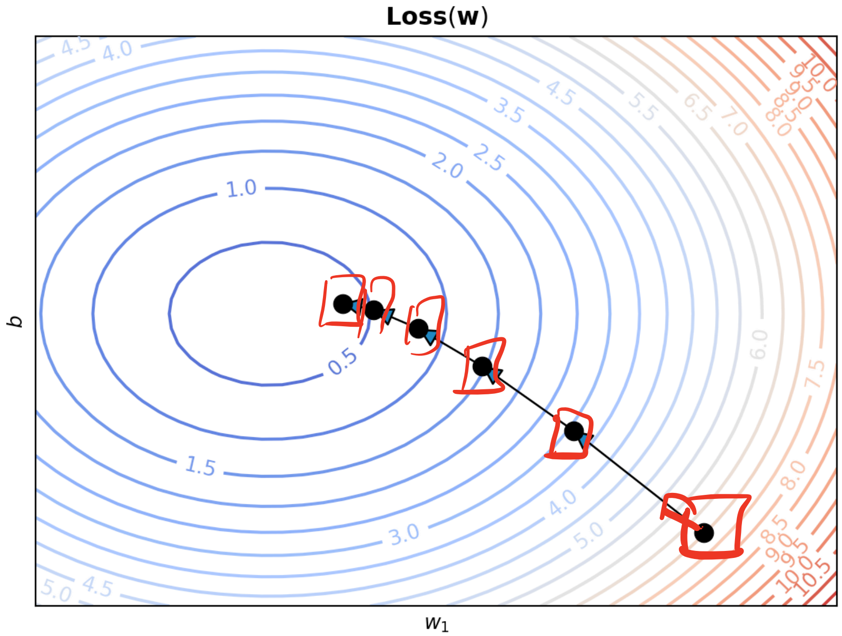
Dataset: $\mathcal{D} = \{(\mathbf{x}_i,\ y_i) \text{ for } i \in 1 \ldots N\}$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i) - y_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmin}} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2$$

While $||\nabla f(\mathbf{w}^{(i)})||_2 > \epsilon$: $\quad \mathbf{w}^{(i+1)} \leftarrow \mathbf{w}^{(i)} - \nabla f(\mathbf{w}^{(i)})$





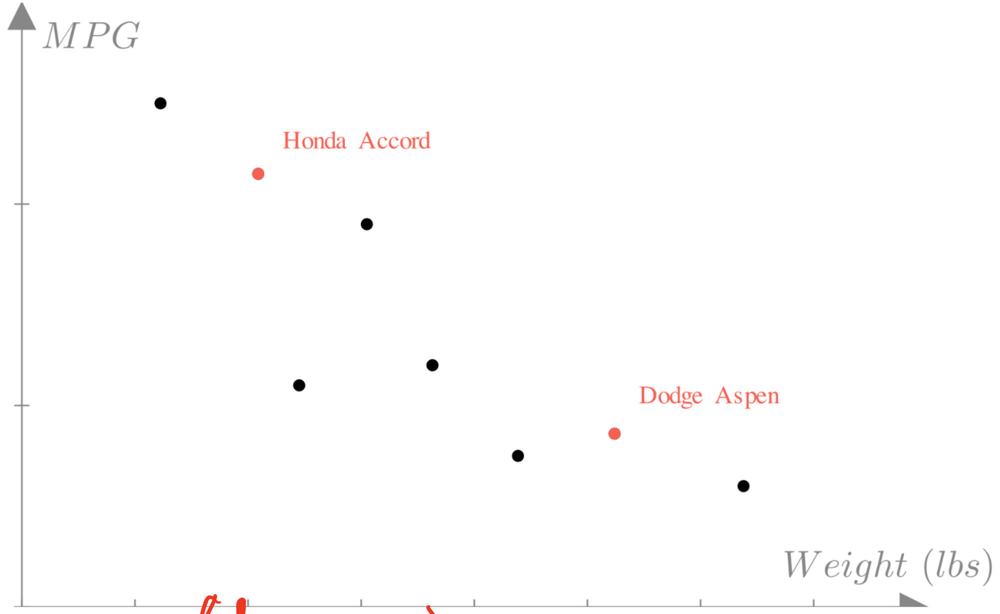
$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \end{bmatrix} = \begin{bmatrix} \text{weight}_i \\ \text{dis.}_i \\ \vdots \\ i \end{bmatrix}$$

Dataset: $\mathcal{D} = \{(\mathbf{x}_i,\ y_i) \text{ for } i \in 1 \dots N\}$

capital

measurement

$$\underline{\mathbf{X}} = \begin{bmatrix} \boxed{\mathbf{x}_1^T} \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} = \begin{array}{c} \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix} \\ \text{weight} \quad \text{displ.} \quad \dots \quad \text{acc.} \end{array} \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \text{obs.}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$



$$MSE = \frac{1}{N} \sum_{i=1}^{N} (f(\mathbf{x}_i) - y_i)^2 = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i)^2$$

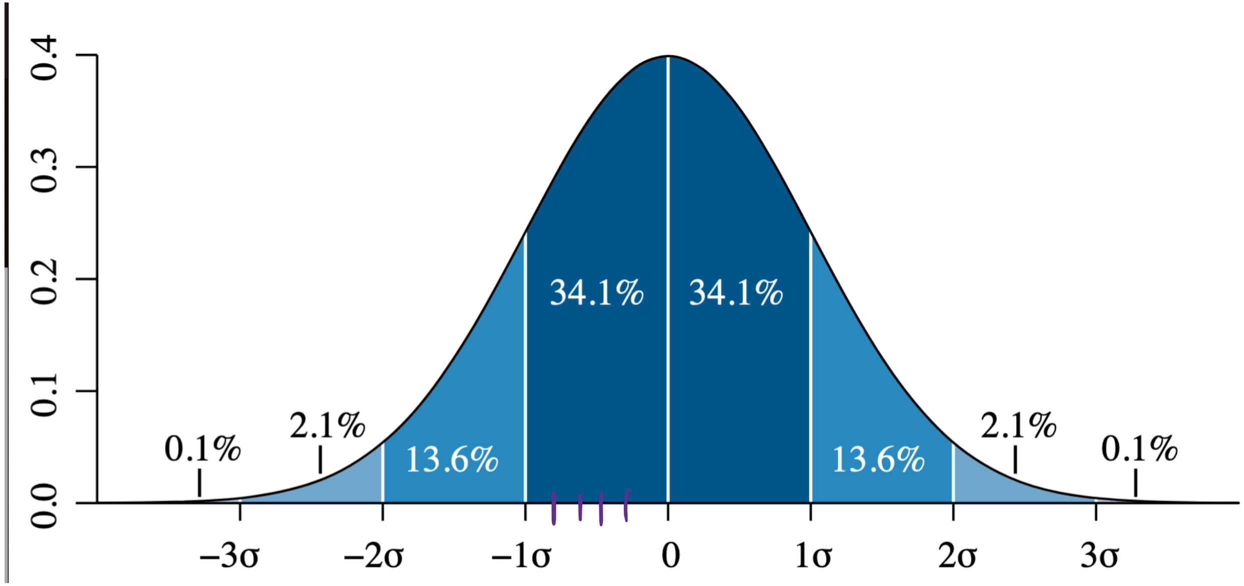
$$= \frac{1}{N} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

$$(Xw)_i = \left(\sum_{j=1}^{d} X_{ij} w_j\right) - y_i$$

Normal distribution

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{1}{2\sigma^2}(y-\mu)^2 \right)$$

e^

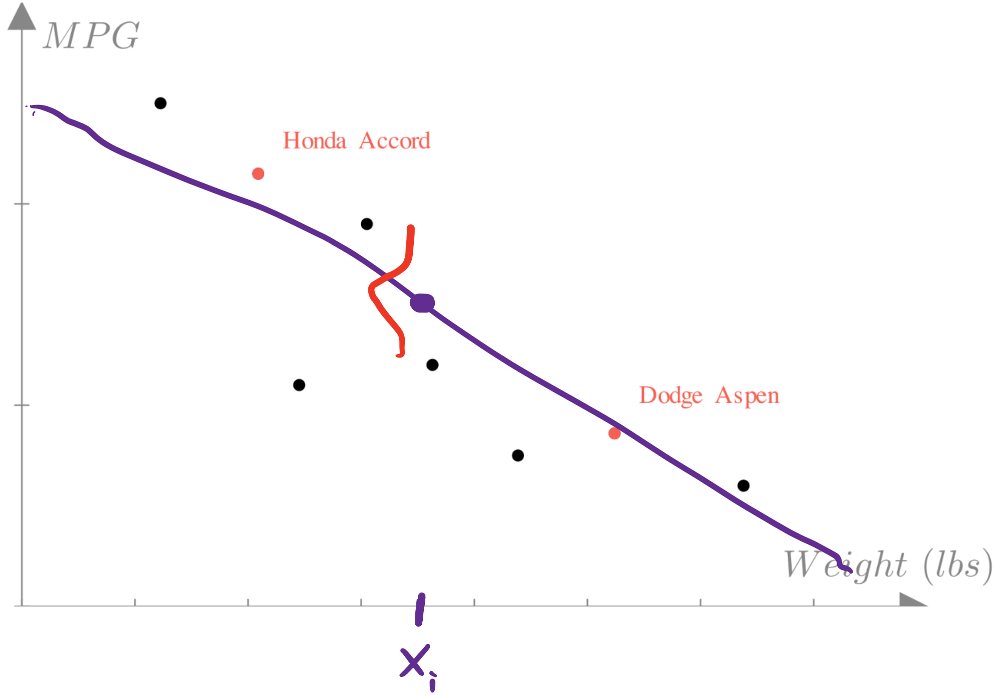


$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{1}{2\sigma^2} (y - \mu)^2 \right)$$

Assume

$$y_i \sim \mathcal{N}\left(\mathbf{x}_i^T \mathbf{w},\ \sigma^2\right)$$

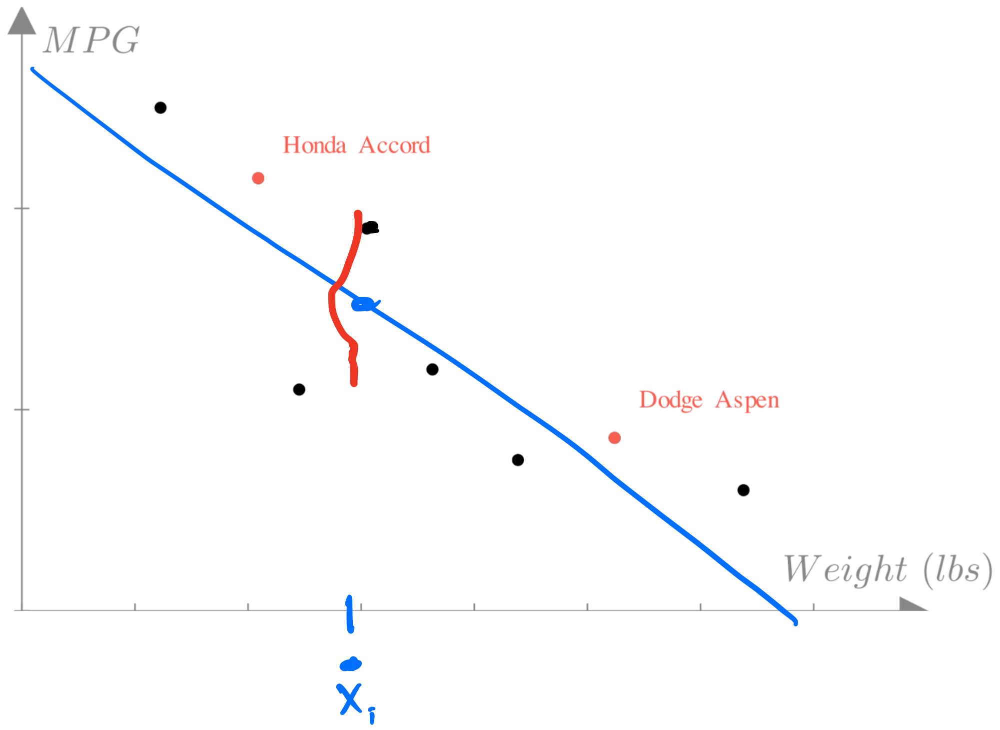
$$p(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{1}{2\sigma^2} (y_i - \mathbf{x}_i^T \mathbf{w})^2 \right)$$



$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2 \right)$$

$$y_i \sim \mathcal{N}\left(\mathbf{x}_i^T \mathbf{w},\ \sigma^2\right)$$

$$p(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2 \right)$$



$$e_i \sim \mathcal{N}\left(\mathbf{x}_i^T \mathbf{w} - y_i,\ \sigma^2\right)$$

$$p(e_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2 \right)$$

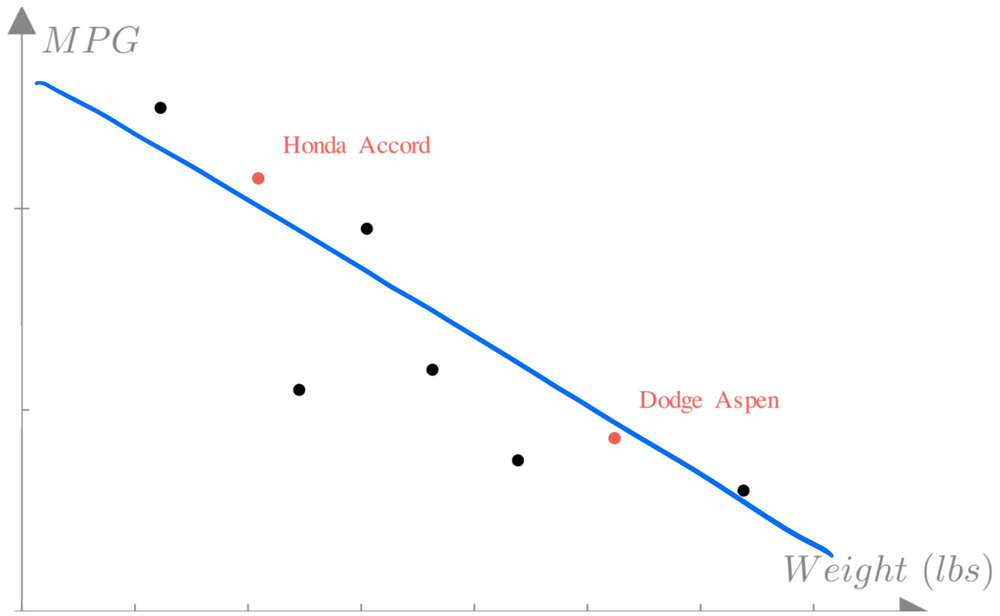
$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2 \right)$$

$$y_i \sim \mathcal{N}\left(\mathbf{x}_i^T\mathbf{w},\ \sigma^2\right)$$

$$p(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T\mathbf{w})^2 \right)$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(y_1, \dots, y_N \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{w})$$

Maximum likelihood

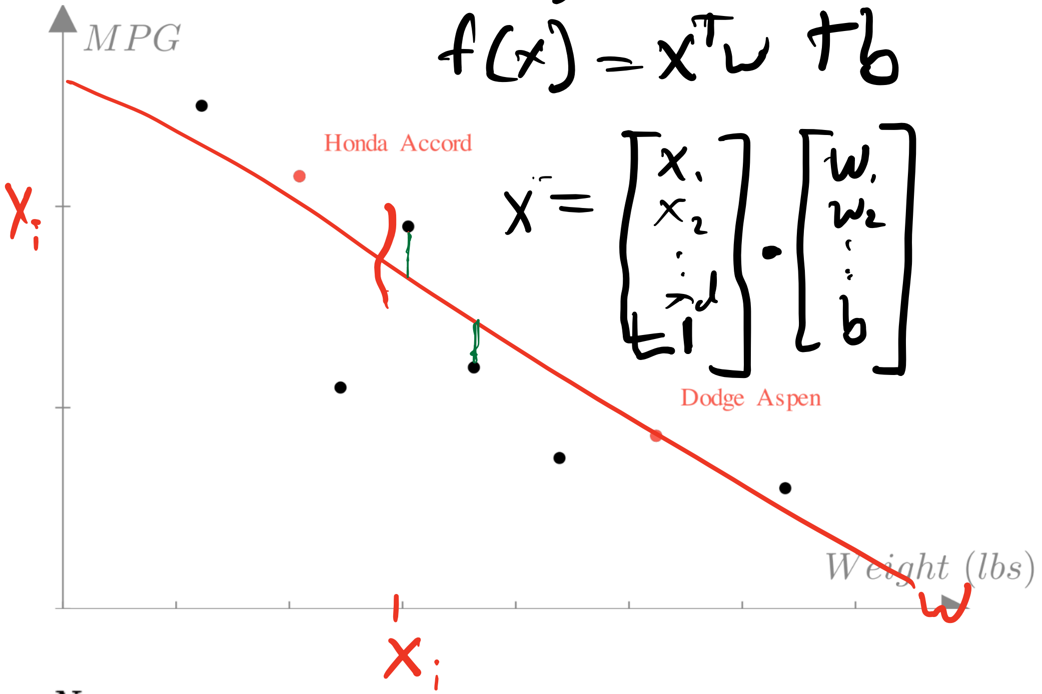


$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(y_1, \dots, y_N \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{w})$$

$f(x) = x^T w$

$f(x) = x^T w + b$

$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \\ 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ b \end{bmatrix}$



Normal assumption

$$p(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2 \right)$$

Independence

$$p(y_1, \dots, y_N \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{w}) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

Negative log likelihood

$$\mathbf{Loss}(\mathbf{w}) = \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$\underbrace{\max_{w^*} P(y_1 \ldots y_2 \mid x_1 \ldots x_N, w)}\; = \prod_{i=1}^{N} P(y_i \mid x_i, w)$$

$$\equiv \max_{w} \log \prod P(y_i \mid x_i, w)$$

$$\equiv \max_{w} \sum_{i=1}^{N} \log P(y_i \mid x_i, w)$$

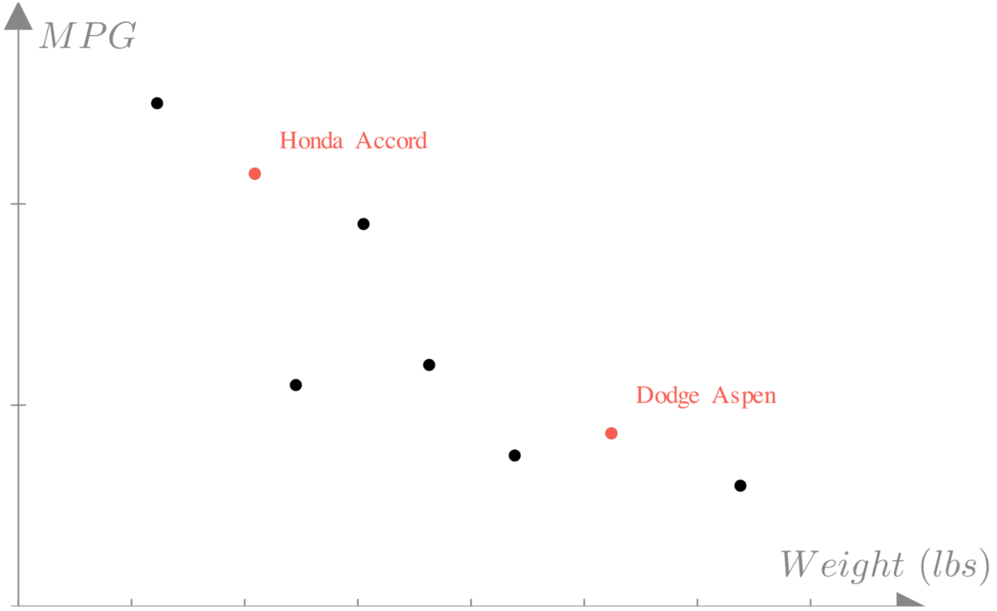
$$\equiv \min_{w} - \sum_{i=1}^{N} \log P(y_i \mid x_i, w) = NLL(x, y, w)$$

Negative log likelihood

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(y_1, \dots, y_N \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{w})$$

$$p(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T\mathbf{w})^2\right)$$

$$p(y_1, \dots, y_N \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{w}) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w})$$



$$\mathbf{Loss}(\mathbf{w}) = \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

we will use natural logs

$\log x = \ln x$

$\log e^x = x$

$\log \frac{1}{a} = -\log a$

$$\mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log \left[\underbrace{\frac{1}{\sigma\sqrt{2\pi}}} \underbrace{\exp\left(-\frac{1}{2\sigma^2} (y_i - \mathbf{x}_i^T \mathbf{w})^2 \right)} \right]$$

$\log ab = \log a + \log b$

$$= -\sum \left[\underbrace{\log \frac{1}{\sigma\sqrt{2\pi}}} + \underbrace{\cancel{\log} \cancel{\exp}}\left(-\frac{1}{2\sigma^2} (y_i - x_i^T w)^2 \right) \right]$$

$$= \underbrace{\sum \log \sigma\sqrt{2\pi} + \left(\frac{1}{2\sigma^2}\right.}(x_i - x_i^T w)^2 \Big)$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T w)^2 + N \log \sigma\sqrt{2\pi} = NLL$$

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

$w^{(i)} \leftarrow w^{(i-1)} - \nabla \mathrm{Loss}(w)$

$$\underbrace{\nabla_{\mathbf{w}} \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y})} = \frac{d}{d\mathbf{w}} \left(\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^T \mathbf{w})^2 + \underbrace{\cancel{N \log \sigma \sqrt{2\pi}}}_{0} \right)$$

$$= \frac{1}{2\sigma^2} \frac{d}{dw} \sum_{i=1}^{N} (y_i - x_i^T w)^2$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{N} \frac{d}{dw} (y_i - x_i^T w)^2$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{N} 2(y_i - x_i^T w) \frac{d}{dw} (y_i - x_i^T w)$$

$$= \boxed{\frac{1}{2\sigma^2}} \sum 2(y_i - x_i^T w)(-x_i)$$

NLL $\rightarrow$ MSS $\frac{1}{N}$

at w^*

$$0 = \frac{1}{\cancel{2}\sigma^2} \sum 2(y_i - x_i^T w)(-x_i)$$

$$\underset{\mathbf{w}}{\operatorname{argmin}}\, MSE(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y})$$

f() ⟶ {Cat, Dog}

Logistic regression

$$y = f(\mathbf{x}), \quad \text{Input: } \mathbf{x} \in \mathbb{R}^n \longrightarrow \text{Output: } \underbrace{y \in \{0, 1\}}$$

Linear Reg.

Input $x \in \mathbb{R}^n$

Vector w/ n entries

output: $y \in \mathbb{R}$

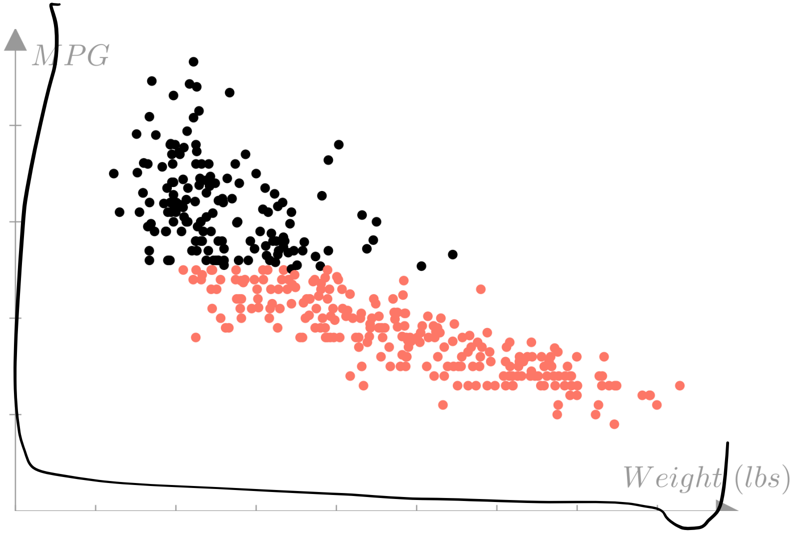
$$y = f(\mathbf{x}), \quad \text{Input: } \mathbf{x} \in \mathbb{R}^n \longrightarrow \text{Output: } y \in \{0, 1\}$$

$$\text{Input: } \mathbf{x}_i = \begin{bmatrix} \text{Weight} \\ \text{Horsepower} \\ \text{Displacement} \\ \text{0-60mph} \end{bmatrix}, \quad \text{Output: } y_i = \begin{cases} 1 : \text{Meets target } (MPG \geq 30) \\ 0 : \text{Fails to meet target } (MPG < 30) \end{cases}$$

Dataset:

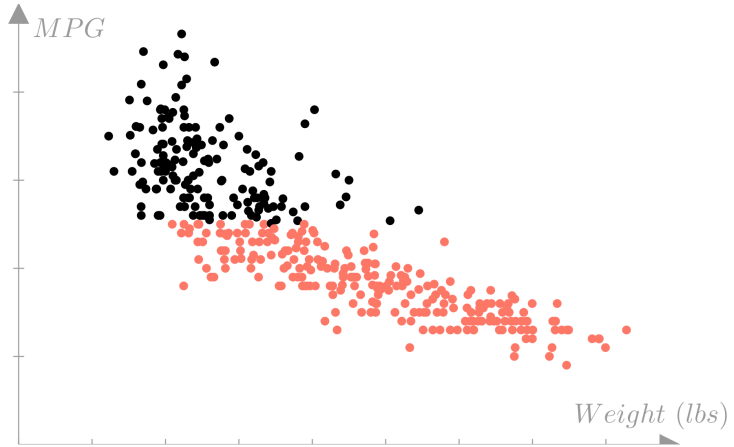
$$\text{Honda Accord: } \begin{bmatrix} \text{Weight:} & \text{2500 lbs} \\ \text{Horsepower:} & \text{123 HP} \\ \text{Displacement:} & \text{2.4 L} \\ \text{0-60mph:} & \text{7.8 Sec} \end{bmatrix} \longrightarrow 1 \text{ (Meets target)}$$

$$\text{Dodge Aspen: } \begin{bmatrix} \text{Weight:} & \text{3800 lbs} \\ \text{Horsepower:} & \text{155 HP} \\ \text{Displacement:} & \text{3.2 L} \\ \text{0-60mph:} & \text{6.8 Sec} \end{bmatrix} \longrightarrow 0 \text{ (Does not meet target)}$$



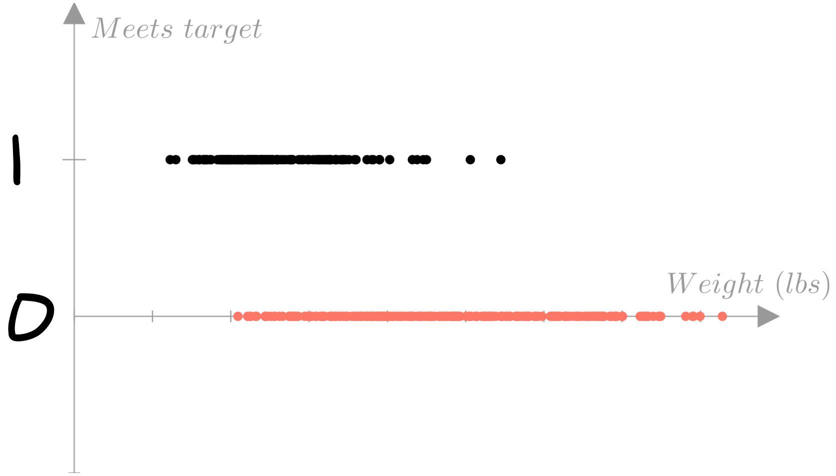
$$y = f(\mathbf{x}), \quad \text{Input: } \mathbf{x} \in \mathbb{R}^n \longrightarrow \text{ Output: } y \in \{0, 1\}$$

$$\text{Input: } \mathbf{x}_i = \begin{bmatrix} \text{Weight} \\ \text{Horsepower} \\ \text{Displacement} \\ \text{0-60mph} \end{bmatrix}, \quad \text{Output: } y_i = \begin{cases} 1 : \text{Meets target } (MPG \geq 30) \\ 0 : \text{Fails to meet target } (MPG < 30) \end{cases}$$



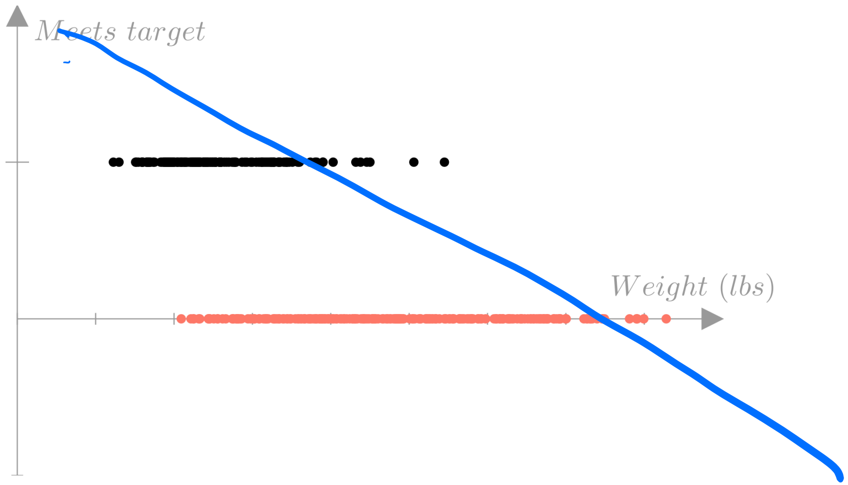
$$\text{Honda Accord: } \begin{bmatrix} \text{Weight:} & \text{2500 lbs} \\ \text{Horsepower:} & \text{123 HP} \\ \text{Displacement:} & \text{2.4 L} \\ \text{0-60mph:} & \text{7.8 Sec} \end{bmatrix} \longrightarrow 1 \text{ (Meets target)}$$

$$\text{Dodge Aspen: } \begin{bmatrix} \text{Weight:} & \text{3800 lbs} \\ \text{Horsepower:} & \text{155 HP} \\ \text{Displacement:} & \text{3.2 L} \\ \text{0-60mph:} & \text{6.8 Sec} \end{bmatrix} \longrightarrow 0 \text{ (Does not meet target)}$$



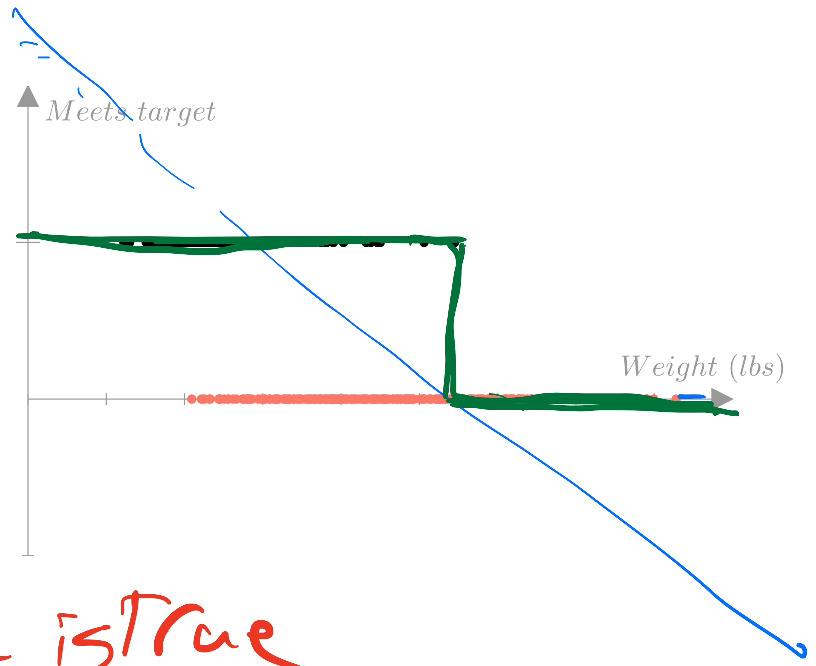
assume $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ 1 \end{bmatrix}$ $w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ b \end{bmatrix}$ $\rightarrow f(x) = x^T w$

$$f(\mathbf{x}) = \underbrace{\mathbf{x}^T\mathbf{w}} + \underbrace{b} = \sum_{i=1}^{d} x_i w_i + b$$



$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} \quad \longrightarrow \quad f(\mathbf{x}) = \begin{cases} 1 \text{ if } \mathbf{x}^T\mathbf{w} > 0 \\ 0 \text{ if } \mathbf{x}^T\mathbf{w} < 0 \end{cases}$$

Threshold



Indicator function

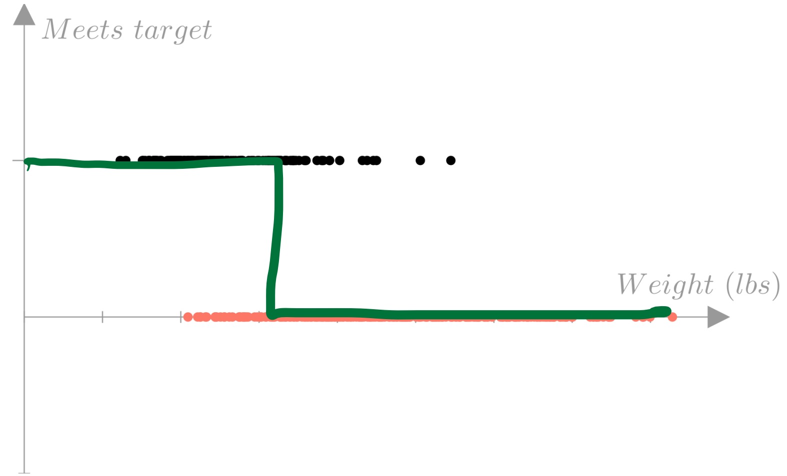
$$f(\mathbf{x}) = \mathbb{I}(\mathbf{x}^T\mathbf{w} \geq 0)$$

$\|$

$$\mathbb{I}(a) = \begin{cases} 1 \text{ if } a \text{ is true} \\ 0 \text{ otherwise} \end{cases}$$

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} \quad \longrightarrow \quad f(\mathbf{x}) = \begin{cases} 1 \text{ if } \mathbf{x}^T\mathbf{w} \geq 0 \\ 0 \text{ if } \mathbf{x}^T\mathbf{w} < 0 \end{cases}$$

$$f(\mathbf{x}) = \mathbb{I}(\mathbf{x}^T\mathbf{w} \geq 0)$$



$$\text{Meets target} = f(\mathbf{x}) =$$

$$\big((\text{weight})w_1 + (\text{horsepower})w_2 + (\text{displacement})w_3 + (\text{0-60mph})w_4 + b\big) \geq 0$$

$$f(\mathbf{x}) = \left(\begin{bmatrix} \text{Weight} \\ \text{Horsepower} \\ \text{Displacement} \\ \text{0-60mph} \\ 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ b \end{bmatrix} \geq 0 \right)$$

X — Inputs

w — parameters

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix}$$

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} \quad \longrightarrow \quad f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x}^T\mathbf{w} \geq 0 \\ 0 & \text{if } \mathbf{x}^T\mathbf{w} < 0 \end{cases}$$

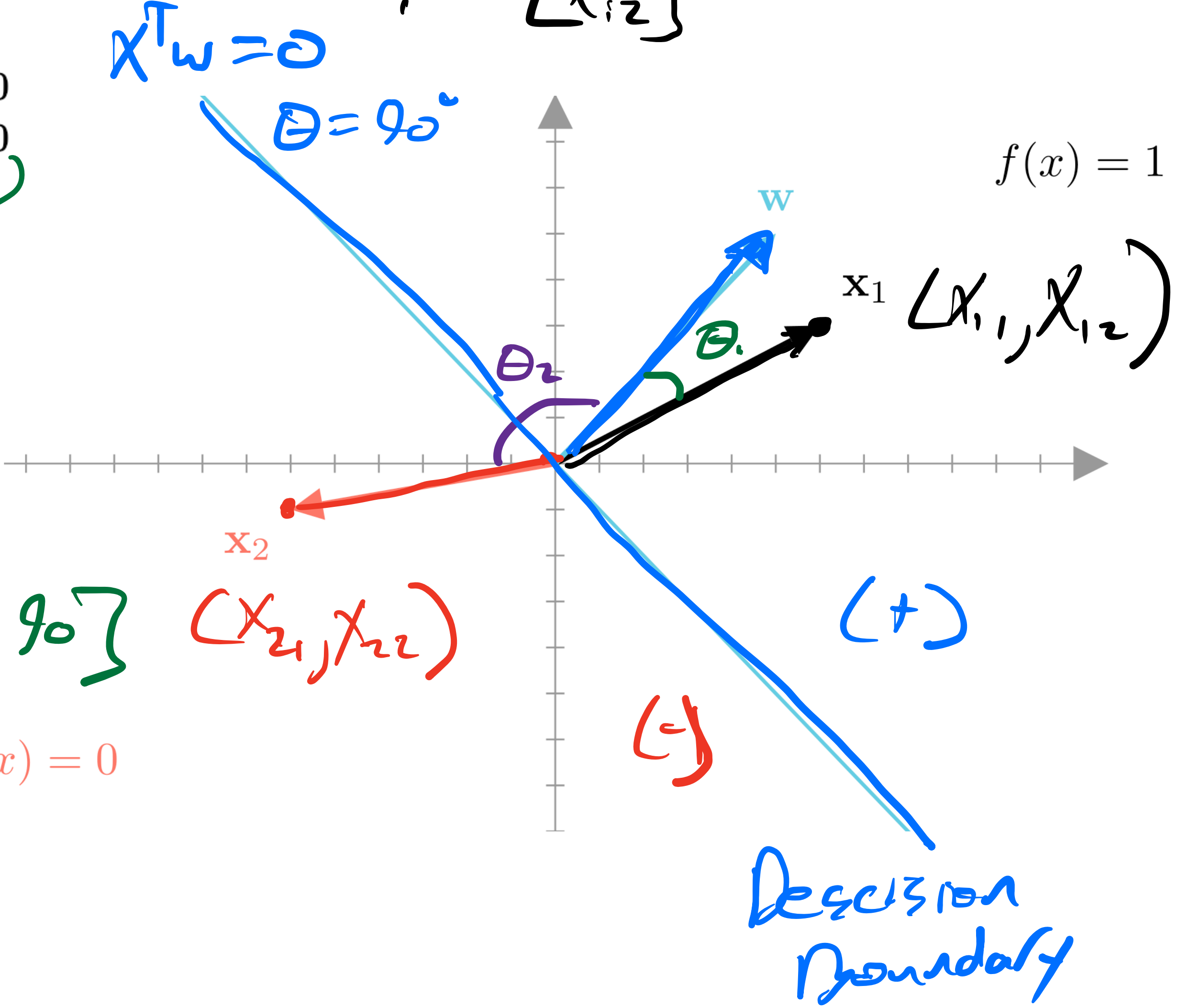
$$\mathbf{x}^T\mathbf{w} = ||\mathbf{x}||_2 ||\mathbf{w}||_2 \cos\theta$$

"

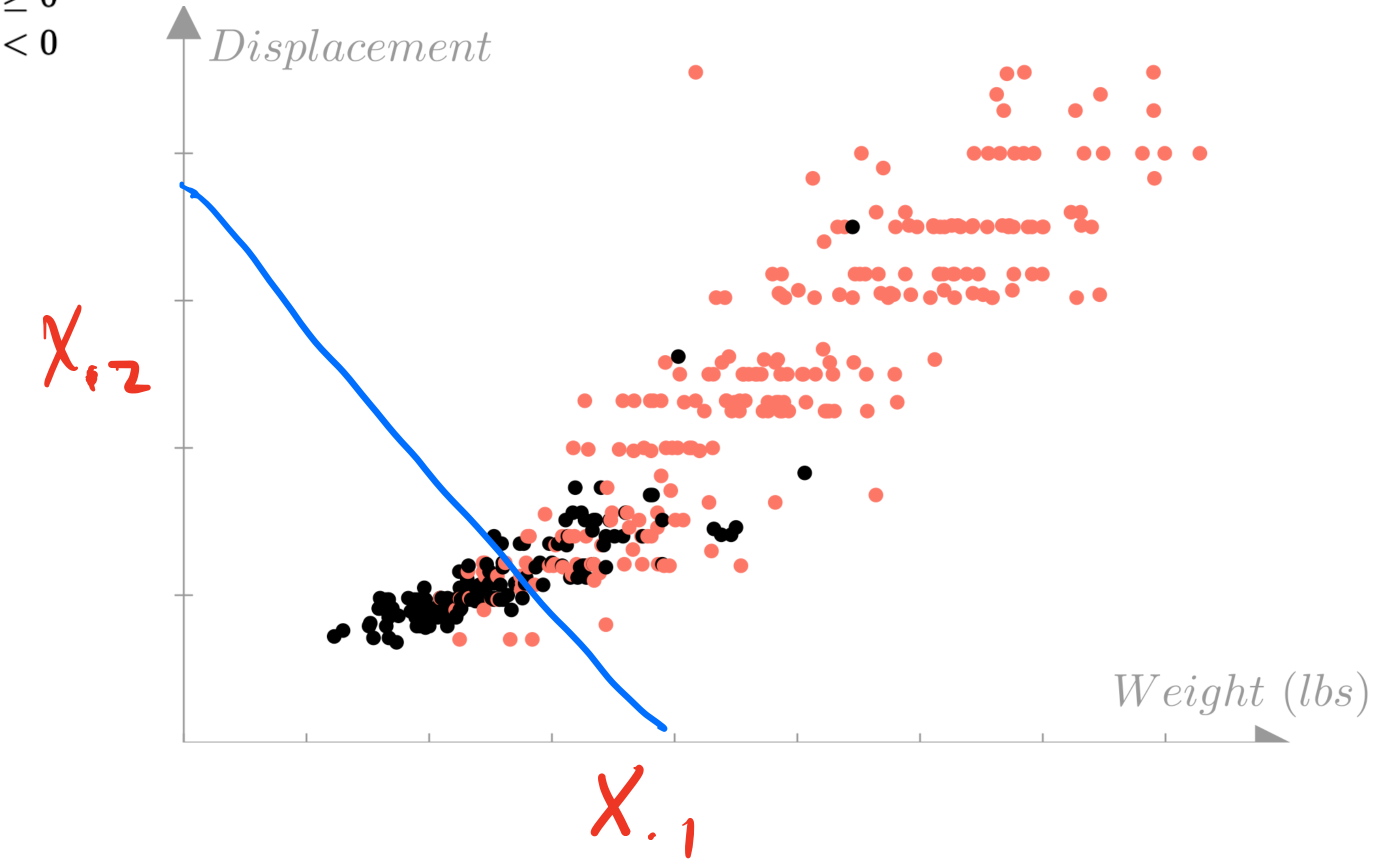
$x_i w$

$x^T w$ (+) if $\theta \in [-90, 90]$

(−) otherwise $f(x) = 0$



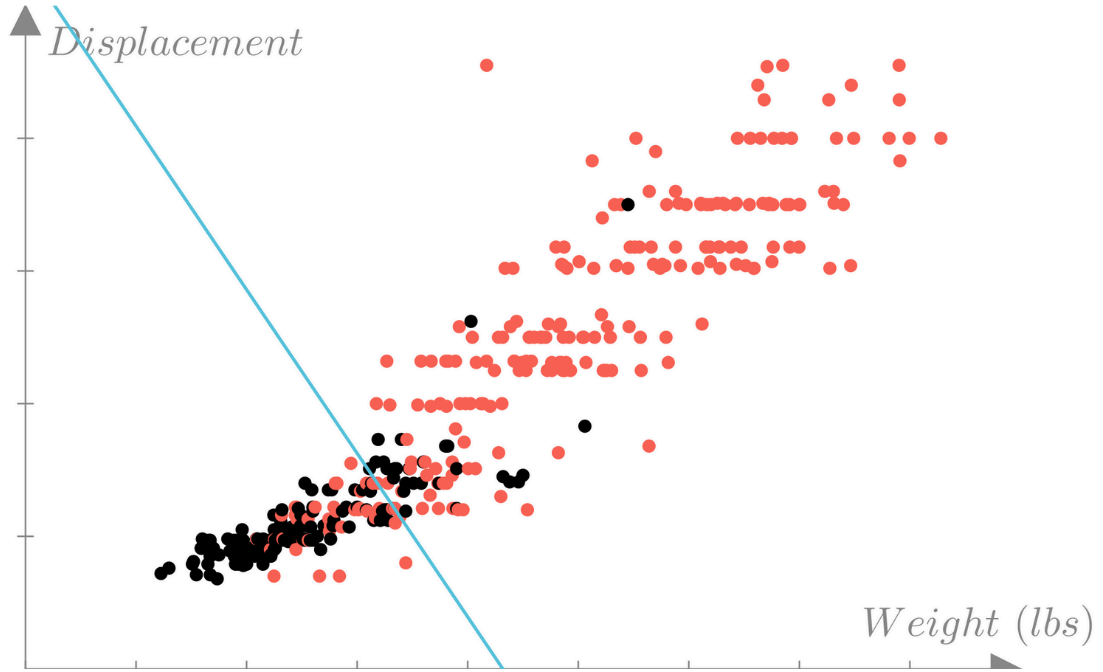
$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} \quad \longrightarrow \quad f(\mathbf{x}) = \begin{cases} 1 \text{ if } \mathbf{x}^T\mathbf{w} \geq 0 \\ 0 \text{ if } \mathbf{x}^T\mathbf{w} < 0 \end{cases}$$



Accuracy: $\dfrac{\text{\# of correct predictions}}{\text{Total predictions}}$

$$\textbf{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(f(\mathbf{x}_i) = y_i\big)$$

Accuracy: 0.8291



Probability of **heads:** q, Probability of **tails:** $1 - q$

$$p(y) = \begin{cases} q & \text{if } y = 1 \\ 1 - q & \text{if } y = 0 \end{cases} \quad q \in [0, 1],\ y \in \{0, 1\}$$

Probability of **heads:** q, Probability of **tails:** $1 - q$

$$p(y) = \begin{cases} q & \text{if } y = 1 \\ 1 - q & \text{if } y = 0 \end{cases} \quad q \in [0, 1],\ y \in \{0, 1\}$$

$$p(y) = q^y (1 - q)^{1-y}$$

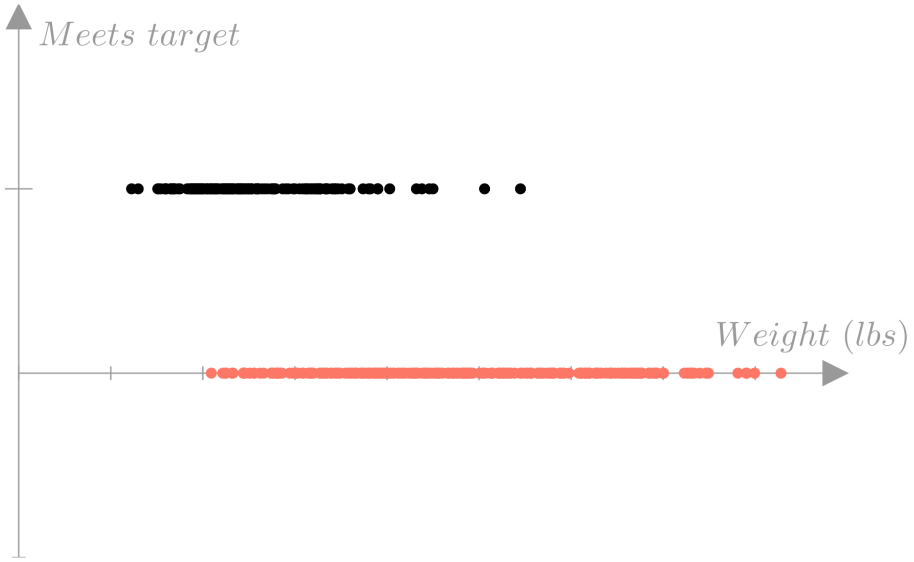
$$\log p(y) = y \log q + (1 - y) \log(1 - q)$$

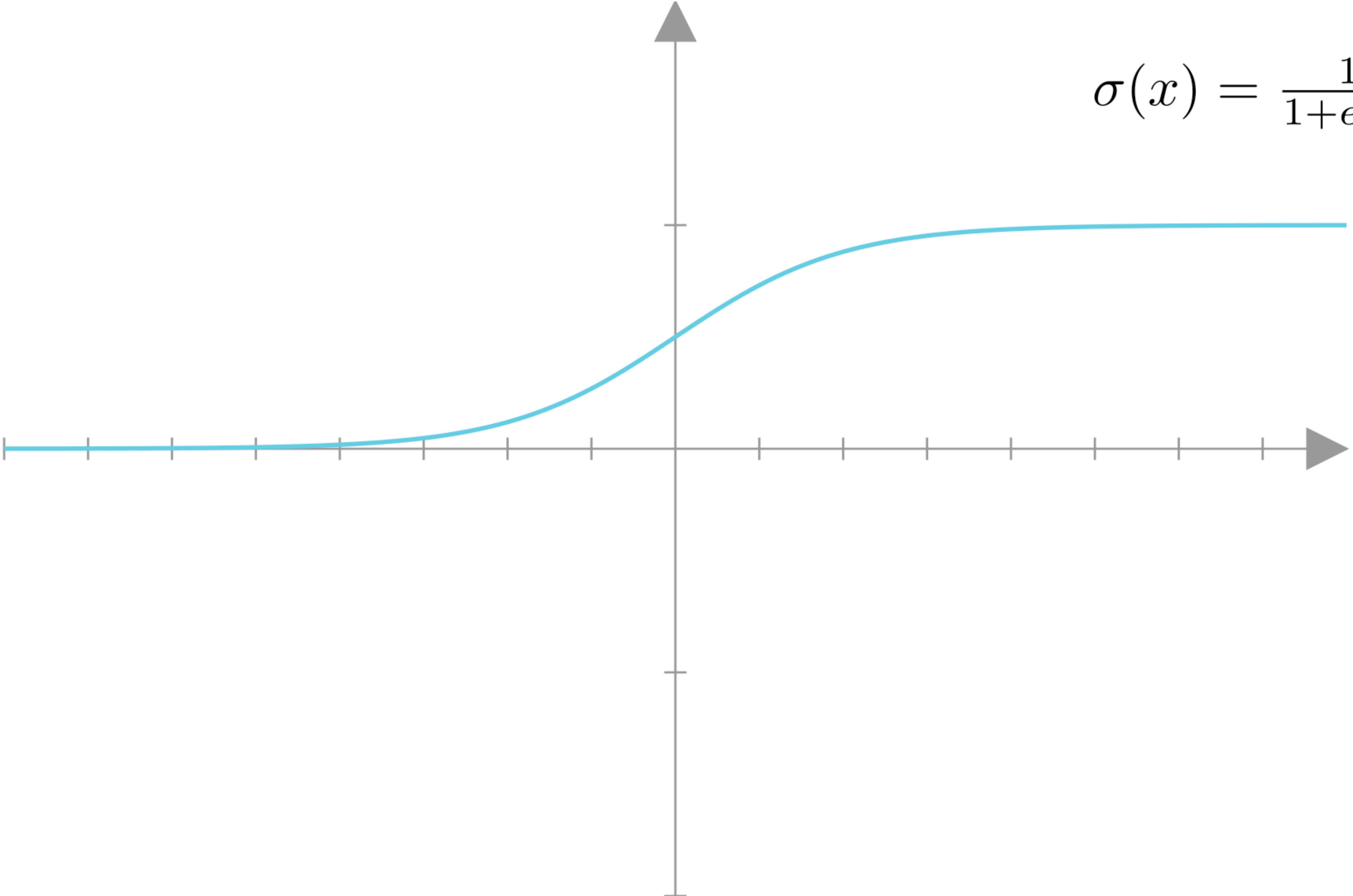
$$y_i \sim \mathcal{N}(\mathbf{x}_i^T\mathbf{w},\ \sigma^2)$$

$$\mathbf{x}^T\mathbf{w} \notin [0,1] \quad \longrightarrow \quad y_i \sim \textbf{Bernoulli}(\mathbf{q} = ?)$$

$$\textbf{Need}\ \ g(x):\ \mathbb{R} \longrightarrow [0,1]$$

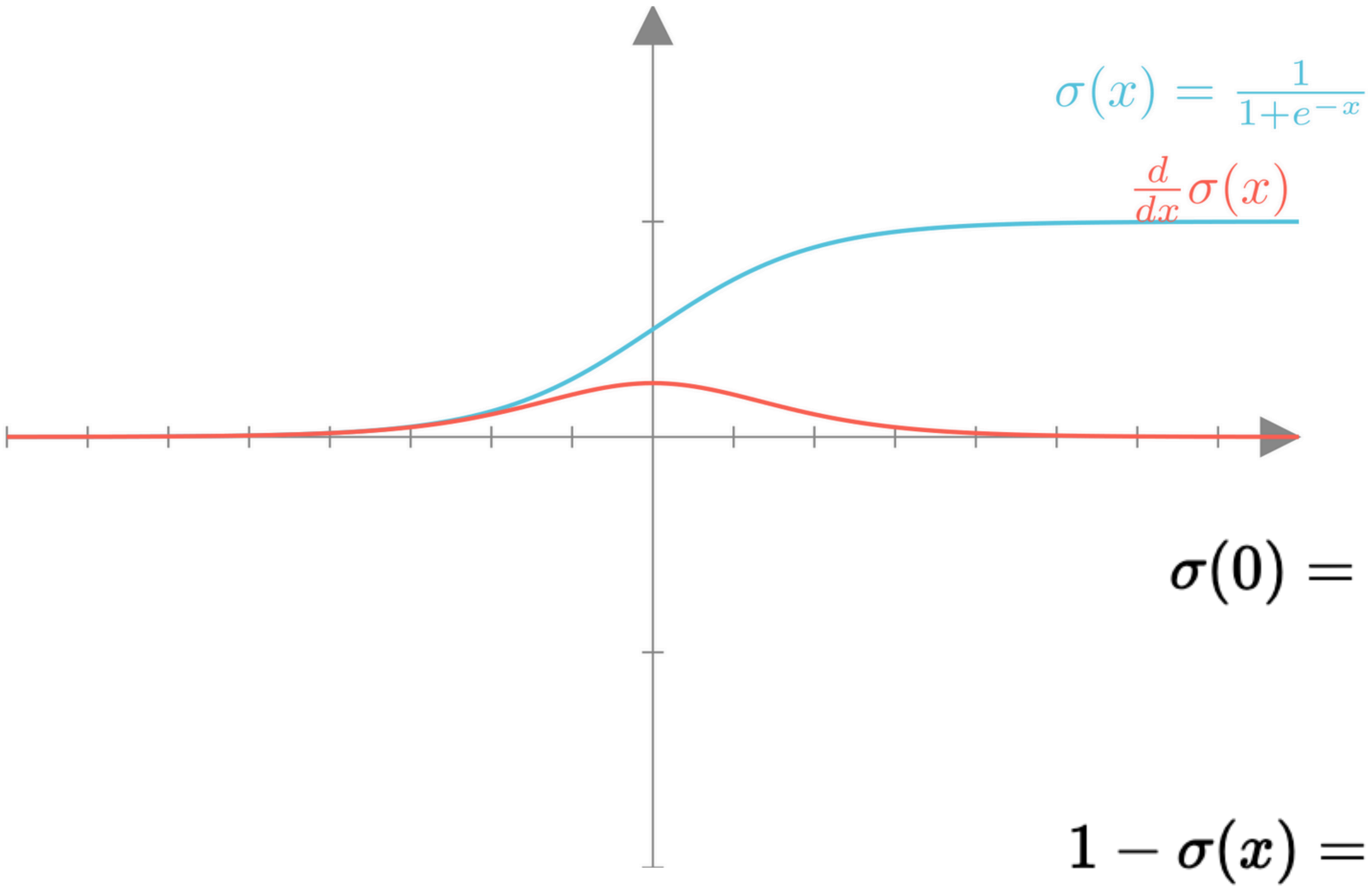
$$\textbf{Input:}\ x \in \mathbb{R} \longrightarrow \textbf{Output:}\ y \in [0,1]$$





$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)\big(1 - \sigma(x)\big)$$



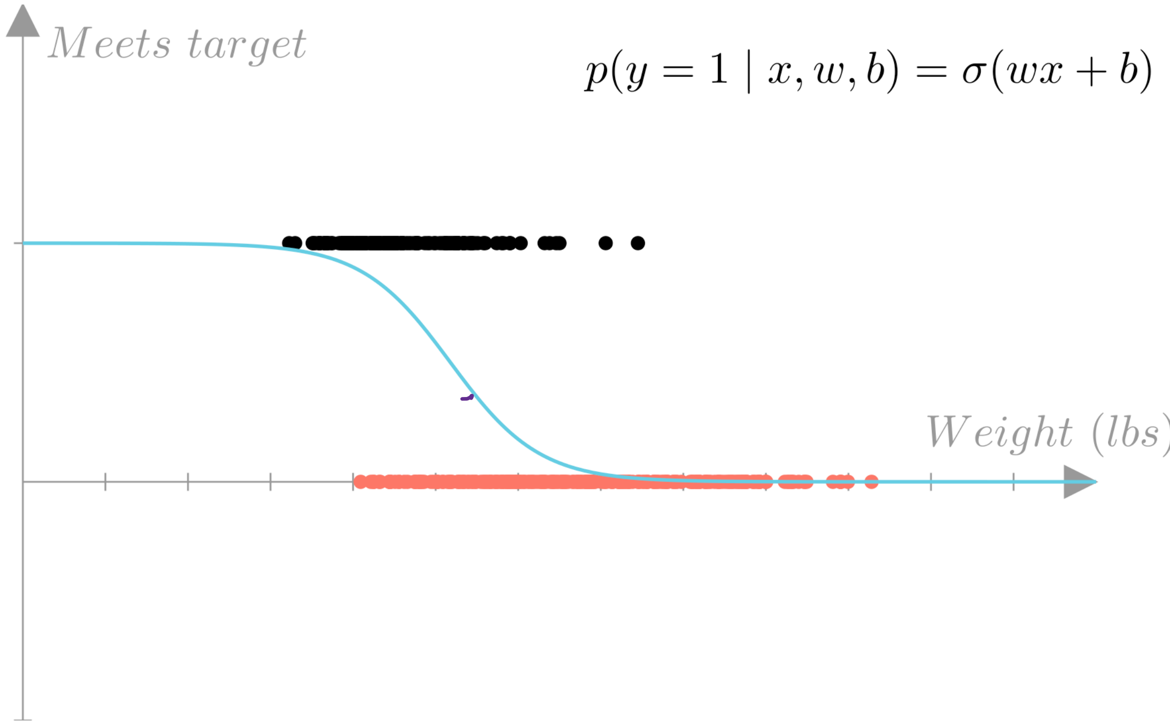
$$\sigma(0) = 0.5$$

$$1 - \sigma(x) = \sigma(-x)$$

$$y_i \sim \textbf{Bernoulli}(\sigma(\mathbf{x_i^T w}))$$

$$p(y_i = 1 \mid \mathbf{x}_i, \mathbf{w}) = \sigma(\mathbf{x}_i^T \mathbf{w}),$$

$$p(y_i = 0 \mid \mathbf{x}_i, \mathbf{w}) = 1 - \sigma(\mathbf{x}_i^T \mathbf{w}) = \sigma(-\mathbf{x}_i^T \mathbf{w})$$

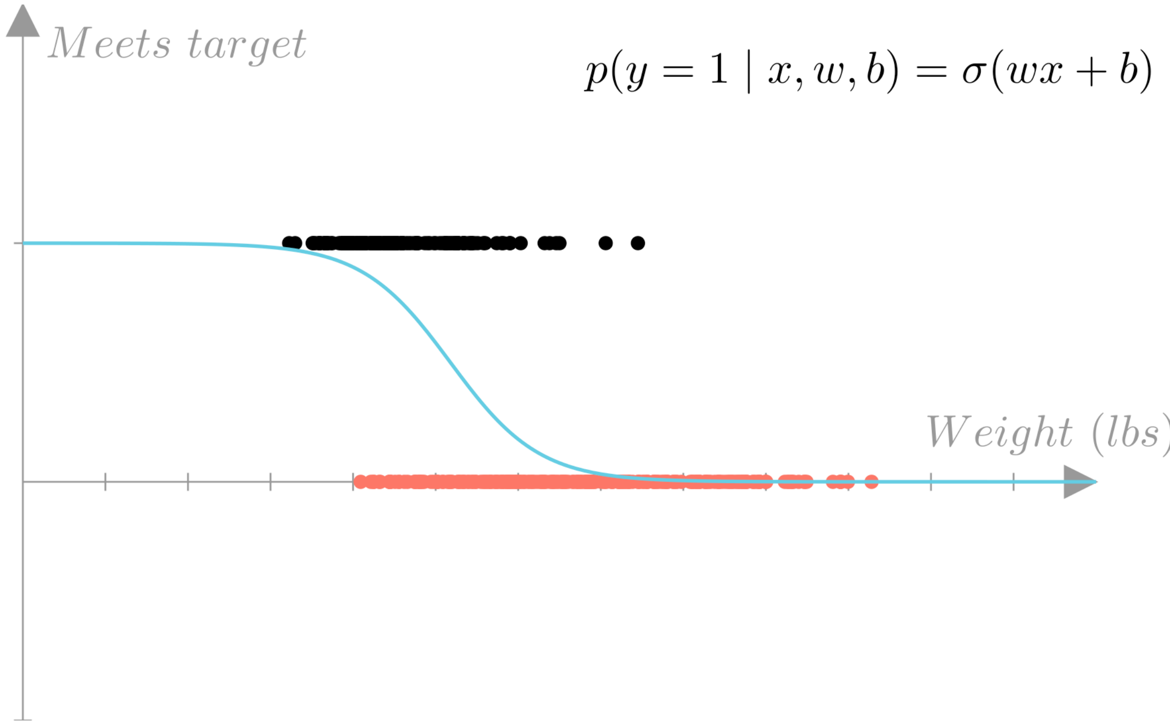


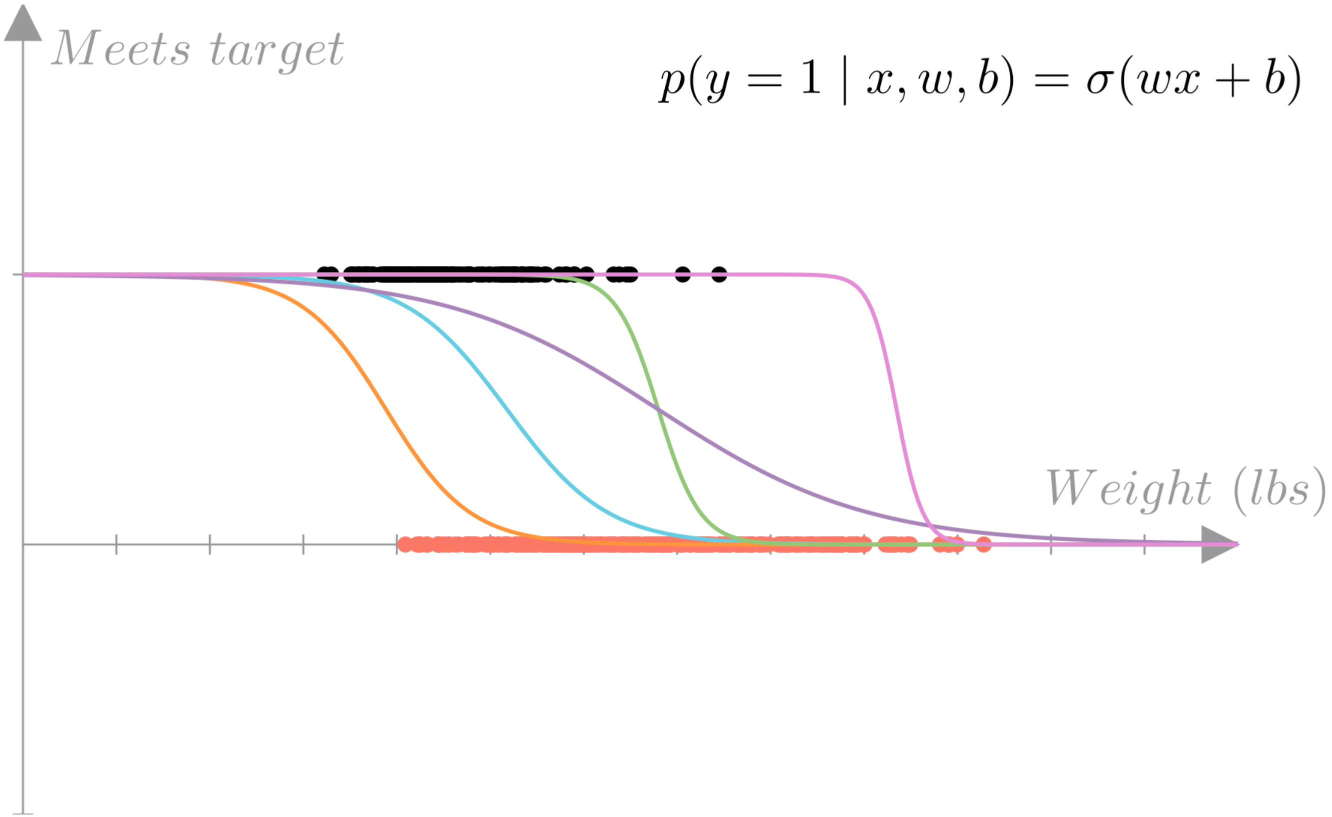
$$\textbf{Prediction function: } f(\mathbf{x}) = \begin{cases} 1 \text{ if } p(y = 1 \mid \mathbf{x}, \mathbf{w}) \geq p(y = 0 \mid \mathbf{x}, \mathbf{w}) \\ 0 \text{ otherwise} \end{cases}$$

Prediction function: $f(\mathbf{x}) = \begin{cases} 1 \text{ if } p(y=1 \mid \mathbf{x}, \mathbf{w}) \geq p(y=0 \mid \mathbf{x}, \mathbf{w}) \\ 0 \text{ otherwise} \end{cases}$

$$p(y=1 \mid \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}^T\mathbf{w}) \geq 0.5$$

$$p(y_i = 1) \geq 0.5 \quad \longrightarrow \quad \mathbf{x}^T\mathbf{w} \geq 0$$





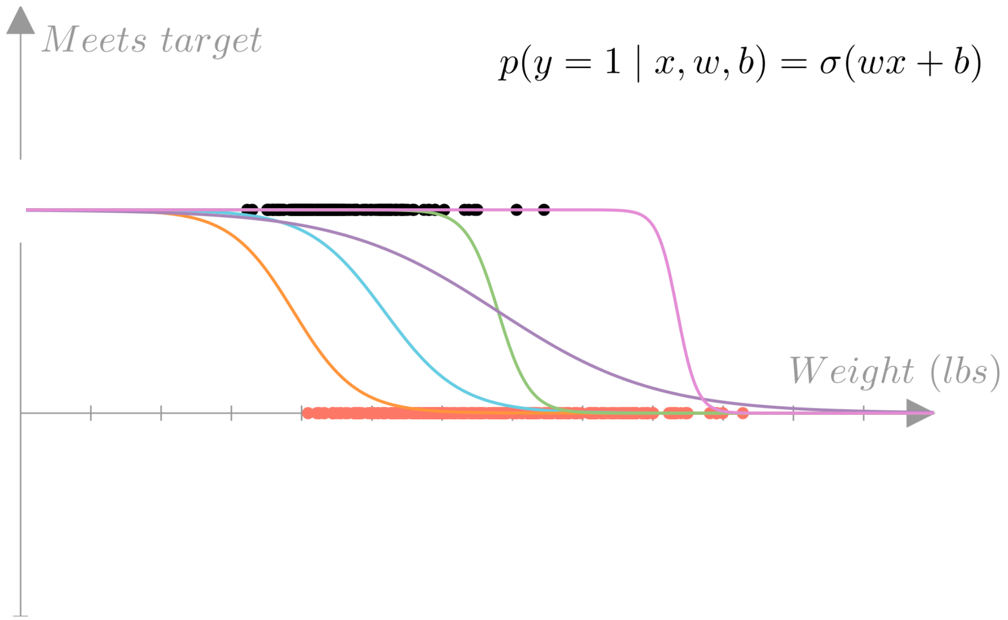
$$p(y = 1 \mid x, w, b) = \sigma(wx + b)$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w})$$

$$\mathbf{Loss}(\mathbf{w}) = \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$p(y_i = 1 \mid \mathbf{x}_i, \mathbf{w}) = \sigma(\mathbf{x}_i^T \mathbf{w}), \quad p(y_i = 0 \mid \mathbf{x}_i, \mathbf{w}) = 1 - \sigma(\mathbf{x}_i^T \mathbf{w}) = \sigma(-\mathbf{x}_i^T \mathbf{w})$$

$$\mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \left[y_i \log \sigma(\mathbf{x}_i^T \mathbf{w}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}_i^T \mathbf{w})) \right]$$



$$\mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \left[y_i \log \sigma(\mathbf{x}_i^T \mathbf{w}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}_i^T \mathbf{w})) \right]$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbf{Loss}(\mathbf{w})$$

$$\nabla_{\mathbf{w}}\mathbf{MSE}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{2}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)\mathbf{x}_i$$

$$\nabla_{\mathbf{w}}\mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)\mathbf{x}_i$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbf{Loss}(\mathbf{w})$$

$$\nabla_{\mathbf{w}} \mathbf{MSE}(\mathbf{w}, \mathbf{X}, \mathbf{y}) \quad = \frac{2}{N} \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i)\mathbf{x}_i$$

$$\nabla_{\mathbf{w}} \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) \quad = \frac{1}{2\sigma^2} \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i)\mathbf{x}_i$$

$$\mathbf{0} = \left(\frac{2}{N}\right) \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i)\mathbf{x}_i$$

$$\underset{\mathbf{w}}{\operatorname{argmin}}\ MSE(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y})$$