Binary Classification

f() ⟶ {Cat, Dog}

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

Linear Reg.

$\mathbb{R}$

Prediction:

predict y as $f(x)$

e.g. efficient? (e.g. horsepower)

$$f(x) = I(wx + b > 0)$$

Vector:

$$f(x) = I(w^T x > 0)$$

$$= I\left(\begin{bmatrix} w \\ b \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} > 0\right)$$

Indicator: true → 1, false → 0

Prediction:

Predict y as $f(x)$

e.g. efficient? e.g. horsepower

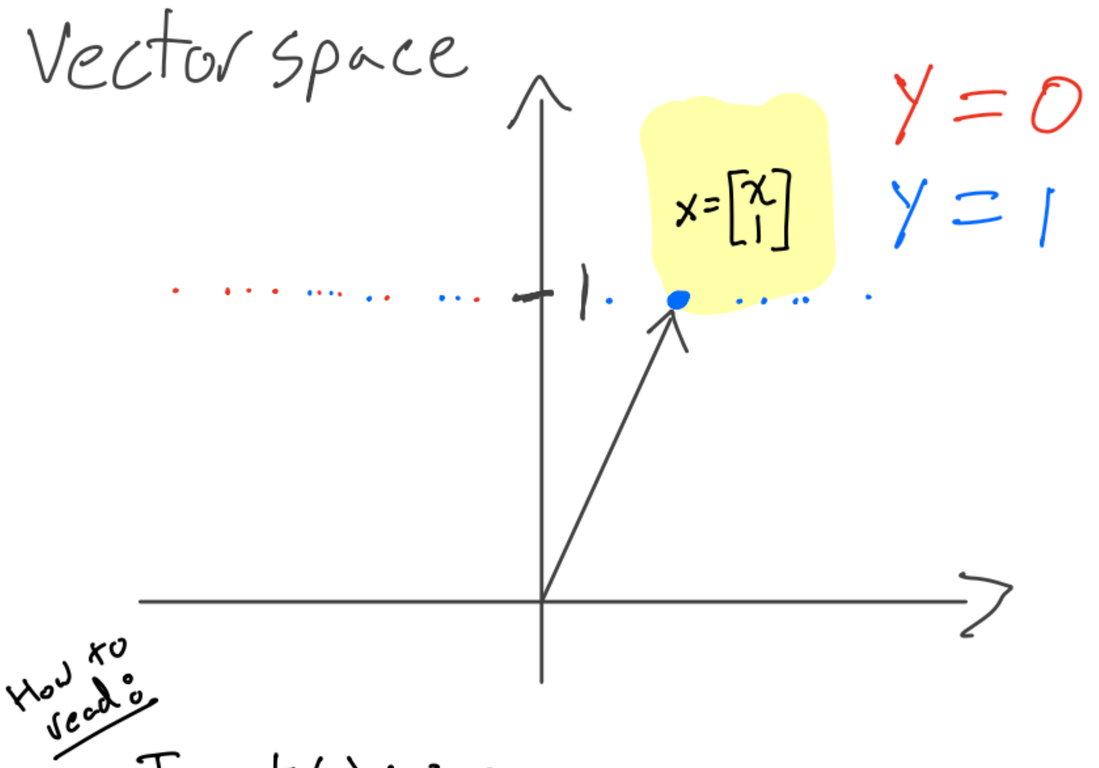
$$f(x) = I(wx + b > 0)$$

Vector:

$$f(x) = I(w^T x > 0)$$

$$= I\left(\begin{bmatrix} w \\ b \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} > 0\right)$$

Indicator: true → 1, false → 0



Prediction:

predict y as $f(x)$

e.g. efficient? (e.g. horsepower)

$$f(x) = I(wx + b > 0)$$

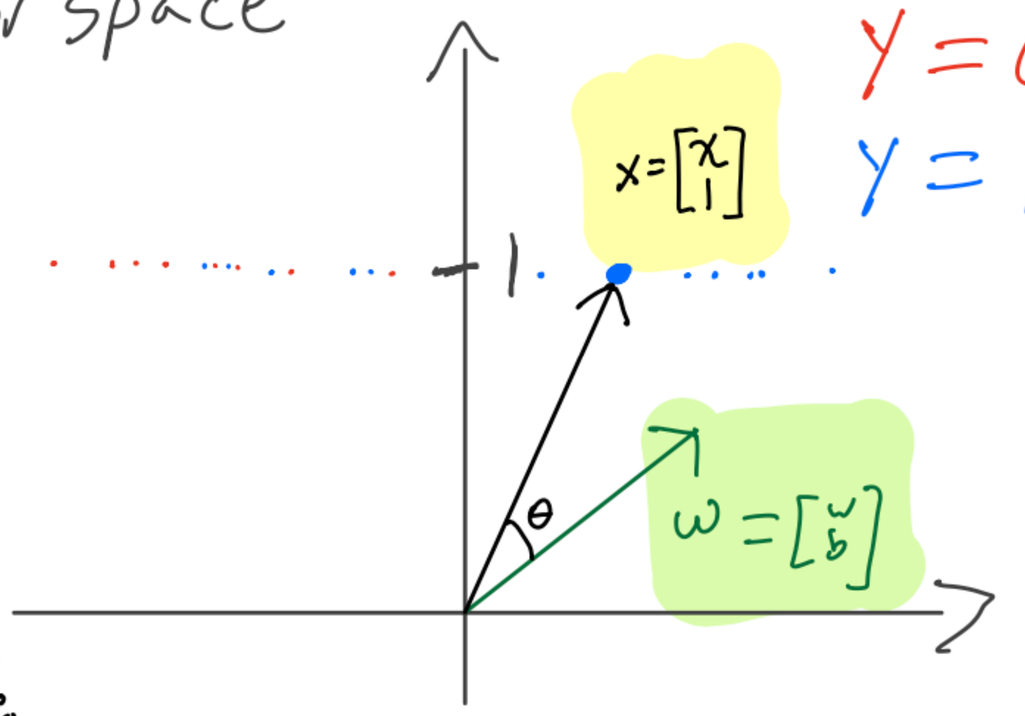
Vector:

$$f(x) = I(w^T x > 0)$$

$$= I\left(\begin{bmatrix} w \\ b \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} > 0\right)$$

Indicator: true → 1, false → 0

Vector space



How to read:

Input (x): ↔

Bias (1): ↕

output (y): color

Prediction:

predict y as $f(x)$

e.g. efficient? e.g. horsepower

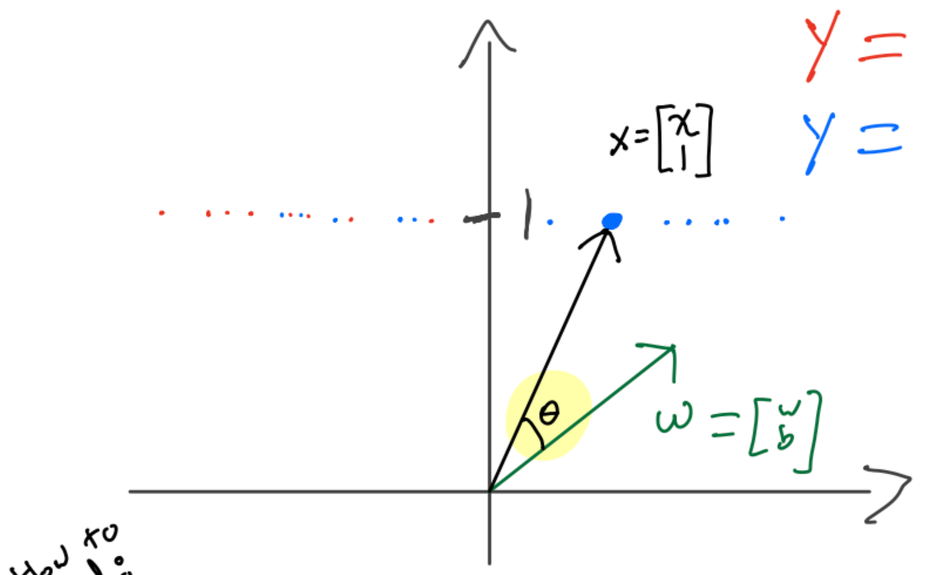
$$f(x) = I(wx + b > 0)$$

Vector:

$$f(x) = I(w^T x > 0)$$

$$= I\left(\begin{bmatrix} w \\ b \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} > 0\right)$$

Indicator: true → 1, false → 0



How to read:

Input (x): ↔

Bias (1): ↕

output (y): color

$$w^T x = \|w\| \, \|x\| \cos\theta$$

Angle

$\cos 90^\circ = 0$

$\cos\theta > 0$, $\theta \in (-90^\circ, 90^\circ)$

Prediction:

predict y as $f(x)$

e.g. efficient? (e.g. horsepower)

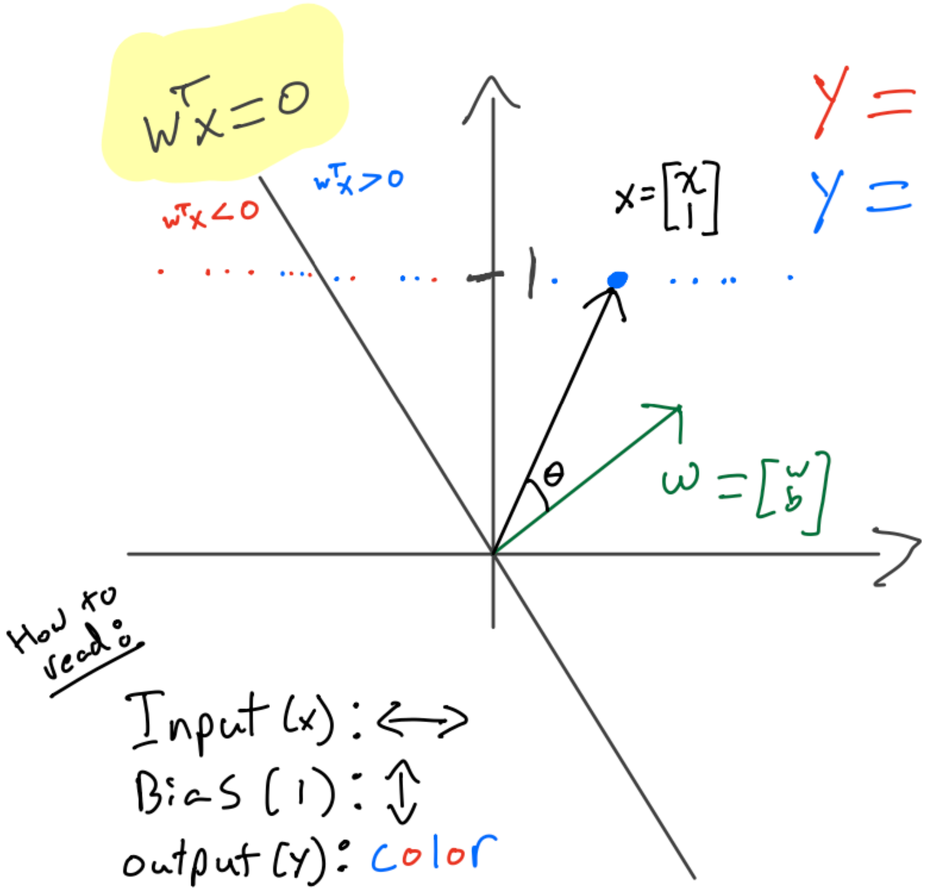
$$f(x) = I(wx + b > 0)$$

Vector:

$$f(x) = I(w^T x > 0)$$

$$= I\left(\begin{bmatrix} w \\ b \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} > 0\right)$$

Indicator: true → 1, false → 0



$$w^T x = \|w\| \, \|x\| \cos\theta$$

Angle

$\cos 90^\circ = 0$

$\cos\theta > 0$, $\theta \in (-90^\circ, 90^\circ)$

Prediction:

Predict y as $f(x)$

e.g. efficient? (e.g. horsepower)

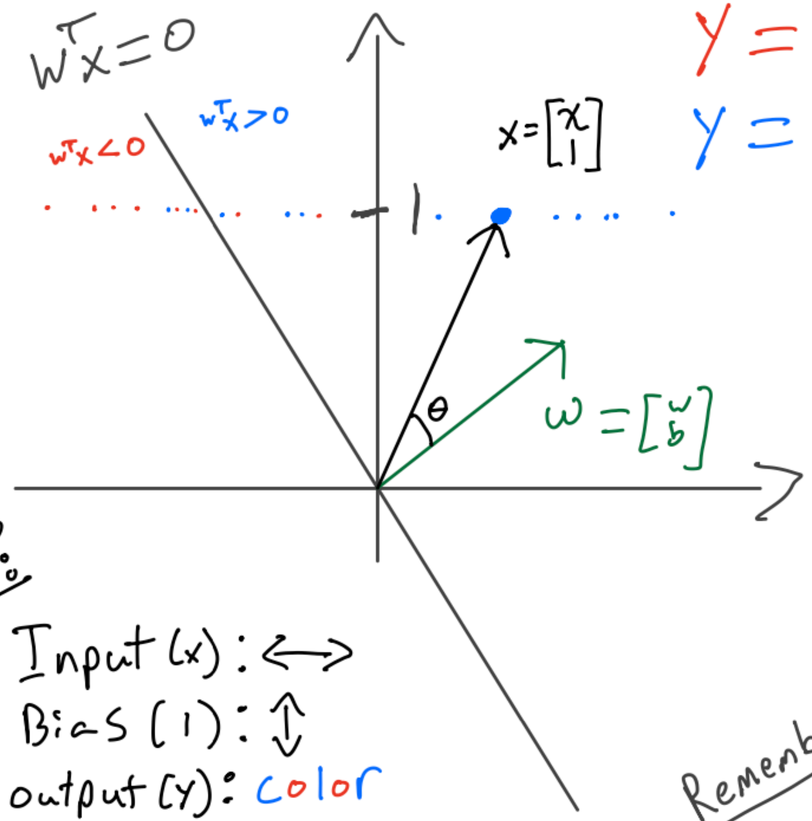
$$f(x) = I(wx + b > 0)$$

Vector:

$$f(x) = I(w^T x > 0)$$

$$= I\left(\begin{bmatrix} w \\ b \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} > 0\right)$$

Indicator: true → 1, false → 0



How to read:

Input (x): ↔

Bias (1): ↕

output (y): color

$y = 0$

$y = 1$

Magnitude: Larger $|w|$ → Larger $w^T x$ (always +!)

$$w^T x = \|w\| \, \|x\| \cos\theta$$

Angle

$\cos 90^\circ = 0$

$\cos\theta > 0$, $\theta \in (-90^\circ, 90^\circ)$

$$P(y=1|x) = \sigma(w^T x) = \sigma(\|w\| \, \|x\| \cos\theta)$$

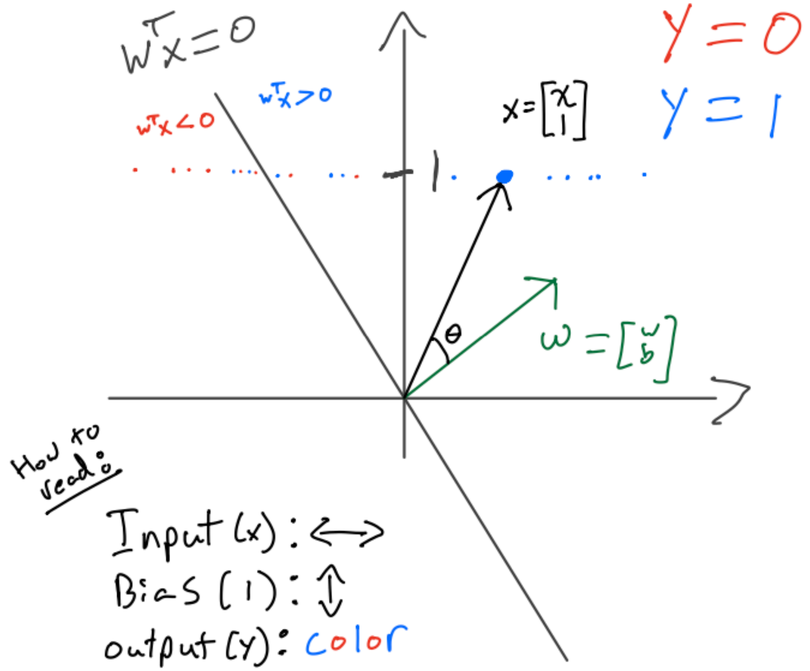
Remember

$\sigma(0) = 0.5$

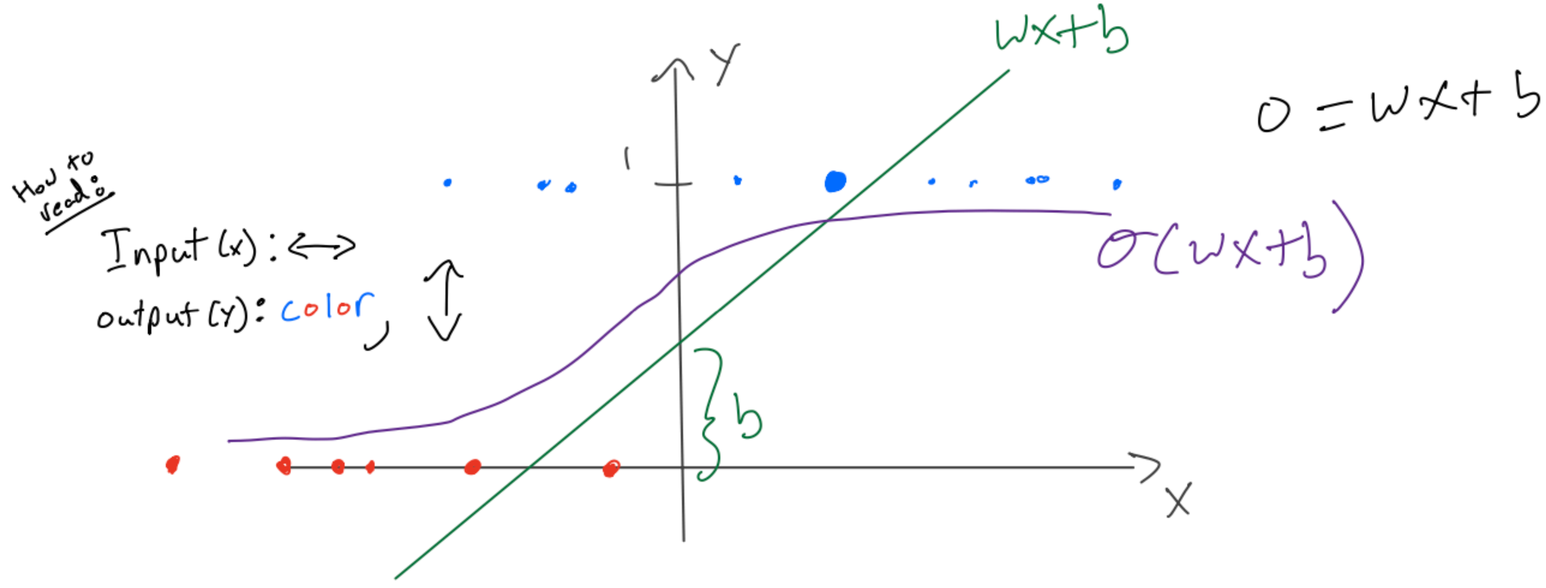
$\sigma(\infty) = 1$

$\sigma(-\infty) = 0$

Vector space



Input/output space



The Bernoulli distribution

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

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

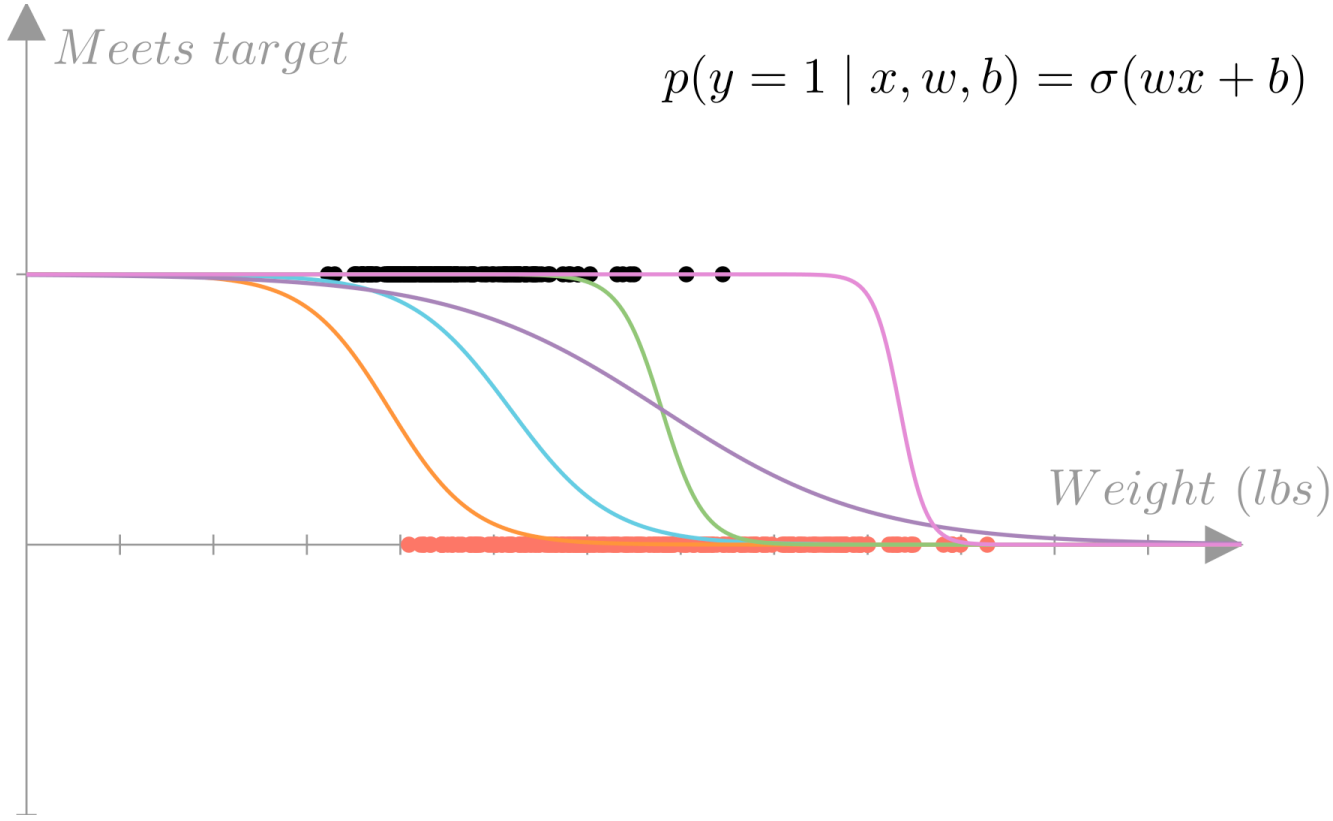
$y=1$

$y=0$

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

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

Maximum likelihood for logistic regression



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

Maximum likelihood for logistic regression

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

Independent $= \prod_{i=1}^{N} p(y_i \mid x_i, w)$

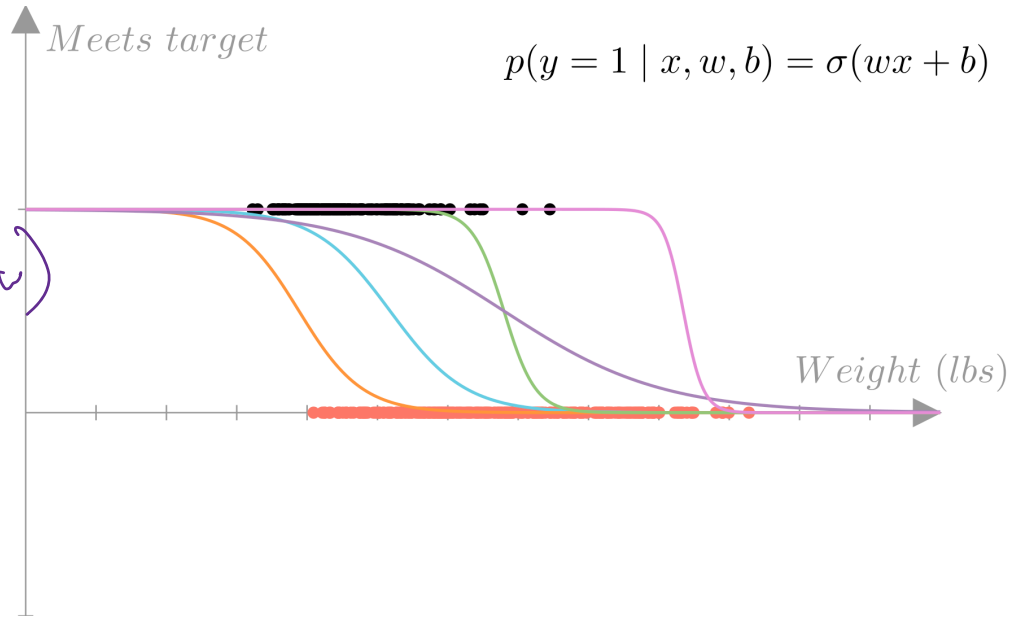
$$\mathbf{Loss}(\mathbf{w}) = \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

Negative log likelihood

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

$p(y=1 \mid x_i, w)$ $\quad$ $p(y=0 \mid x_i, w)$

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



Deriving the gradient

parameters $\mathbf{w}$ — data $\mathbf{X}, \mathbf{y}$ (inputs $\mathbf{x}_i$, outputs y_i)

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

$$\frac{d}{dw} - \sum_{i=1}^{N} \left[y_i \log \sigma(x_i^T w) + (1 - y_i) \log(1 - \sigma(x_i^T w)) \right]$$

$$= -\sum_{i=1}^{N} \frac{d}{dw} \left[y_i \log \sigma(x_i^T w) + (1 - y_i) \log(1 - \sigma(x_i^T w)) \right]$$

$$= -\sum_{i=1}^{N} \left[y_i \frac{d}{dw} \log \sigma(x_i^T w) + (1 - y_i) \frac{d}{dw} \log(1 - \sigma(x_i^T w)) \right]$$

$$= -\sum_{i=1}^{N} \left[y_i \frac{1}{\sigma(x_i^T w)} \frac{d}{dw} \sigma(x_i^T w) + \frac{(1 - y_i)}{1 - \sigma(x_i^T w)} \frac{d}{dw} (1 - \sigma(x_i^T w)) \right]$$

$$\frac{1}{\sigma(x_i^T w)} \sigma(x_i^T w)(1-\sigma(x_i^T w)) \frac{d}{dw} x_i^T w$$

$$= (1-\sigma(x_i^T w)) x_i$$

$$= \sigma(-x_i^T w) x_i$$

$x_1 w_1 + x_2 w_2 + x_3 w_3 \cdots$

$$\begin{bmatrix} \frac{\partial}{\partial w_1} \\ ? \\ \frac{\partial}{\partial w_2} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Maximum likelihood for logistic regression

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

Independent $= \prod_{i=1}^{N} p(y_i \mid x_i, w)$

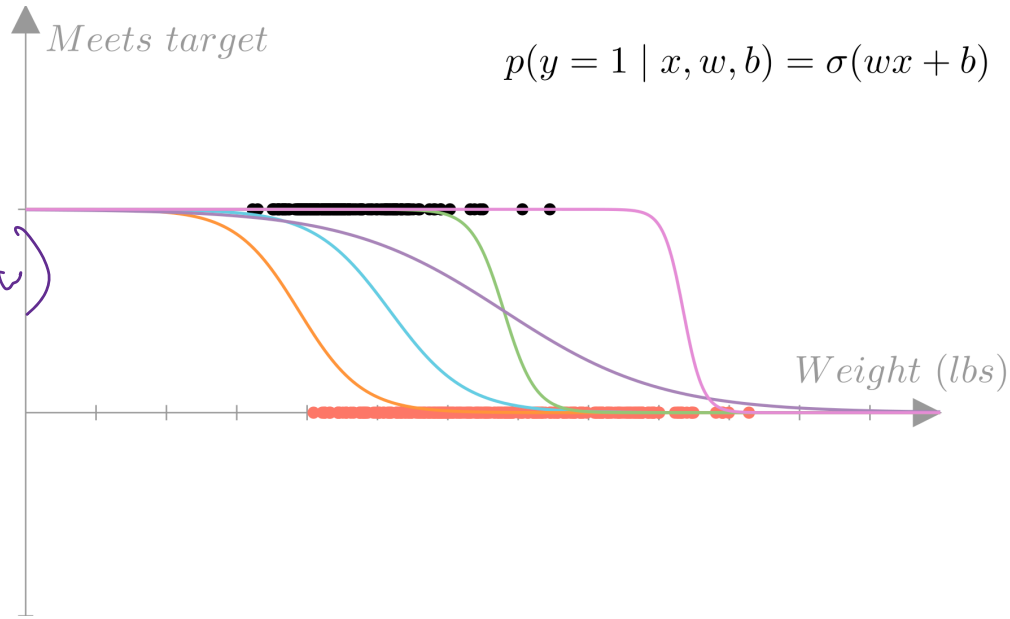
$$\mathbf{Loss}(\mathbf{w}) = \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

Negative log likelihood

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

$p(y=1 \mid x_i, w)$ $\quad$ $p(y=0 \mid x_i, w)$

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



Multi-class classification

f() ⟶ {Cat, Dog, Mouse,…}

$$y = f(\mathbf{x}), \quad \text{Input: } \mathbf{x} \in \mathbb{R}^n \longrightarrow \text{Output: } y \in \{1, 2, \ldots, C\}$$

Ordering irrelevant!

1: Cat, 2: Dog, 3: Mouse

1: Dog, 2: Mouse, 3: Cat

Multi-class prediction functions

Binary thresholding

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

Multi-class thresholding

$$f(\mathbf{x}) = \underset{c \in \{1 \ldots C\}}{\operatorname{argmax}} \mathbf{x}^T\mathbf{w}_c$$

Multi-class:

Prediction:

predict y as $f(x)$

e.g. Brand ?. (e.g. horsepower)

$$f(x) = \arg\max_c w_c x + b_c$$

Vector:

$$f(x) = \arg\max_c w_c^T x$$

$$= \arg\max_c \begin{bmatrix} w_c \\ b_c \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix}$$

Multi-class:

Prediction:

predict y as $f(x)$

e.g. Brand ? e.g. horsepower

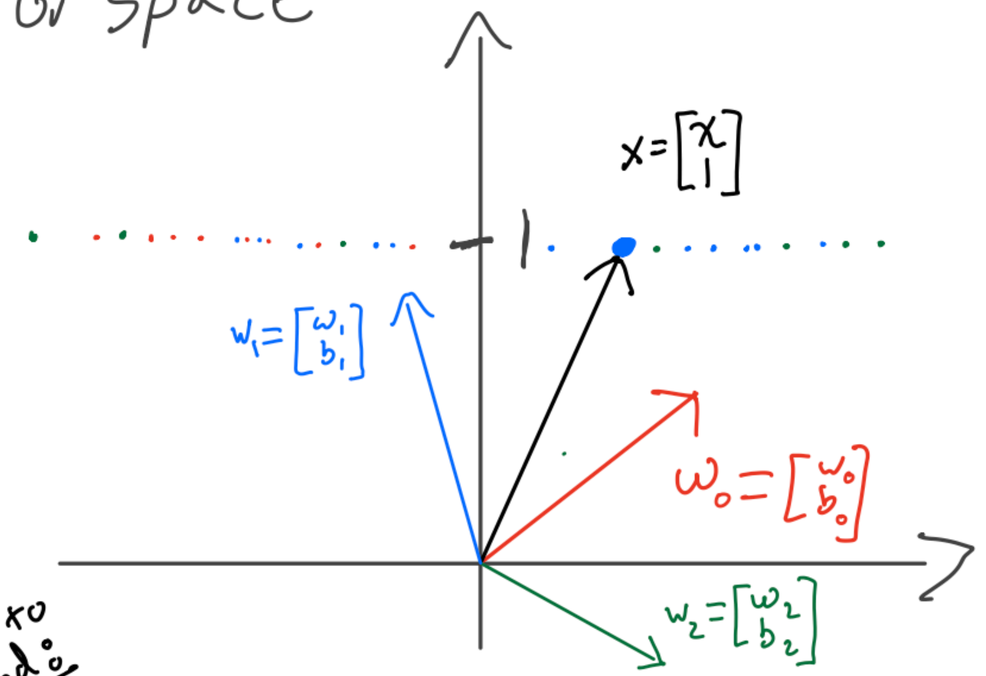
$$f(x) = \arg\max_c w_c x + b_c$$

Vector:

$$f(x) = \arg\max_c w_c^T x$$

$$= \arg\max_c \begin{bmatrix} w_c \\ b_c \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix}$$

Vector space



y = 0
y = 1
y = 2

How to read:

Input (x): ⟷
Bias (1): ↕
output (y): color

Multi-class!

Prediction:

Predict y as $f(x)$

e.g. Brand ? e.g. horsepower

$$f(x) = \arg\max_c w_c x + b_c$$

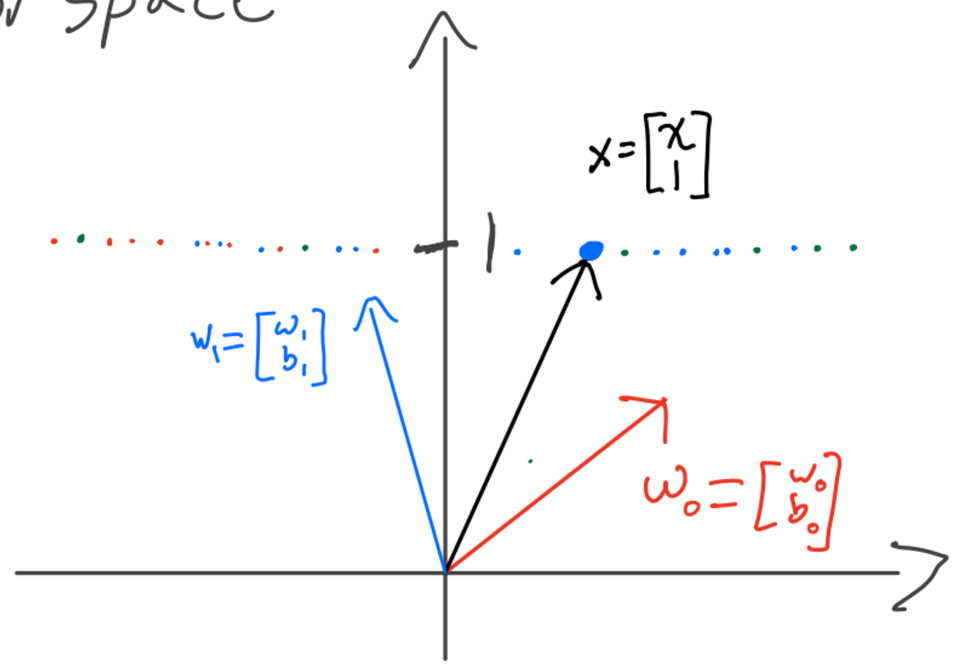
Vector:

$$f(x) = \arg\max_c w_c^T x$$

$$= \arg\max_c \begin{bmatrix} w_c \\ b_c \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix}$$

2-class: $w_1^T x > w_0^T x$

Vector space



$y = 0$

$y = 1$

Multi-class:

Prediction:

Predict y as $f(x)$

e.g. Brand ? e.g. horsepower

$$f(x) = \arg\max_c w_c x + b_c$$

Vector:

$$f(x) = \arg\max_c w_c^T x$$

$$= \arg\max_c \begin{bmatrix} w_c \\ b_c \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix}$$

2-class:

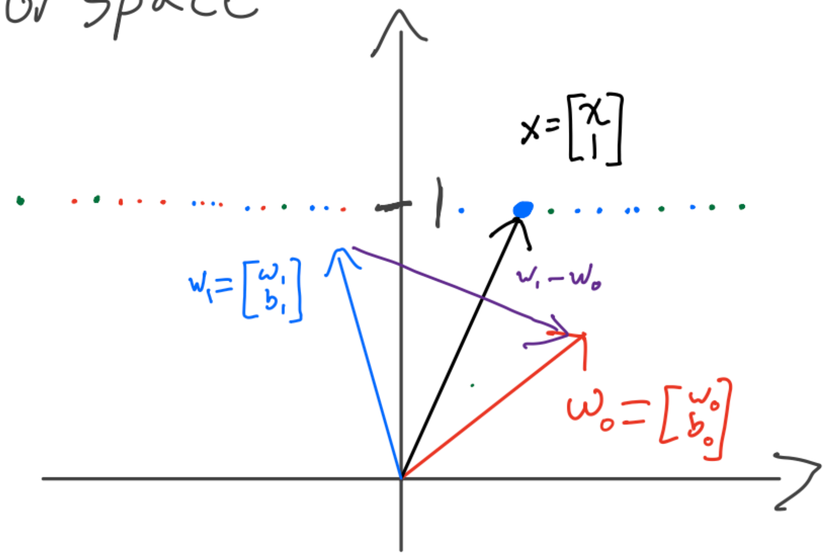
$$w_1^T x > w_0^T x$$

$$0 > w_1^T x - w_0^T x$$

$$0 > (w_1 - w_0)^T x$$

w

Vector space



$y = 0$

$y = 1$

Multi-class:

Prediction:

Predict y as $f(x)$

e.g. Brand ? e.g. horsepower

$$f(x) = \arg\max_c w_c x + b_c$$

Vector:

$$f(x) = \arg\max_c w_c^T x$$

$$= \arg\max_c \begin{bmatrix} w_c \\ b_c \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix}$$

2-class:

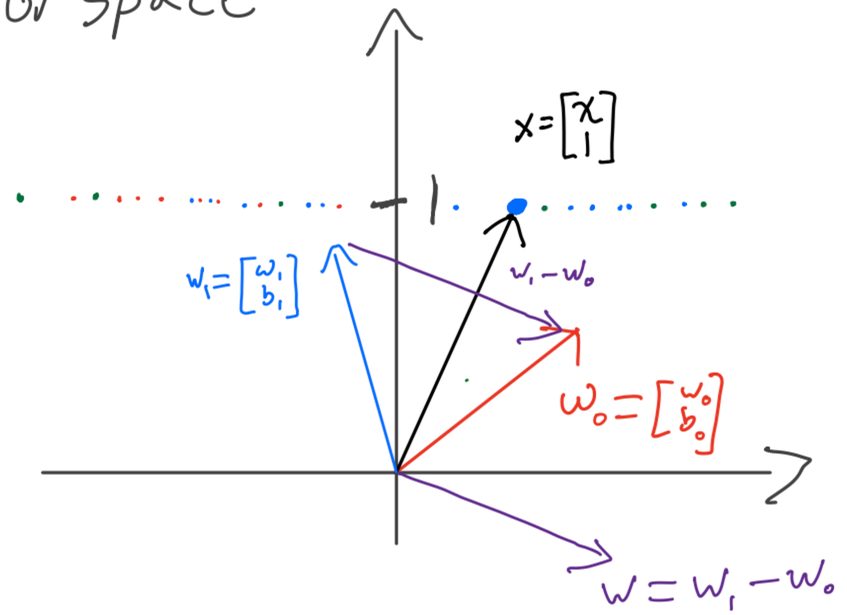
$$w_1^T x > w_0^T x$$

$$0 > w_1^T x - w_0^T x$$

$$0 > (w_1 - w_0)^T x$$

$$w$$

Vector space



y=0

y=1

Multi-class:

Prediction:

predict y as $f(x)$

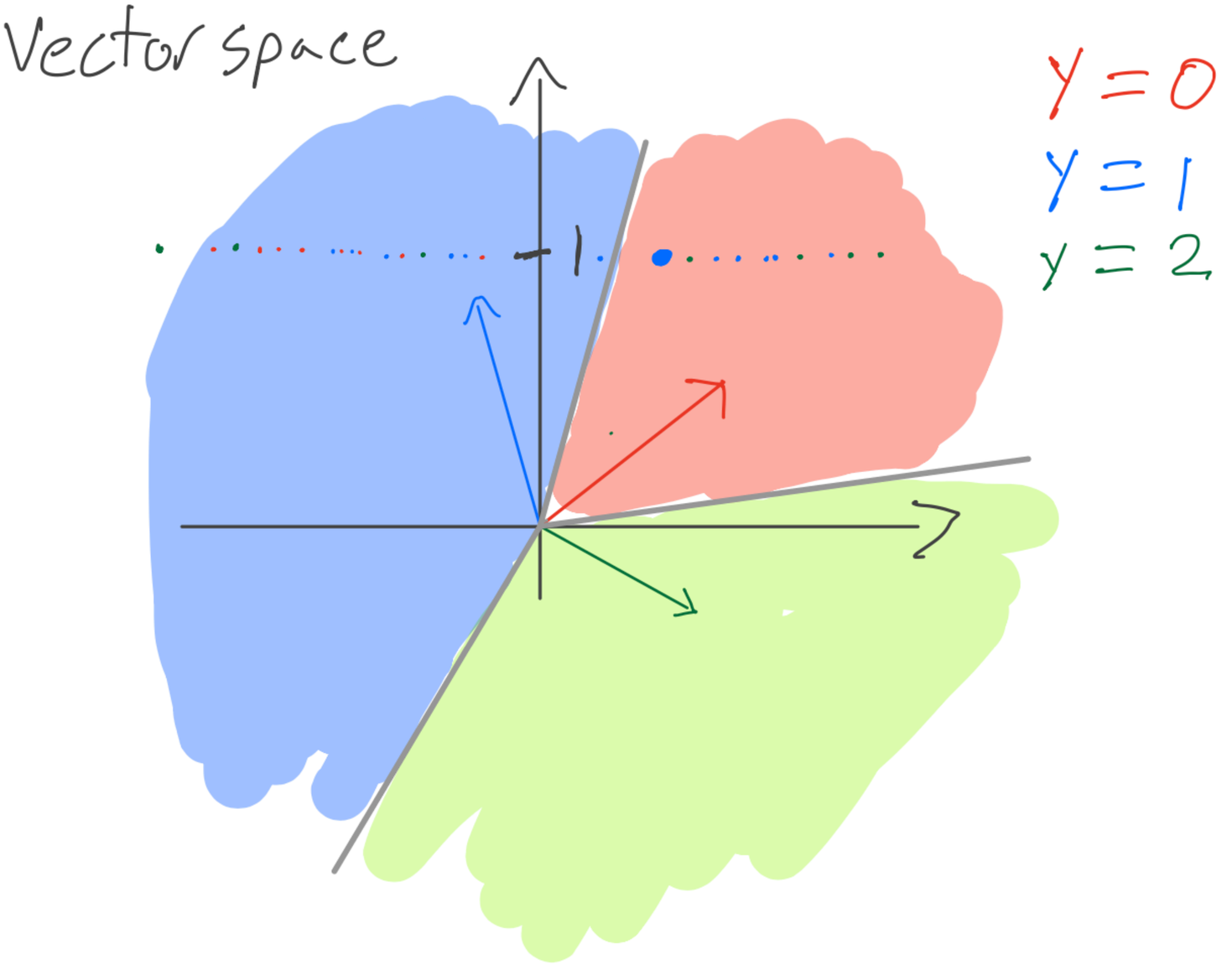
e.g. Brand ? e.g. horsepower

$$f(x) = \arg\max_c w_c x + b_c$$

Vector:

$$f(x) = \arg\max_c w_c^T x$$

$$= \arg\max_c \begin{bmatrix} w_c \\ b_c \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix}$$



Multi-class prediction functions

Multi-class thresholding

$$f(\mathbf{x}) = \underset{c \in \{1 \ldots C\}}{\operatorname{argmax}} \mathbf{x}^T \mathbf{w}_c$$

For convenience, we can also define a matrix that contains all C parameter vectors:

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_C^T \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} & \ldots & W_{1d} \\ W_{21} & W_{22} & \ldots & W_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ W_{C1} & W_{C2} & \ldots & W_{Cd} \end{bmatrix}$$

With this notation, our prediction function becomes:

$$f(\mathbf{x}) = \underset{c \in \{1 \ldots C\}}{\operatorname{argmax}} \left(\mathbf{x}^T \mathbf{W}^T\right)_c, \quad \mathbf{W} \in \mathbb{R}^{C \times d}$$

Categorical distribution

$$p(y = c) = q_c, \quad y \in \{1 \dots C\}$$

$$\mathbf{q} \in \mathbb{R}^C \quad q_c \geq 0 \; \forall c \in \{1 \dots C\} \quad \sum_{c=1}^{C} q_c = 1$$

A probabilistic model for multi-class classification

$$y_i \sim \mathbf{Categorical}(\mathbf{q} = ?), \quad \mathbf{q} = \mathbf{x}_i^T \mathbf{W}^T ?$$

$$\mathbf{x}^T \mathbf{W}^T \in \mathbb{R}^C, \quad q_c \ngeq 0 \; \forall c \in \{1 \dots C\}, \quad \sum_{c=1}^{C} q_c \neq 1$$

$$\textbf{Need} \;\; f(\mathbf{x}) : \; \mathbb{R}^C \longrightarrow [0, \infty)^C, \; \sum_{i=1}^{C} f(\mathbf{x})_c = 1$$

Categorical distribution

$$p(y = c) = q_c, \quad y \in \{1 \dots C\}$$

$$\mathbf{q} \in \mathbb{R}^C \quad q_c \geq 0 \; \forall c \in \{1 \dots C\} \quad \sum_{c=1}^{C} q_c = 1$$

$$p(y) = \prod q_c^{\mathbb{I}(y=c)} = q_y$$

$$\log p(y) =$$

Softmax function

$$\mathbf{softmax}(\mathbf{x})_c = \frac{e^{x_c}}{\sum_{j=1}^{C} e^{x_j}}$$

$$\mathbf{softmax}(\mathbf{x}) = \begin{bmatrix} \frac{e^{x_1}}{\sum_{j=1}^{C} e^{x_j}} \\ \frac{e^{x_2}}{\sum_{j=1}^{C} e^{x_j}} \\ \vdots \\ \frac{e^{x_C}}{\sum_{j=1}^{C} e^{x_j}} \end{bmatrix}$$

Softmax function

$$\mathbf{softmax}(\mathbf{x})_c = \frac{e^{x_c}}{\sum_{j=1}^{C} e^{x_j}}$$

$$\sum_{i=1}^{C} \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} = \frac{\sum_{i=1}^{C} e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} = 1$$

$$\underset{c \in \{1,\ldots,C\}}{\mathbf{argmax}}\ \mathbf{x}_c = \underset{c \in \{1,\ldots,C\}}{\mathbf{argmax}}\ \mathbf{softmax}(\mathbf{x})_c$$

A probabilistic model for multi-class classification: Multinomial Logistic regression

$$y_i \sim \mathbf{Categorical}\big(\mathrm{softmax}(\mathbf{x}^T\mathbf{W})\big)$$

$$p(y_i = c \mid \mathbf{x}, \mathbf{W}) = \mathrm{softmax}(\mathbf{x}^T\mathbf{W})_c = \frac{e^{\mathbf{x}^T\mathbf{w}_c}}{\sum_{j=1}^{C} e^{\mathbf{x}^T\mathbf{w}_j}}$$

Maximum likelihood estimation for Multinomial Logistic regression

Model

$$y_i \sim \textbf{Categorical}\big(\text{softmax}(\mathbf{x}^T\mathbf{W})\big)$$

$$p(y_i = c \mid \mathbf{x}, \mathbf{W}) = \text{softmax}(\mathbf{x}^T\mathbf{W})_c = \frac{e^{\mathbf{x}^T\mathbf{w}_c}}{\sum_{j=1}^{C} e^{\mathbf{x}^T\mathbf{w}_j}}$$

Loss

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

$$\textbf{NLL}(\mathbf{W}, \mathbf{X}, \mathbf{y}) =$$

Maximum likelihood estimation for Multinomial Logistic regression

Model

$$y_i \sim \textbf{Categorical}\big(\text{softmax}(\mathbf{x}^T\mathbf{W})\big)$$

$$p(y_i = c \mid \mathbf{x}, \mathbf{W}) = \text{softmax}(\mathbf{x}^T\mathbf{W})_c = \frac{e^{\mathbf{x}^T\mathbf{w}_c}}{\sum_{j=1}^{C} e^{\mathbf{x}^T\mathbf{w}_j}}$$

Loss

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

$$\textbf{NLL}(\mathbf{W}, \mathbf{X}, \mathbf{y}) =$$

Models so far

We saw that a reasonable model for continuous outputs ($y \in \mathbb{R}$) is **linear regression**.

$$\textbf{Predict } y \in \textbf{ as } \begin{cases} y = \mathbf{x}^T\mathbf{w} & \text{(prediction function)} \\ p(y \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(y \mid \mathbf{x}^T\mathbf{w}, \sigma^2) & \text{(probabilistic view)} \end{cases}$$

A reasonable model for *binary* outputs ($y \in \{0, 1\}$) is **logistic regression**:

$$\textbf{Predict } y \in \textbf{ as } \begin{cases} y = \mathbb{I}(\mathbf{x}^T\mathbf{w} > 0) & \text{(prediction function)} \\ p(y = 1 \mid \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}^T\mathbf{w}) & \text{(probabilistic view)} \end{cases}$$

A reasonable model for *categorical* outputs ($y \in \{0, 1, \ldots, C\}$) is **multinomial logistic regression**:

$$\textbf{Predict } y \in \textbf{ as } \begin{cases} y = \underset{c}{\operatorname{argmax}}\ \mathbf{x}^T\mathbf{w}_c & \text{(prediction function)} \\ p(y = c \mid \mathbf{x}, \mathbf{w}) = \operatorname{softmax}(\mathbf{x}^T\mathbf{W})_c & \text{(probabilistic view)} \end{cases}$$