

Binary Classification

$f($  $) \rightarrow \{\text{Cat, Dog}\}$

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

Linear Log.

\mathbb{R}

Prediction:

Predict y as $f(x)$

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

$$f(x) = I(\omega x + b > 0) \quad w_1x_1 + w_2x_2 + \dots$$

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 $\rightarrow 1$, false $\rightarrow 0$

Prediction:

Predict y as $f(x)$
 e.g. efficient? (e.g. horsepower)

$$f(x) = I(\omega x + b > 0)$$

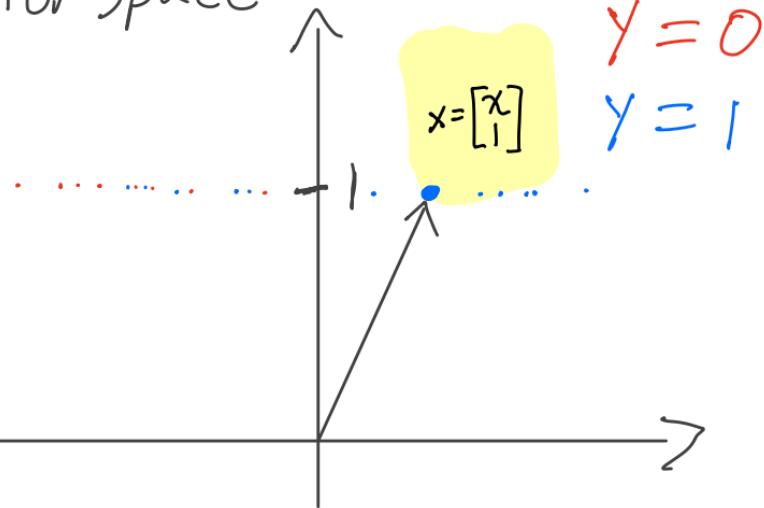
Vector:

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

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

Indicator: true $\rightarrow 1$, false $\rightarrow 0$

Vector Space



How to
read it

Input (x): \leftrightarrow

Bias (1): \uparrow

Output (y): color

Prediction:

Predict y as $f(x)$
 e.g. efficient? e.g. horsepower

$$f(x) = I(\omega x + b > 0)$$

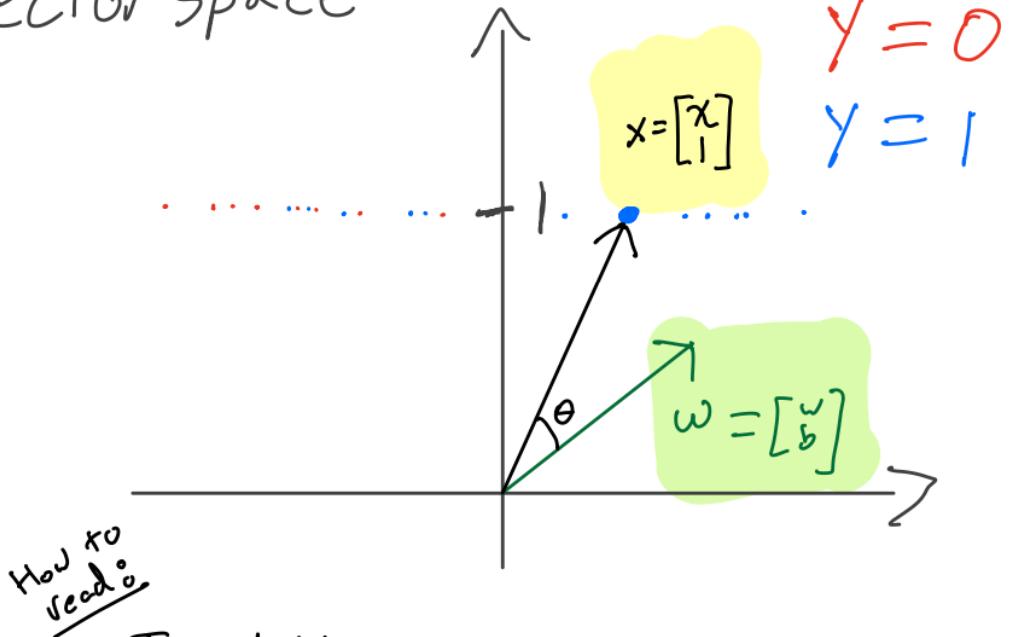
Vector:

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

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

Indicator: true $\rightarrow 1$, false $\rightarrow 0$

Vector Space



Input (x): \leftrightarrow

Bias (1): \uparrow

Output (y): color

Prediction:

Predict y as $f(x)$
 e.g. efficient? (e.g. horsepower)

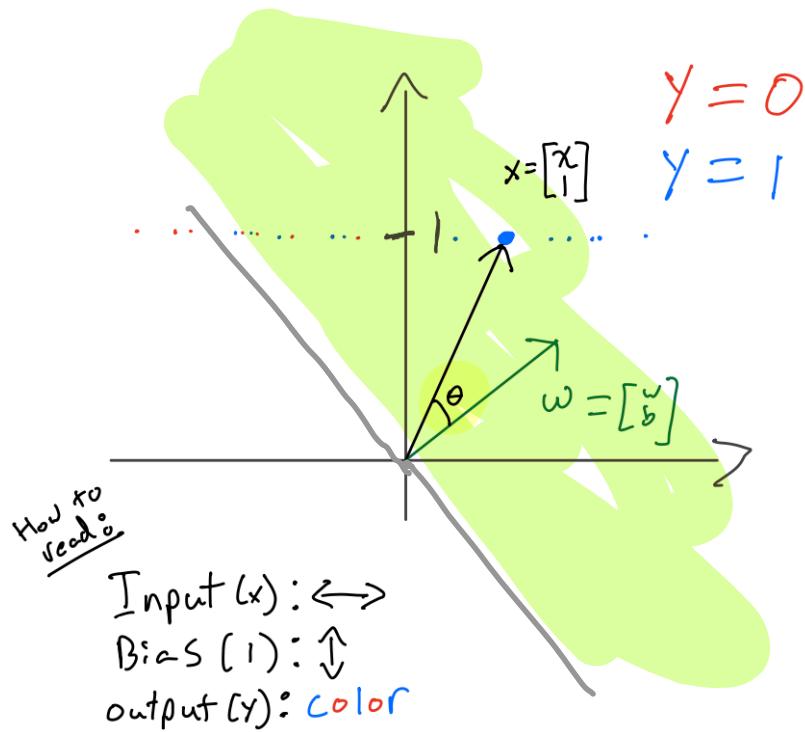
$$f(x) = I(\omega x + b > 0)$$

Vector:

$$f(x) = I(\omega^T x > 0)$$

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

Indicator: true $\rightarrow 1$, false $\rightarrow 0$



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

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

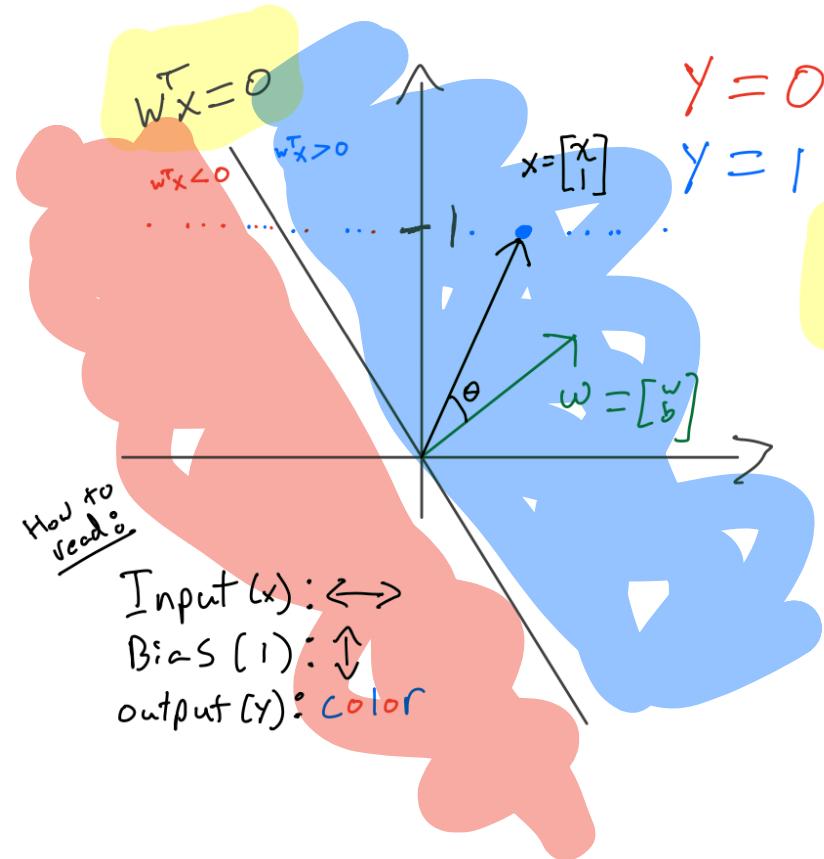
$$f(x) = I(w^T x + b > 0)$$

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 $\rightarrow 1$, false $\rightarrow 0$



Prediction:

Predict y as $f(x)$
e.g. efficient? (e.g. horsepower)

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

Vector:

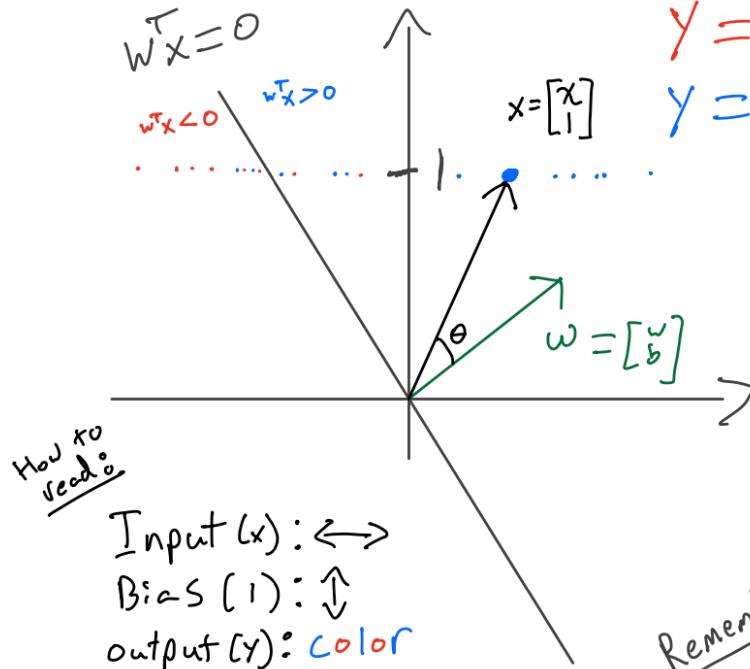
$$f(x) = I(w^T x > 0)$$

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

Indicator: true $\rightarrow 1$, false $\rightarrow 0$

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

$$P(y=1|x) = 0.5$$



$y = 0$ Magnitude: Larger $|w|$ \rightarrow
 $y = 1$ Larger $w^T x$
(always +!)

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

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

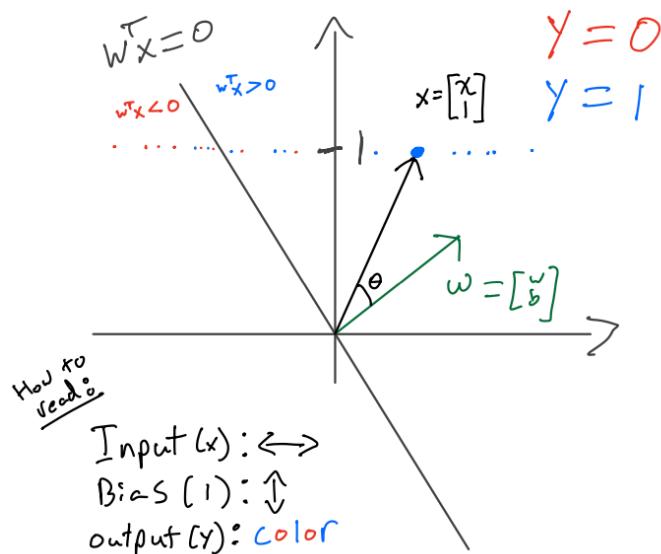
Remember

$$\sigma(0) = 0.5$$

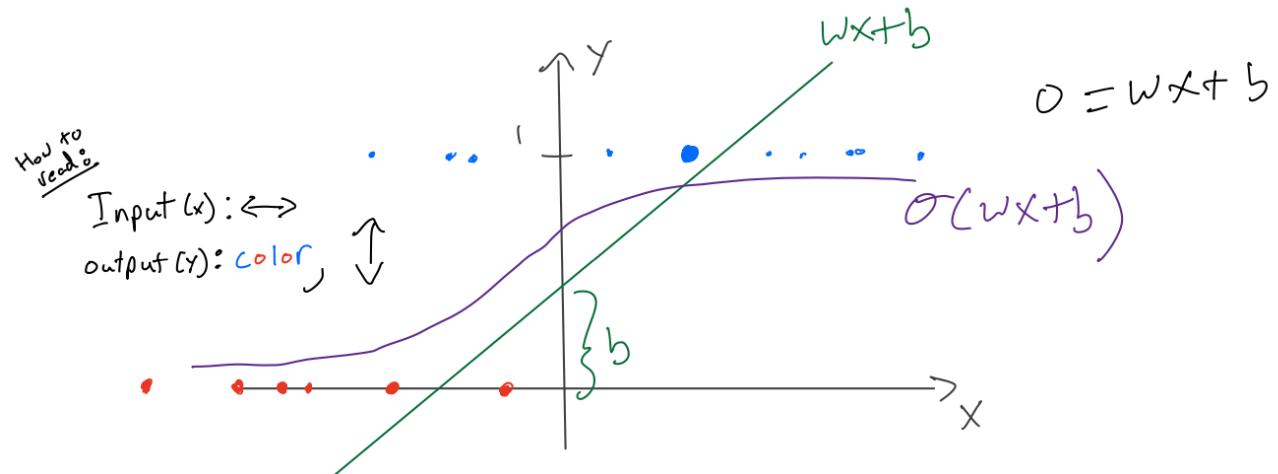
$$\sigma(\infty) = 1$$

$$\sigma(-\infty) = 0$$

Vector space



Input/output Space



The Bernoulli distribution

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

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

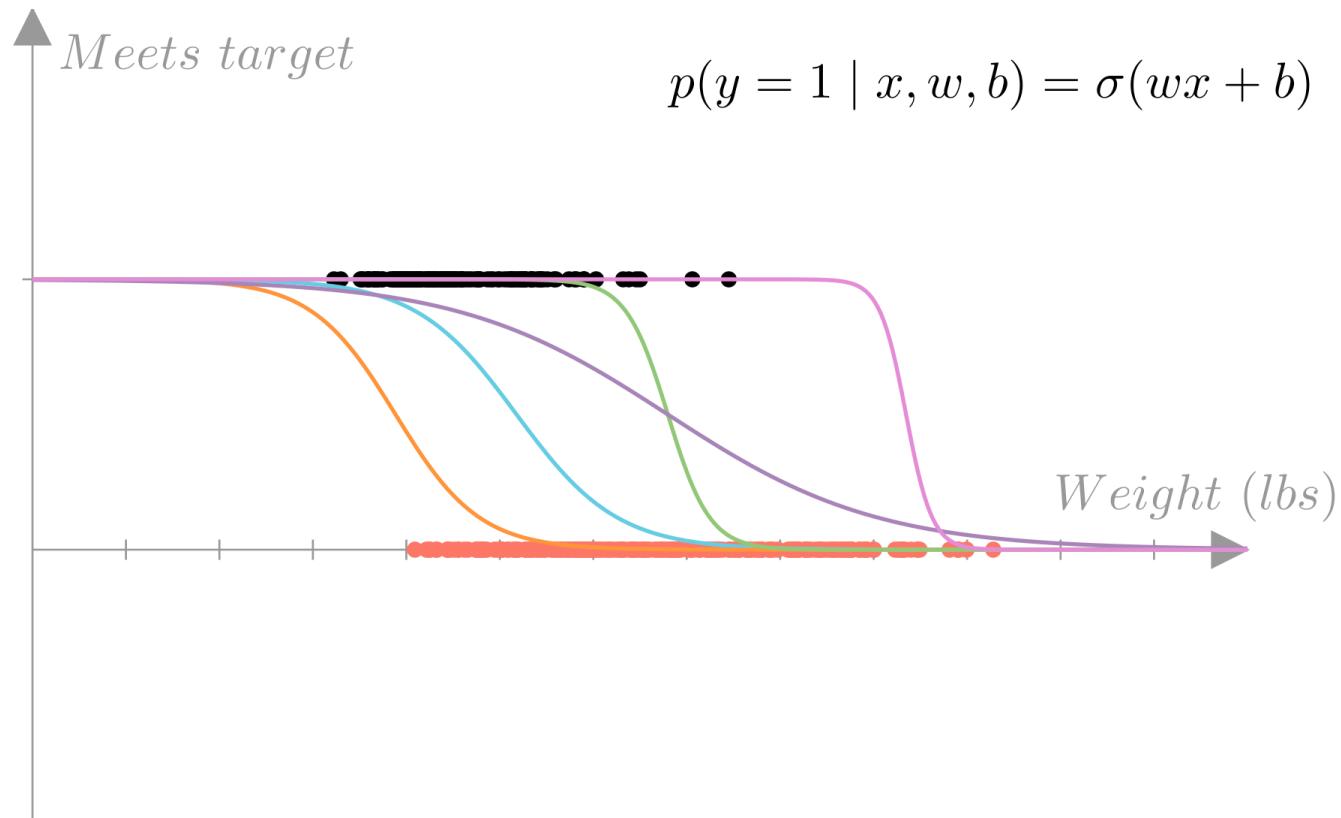
$$p(y) = q^y (1 - q)^{1-y}$$

$y=1$
 $y=0$

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

$\brace{q \in (0, 1)}$

Maximum likelihood for logistic regression



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

Maximum likelihood for logistic regression

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} p(y_1, \dots, y_N \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{w}) \quad (1)$$

Independent $\sim \prod_{i=1}^N p(y_i \mid \mathbf{x}_i, \mathbf{w})$

2)

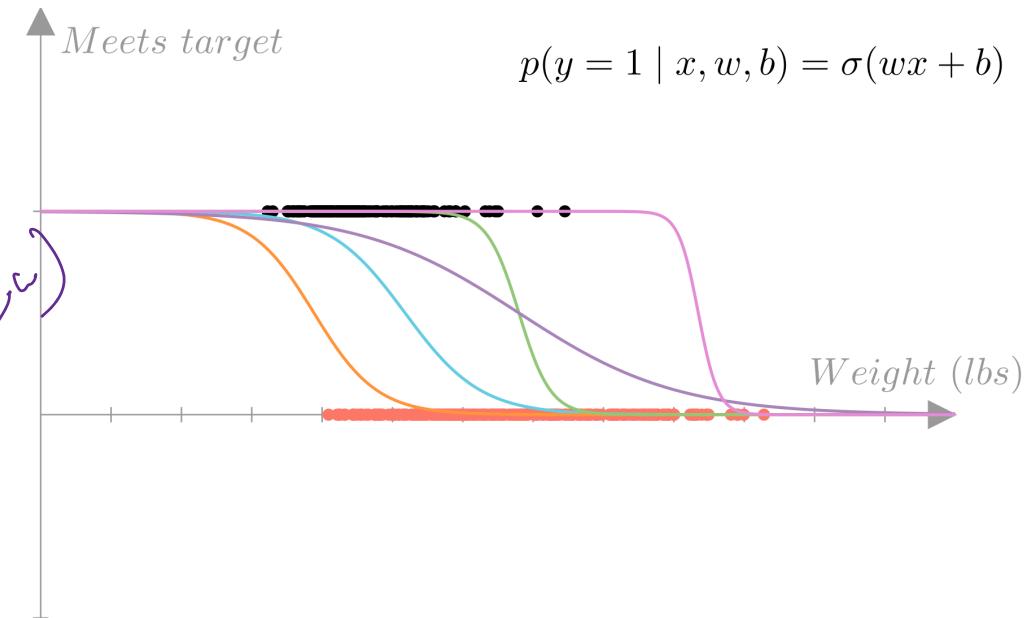
$$\text{Loss}(\mathbf{w}) = \text{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^N \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

Negative log likelihood

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

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

$p(y_i = 1 \mid \mathbf{x}_i, \mathbf{w})$ *$p(y_i = 0 \mid \mathbf{x}_i, \mathbf{w})$*



Deriving the gradient

parameters ↗
data ↗

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

inputs outputs

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

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

$$= - \sum_{i=1}^N \left[y_i \underbrace{\frac{d}{d \mathbf{w}} \log \sigma(\mathbf{x}_i^T \mathbf{w})}_{\text{green box}} + (1 - y_i) \underbrace{\frac{d}{d \mathbf{w}} \log(1 - \sigma(\mathbf{x}_i^T \mathbf{w}))}_{\text{red box}} \right]$$

$$= - \sum_{i=1}^N \left[y_i \underbrace{\frac{1}{\sigma(\mathbf{x}_i^T \mathbf{w})} \frac{d}{d \mathbf{w}} \sigma(\mathbf{x}_i^T \mathbf{w})}_{\text{purple box}} + \frac{(1 - y_i)}{1 - \sigma(\mathbf{x}_i^T \mathbf{w})} \underbrace{\frac{1}{\sigma(\mathbf{x}_i^T \mathbf{w})} \frac{d}{d \mathbf{w}} \log(1 - \sigma(\mathbf{x}_i^T \mathbf{w}))}_{\text{purple box}} \right]$$

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

$$= \boxed{(\sigma(x_i^T w)) x_i}$$

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

$x_1 w_1 + \underbrace{x_2 w_2}_{} + x_3 w_3 \dots$

$$\begin{bmatrix} \frac{\partial v_1}{\partial w_1} \\ \vdots \\ \frac{\partial v_n}{\partial w_1} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\frac{d}{dw} \log \sigma(x_i^T w) = (\sigma(x_i^T w)) x_i$$

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

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

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

d

Maximum likelihood for logistic regression

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} p(y_1, \dots, y_N \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{w}) \quad (1)$$

Independent $\sim \prod_{i=1}^N p(y_i \mid \mathbf{x}_i, \mathbf{w})$

2)

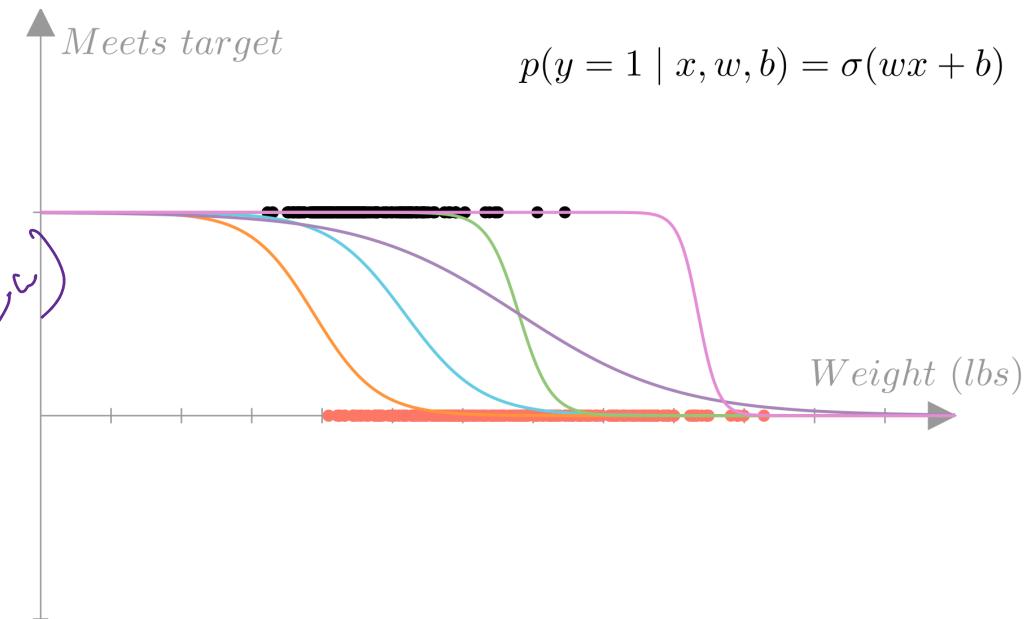
$$\text{Loss}(\mathbf{w}) = \text{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^N \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

Negative log likelihood

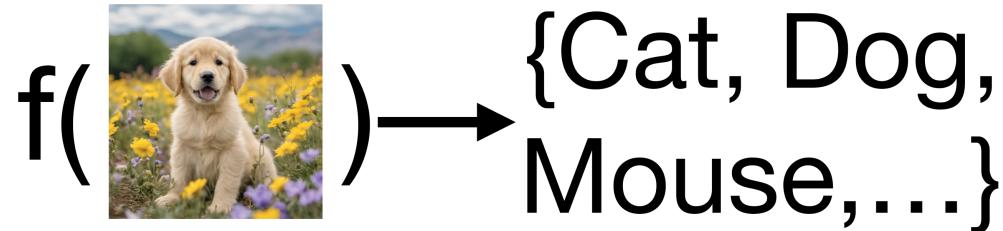
$$p(y_i = 1 \mid \mathbf{x}_i, \mathbf{w}) = \sigma(\mathbf{x}_i^T \mathbf{w}), \quad p(y_i = 0 \mid \mathbf{x}_i, \mathbf{w}) = 1 - \sigma(\mathbf{x}_i^T \mathbf{w}) = \sigma(-\mathbf{x}_i^T \mathbf{w})$$

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

$p(y_i = 1 \mid \mathbf{x}_i, \mathbf{w})$ *$p(y_i = 0 \mid \mathbf{x}_i, \mathbf{w})$*



Multi-class classification



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

L_{classes}

Ordering irrelevant!

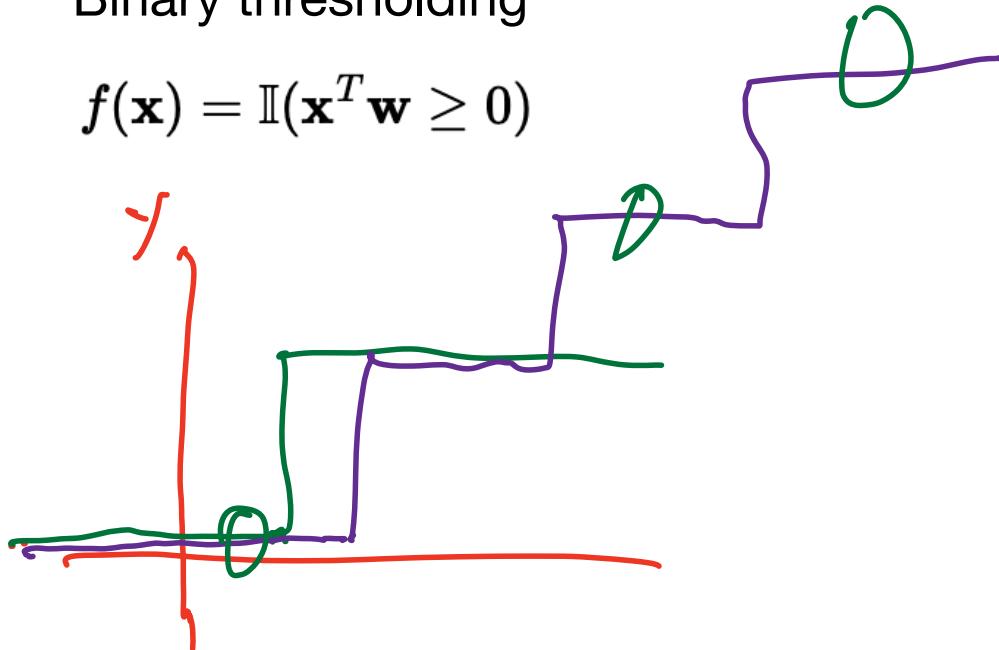
1: Cat, 2: Dog, 3: Mouse

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

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

Linear func.
 w_c for
each class

Multi-class:

Prediction:

Predict y as $f(x)$
e.g. Brand? e.g. horsepower

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

$$\begin{cases} c_1 : (w_1, b_1) \\ c_2 : (w_2, b_2) \\ c_3 : (w_3, b_3) \end{cases}$$

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

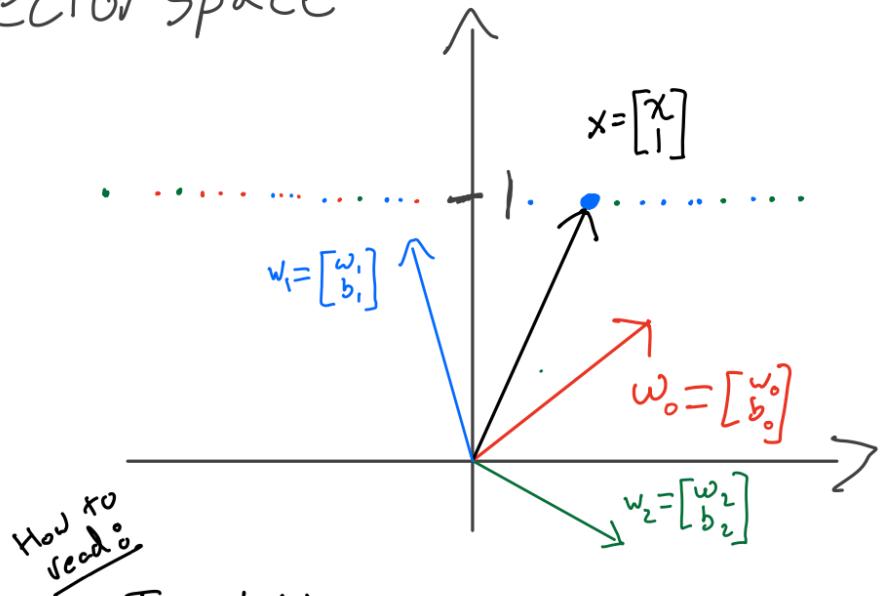
$$f(x) = \arg \max_c w_c x + b_c$$

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



$y = 0$
 $y = 1$
 $y = 2$

Input (x): \leftrightarrow
Bias (1): \uparrow
Output (y): color

Multi-class!

Prediction:

Predict y as $f(x)$
e.g. Brand? (e.g. horsepower)

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

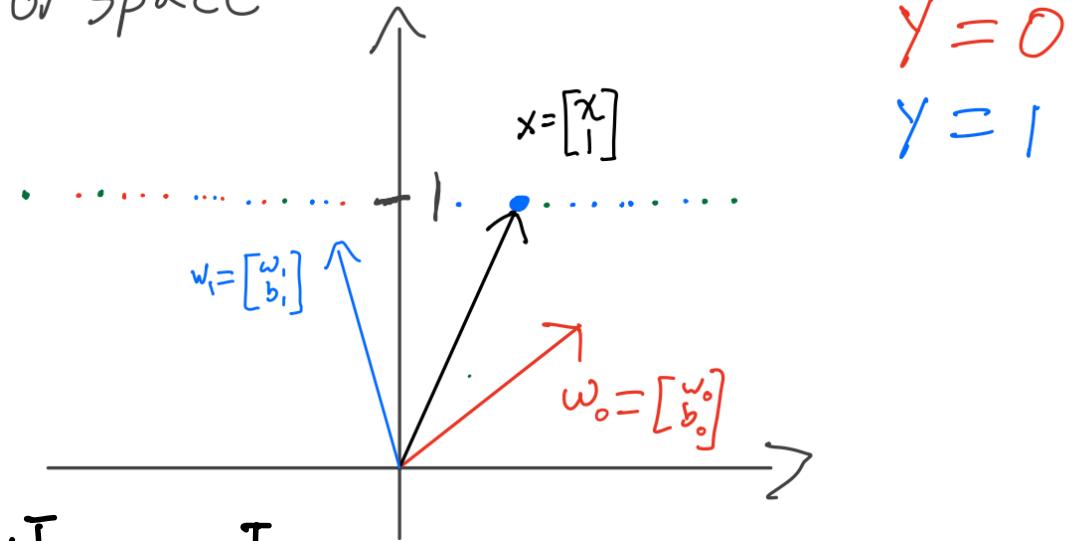
Vector:

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

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

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

Vector Space



$$\begin{aligned} y &= 0 \\ y &= 1 \end{aligned}$$

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:

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

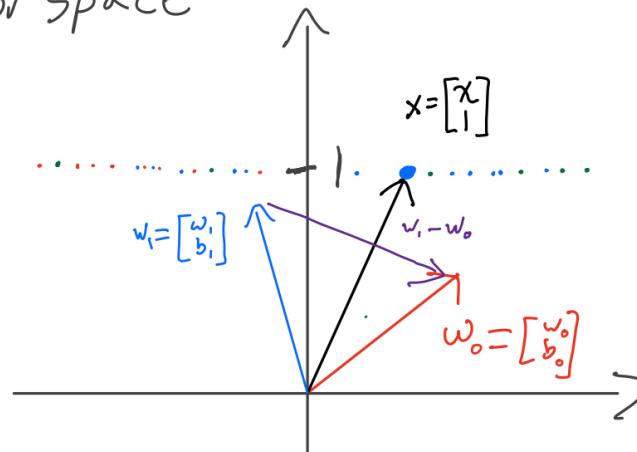
$$= \arg \max_c \begin{bmatrix} w_c \\ b_c \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix}$$

$$2\text{-class: } w_i^T x > w_o^T x$$

$$0 > w_i^T x - w_o^T x$$

$$0 > (w_i - w_o)^T x$$

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:

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

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

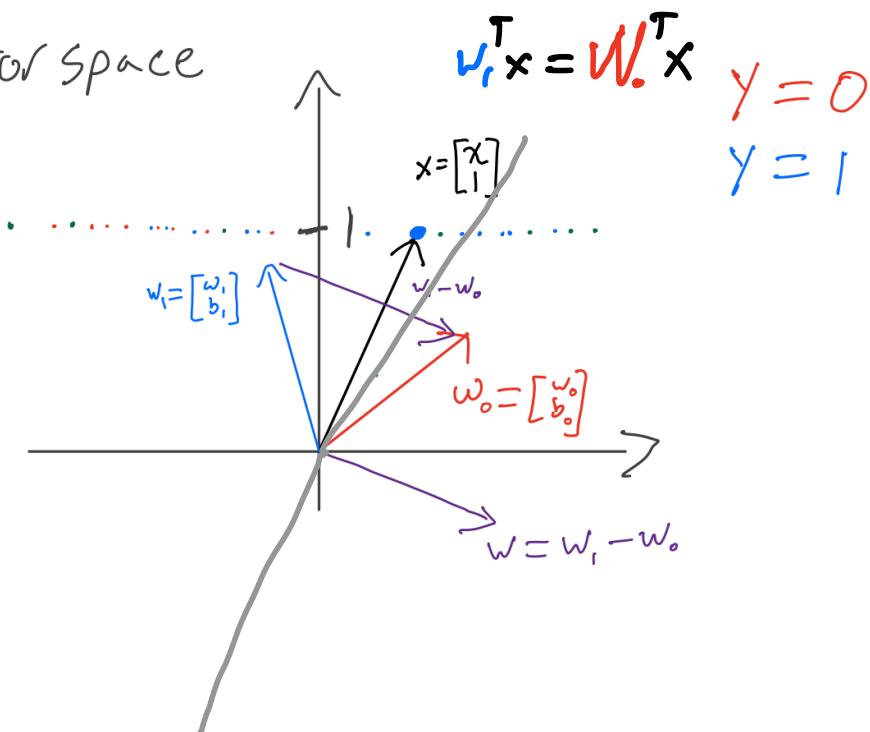
2-class: $w_i^T x > w_o^T x$

$$0 > w_i^T x - w_o^T x$$

$$0 > (w_i - w_o)^T x$$

w

Vector Space



Multi-class!

Prediction:

Predict y as $f(x)$
e.g. Brand? (e.g. horsepower)

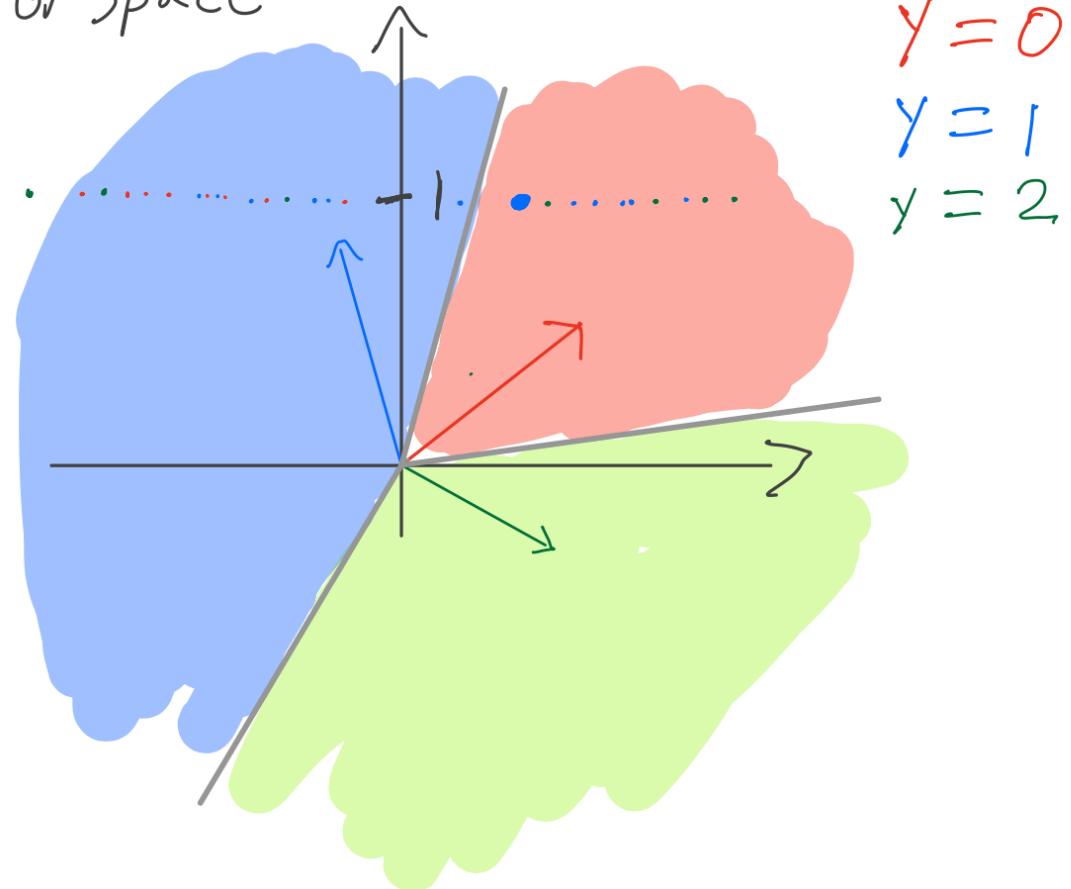
$$f(x) = \arg \max_c w_c x + b_c$$

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



$y = 0$
 $y = 1$
 $y = 2$

Multi-class prediction functions

$$\left[\begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{matrix} \right] = \left[\begin{matrix} x^T w_1 \\ x^T w_2 \\ \vdots \\ x^T w_C \end{matrix} \right]$$

Multi-class thresholding

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

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

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

columns

rows

d: dimension of inputs
(# features)

C: # classes

With this notation, our prediction function becomes:

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

Categorical distribution

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

Fair dice; $\left\{ \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \right\}$

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

$$P(\cdot = 1) = \frac{1}{6}$$

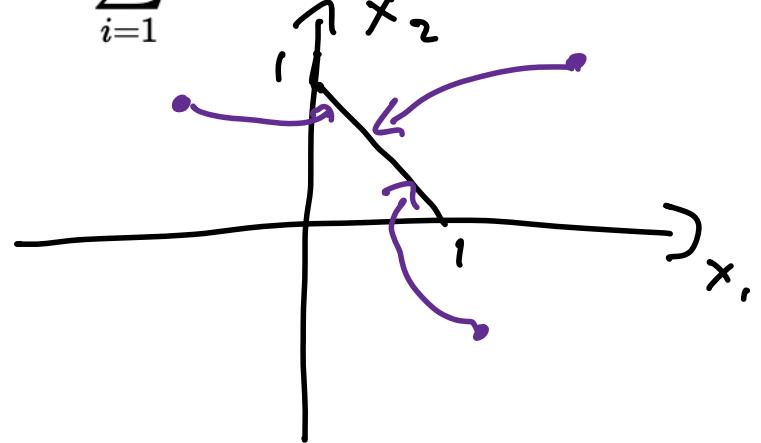
A probabilistic model for multi-class classification

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

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

Need $f(\mathbf{x}) : \mathbb{R}^C \rightarrow [0, \infty)^C, \quad \sum_{i=1}^C f(\mathbf{x})_i = 1$

$$\mathbb{R} \rightarrow [0, 1]$$
$$\sigma(x) := \frac{1}{1 + e^{-x}}$$



Categorical distribution

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

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

C , if $y=c \rightarrow 1$ otherwise 0

$$p(y) = \prod_{c=1}^C q_c^{\mathbb{I}(y=c)} = q_y$$

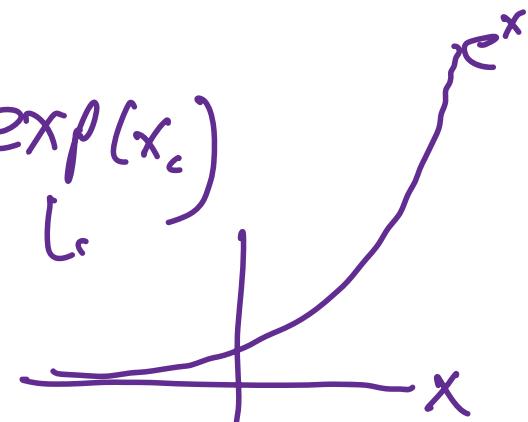
$$\log p(y) = \underbrace{\log \prod_{c=1}^C q_c^{\mathbb{I}(y=c)}}_{\text{L}} = \sum_{c=1}^C \mathbb{I}(y=c) \log q_c$$

Softmax function

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

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

$$| : x_c \rightarrow \exp(x_c)$$



\therefore

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

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

$$\text{softmax}(\mathbf{x})_c > 0$$

$$\sum_{i=1}^C \text{softmax}(\mathbf{x}) = 1$$

Softmax function

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

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

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

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

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

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

Maximum likelihood estimation for Multinomial Logistic regression

Model

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

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

Loss

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

$$\begin{aligned} \text{NLL}(\mathbf{W}, \mathbf{X}, \mathbf{y}) &= - \sum_{i=1}^N \log \frac{e^{\mathbf{x}_i^T \mathbf{w}_y}}{\sum_{j=1}^C e^{\mathbf{x}_i^T \mathbf{w}_j}} \\ &= - \sum_{i=1}^N \left[\mathbf{x}_i^T \mathbf{w}_y - \log \sum_{j=1}^C e^{\mathbf{x}_i^T \mathbf{w}_j} \right] \end{aligned}$$

Maximum likelihood estimation for Multinomial Logistic regression

Model

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

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

Loss

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

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

Models so far

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

Predict $y \in \mathbb{R}$ as

$$\begin{cases} y = \mathbf{x}^T \mathbf{w} \\ p(y | \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(y | \mathbf{x}^T \mathbf{w}, \sigma^2) \end{cases}$$

(prediction function) (probabilistic view)

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

Predict $y \in \{0, 1\}$ as

$$\begin{cases} y = \mathbb{I}(\mathbf{x}^T \mathbf{w} > 0) \\ p(y = 1 | \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}^T \mathbf{w}) \end{cases}$$

(prediction function) (probabilistic view)

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

Predict $y \in \{0, 1, \dots, C\}$ as

$$\begin{cases} y = \underset{c}{\operatorname{argmax}} \mathbf{x}^T \mathbf{w}_c \\ p(y = c | \mathbf{x}, \mathbf{w}) = \text{softmax}(\mathbf{x}^T \mathbf{W})_c \end{cases}$$

(prediction function) (probabilistic view)