Logistic Regression

CS344
Deep Learning

WELLESLEY

Common Problems ML May Help Solve

Logistic regression

BINARY CLASSIFICATION	MULTICLASS CLASSIFICATION	REGRESSION
Predicting 2 categorical outcomes	Predicting >2 categorical outcomes	Predicting a continuous outcome
Email is spam or not	Song is pop, rap, or country	Stock price
Someone has a disease or not	Flower is daisy, rose, sunflower, or tulip	Hours of sleep per night

Linear Classifier

$d = 2$

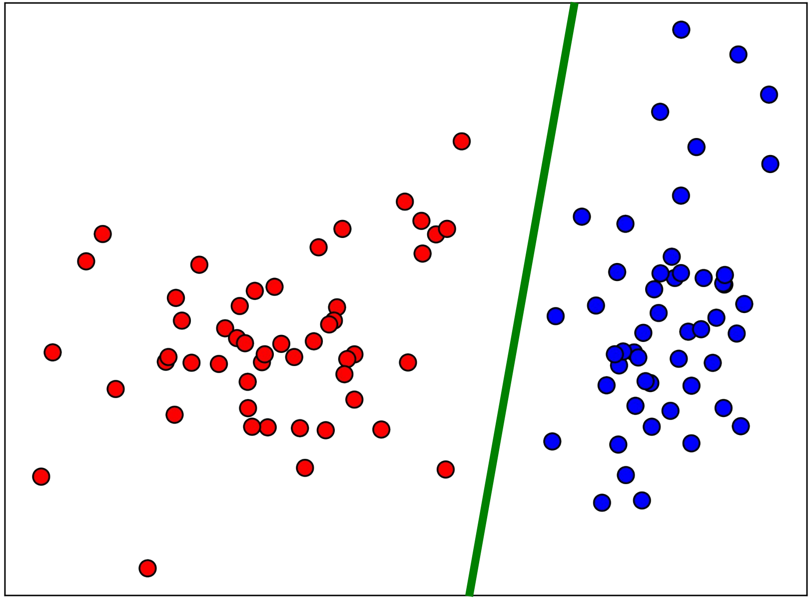
$m = 100$

35	59	0
72	63	1
13	77	0
22	33	0
58	19	1
…	…	…
80	37	1
44	51	0

Linear Classifier

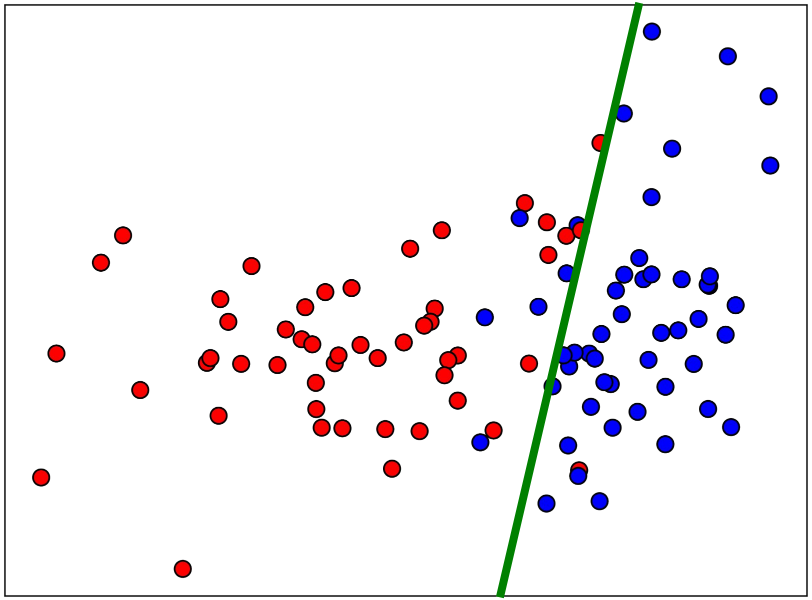
$$X = \begin{pmatrix} 35 & 59 \\ 72 & 63 \\ 13 & 77 \\ 22 & 33 \\ 58 & 19 \\ \dots & \dots \\ 80 & 37 \\ 44 & 51 \end{pmatrix} \qquad y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \dots \\ 1 \\ 0 \end{bmatrix}$$

(100, 2) (100,)



Logistic regression learns a linear decision boundary, i.e., a *hyperplane* that divides the two classes

Linear Classifier

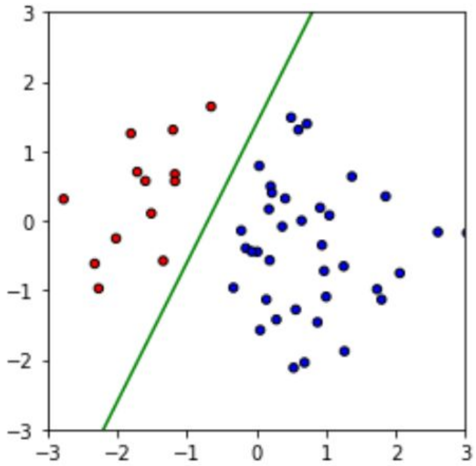


Data are not linearly separable

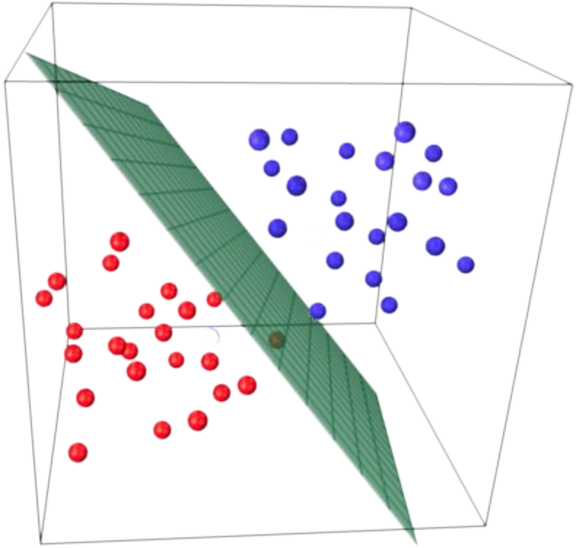
Hyperplane

A hyperplane in $\mathbb{R}^n$ is an n-1 dimensional subspace

A hyperplane in $\mathbb{R}^1$ is a point



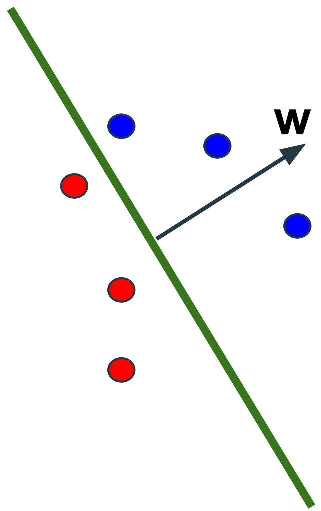
A hyperplane in $\mathbb{R}^2$ is a line



A hyperplane in $\mathbb{R}^3$ is a plane

What is a Hyperplane?

Parameterized by a “weight” vector **w** *orthogonal* to the hyperplane, centered at the origin

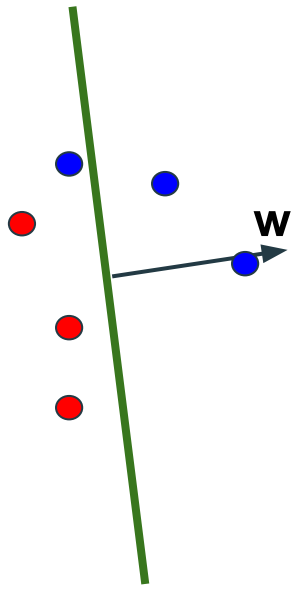


What range is

- the dot product of **w** with any of the blue points?
- the dot product of **w** with any of the red points?

What is a Hyperplane?

Parameterized by a “weight” vector **w** *orthogonal* to the hyperplane, centered at the origin

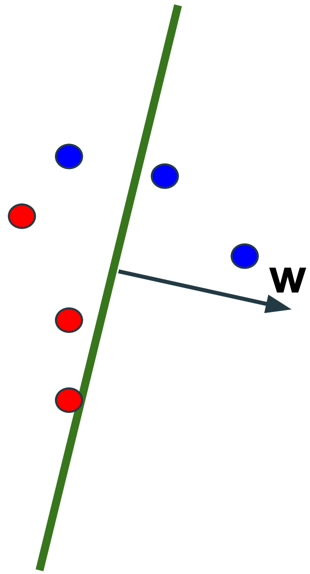


What range is

- the dot product of **w** with any of the blue points?
- the dot product of **w** with any of the red points?

What is a Hyperplane?

Parameterized by a “weight” vector **w** *orthogonal* to the hyperplane, centered at the origin



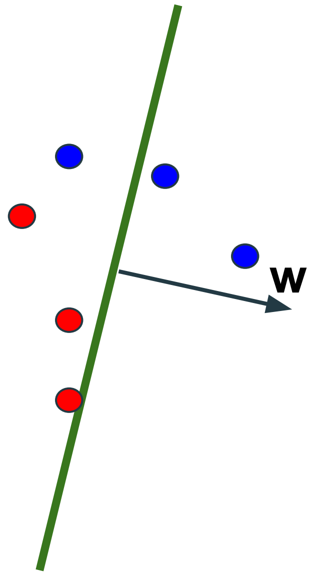
What range is

- the dot product of **w** with any of the blue points?
- the dot product of **w** with any of the red points?

What is a Hyperplane?

Adding a bias term *b*

Parameterized by a "weight" vector **w** *orthogonal* to the hyperplane, centered at the origin



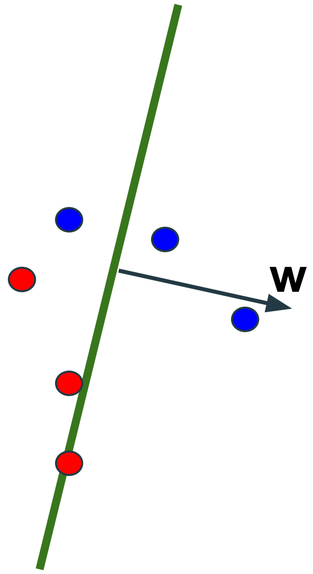
What range is

- the dot product of **w** with any of the blue points?
- the dot product of **w** with any of the red points?

What is a Hyperplane?

Adding a bias term *b*

Parameterized by a "weight" vector **w** *orthogonal* to the hyperplane, centered at the origin



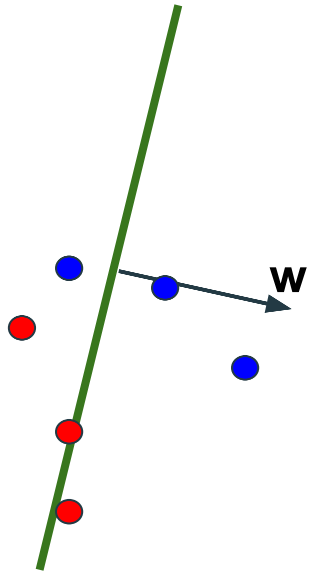
What range is

- the dot product of **w** with any of the blue points?
- the dot product of **w** with any of the red points?

What is a Hyperplane?

Adding a bias term *b*

Parameterized by a "weight" vector **w** *orthogonal* to the hyperplane, centered at the origin



What range is

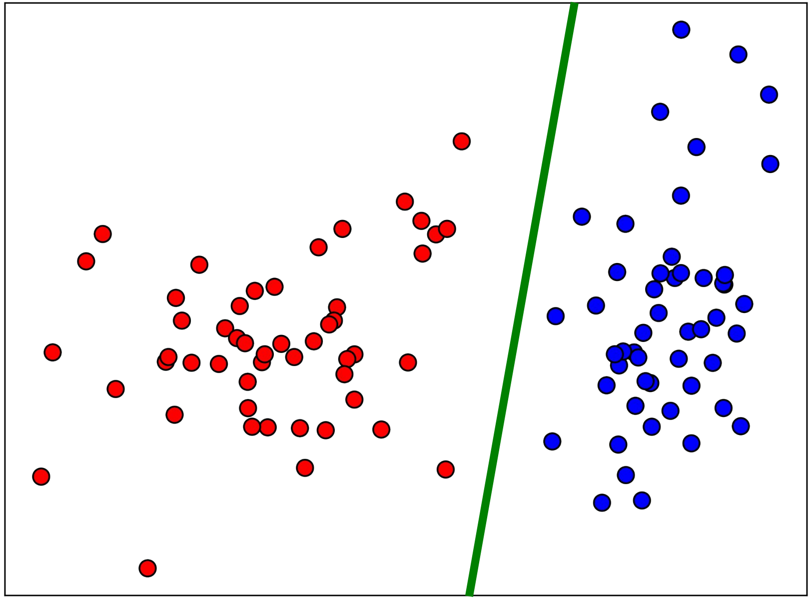
- the dot product of **w** with any of the blue points?
- the dot product of **w** with any of the red points?

Training

$b = 3$

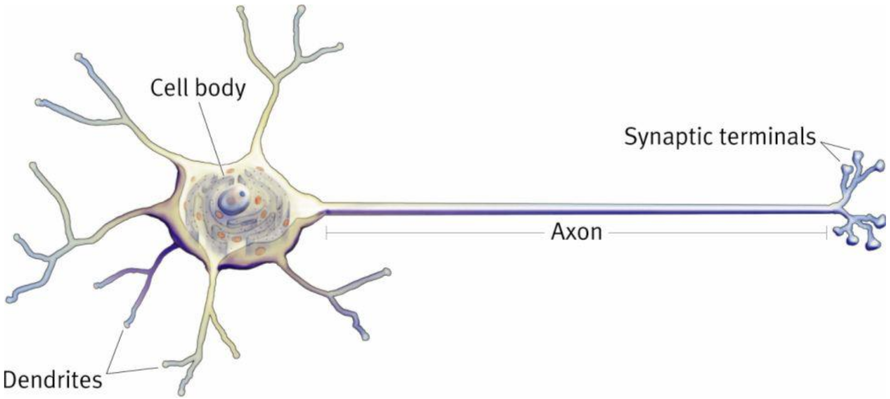
$w = (23 \quad 2)$

$$X = \begin{pmatrix} 35 & 59 \\ 72 & 63 \\ 13 & 77 \\ 22 & 33 \\ 58 & 19 \\ \dots & \dots \\ 80 & 37 \\ 44 & 51 \end{pmatrix} \quad y = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \dots \\ 1 \\ 0 \end{pmatrix}$$

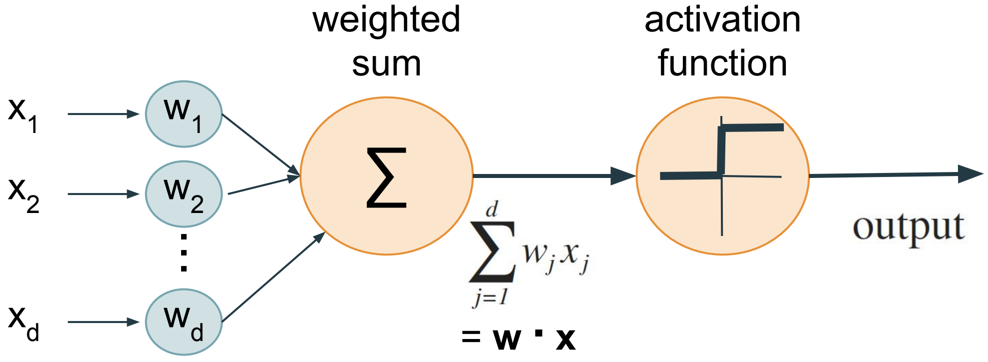


During *training*, the parameters of the model are learned from the *training data*

Neural Inspiration

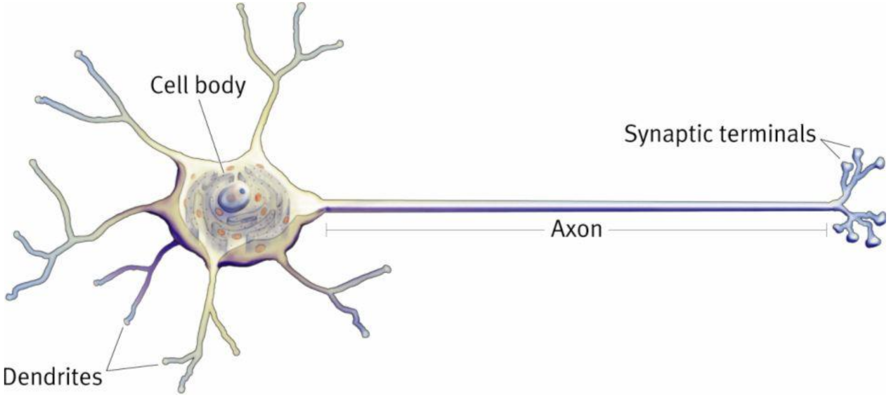


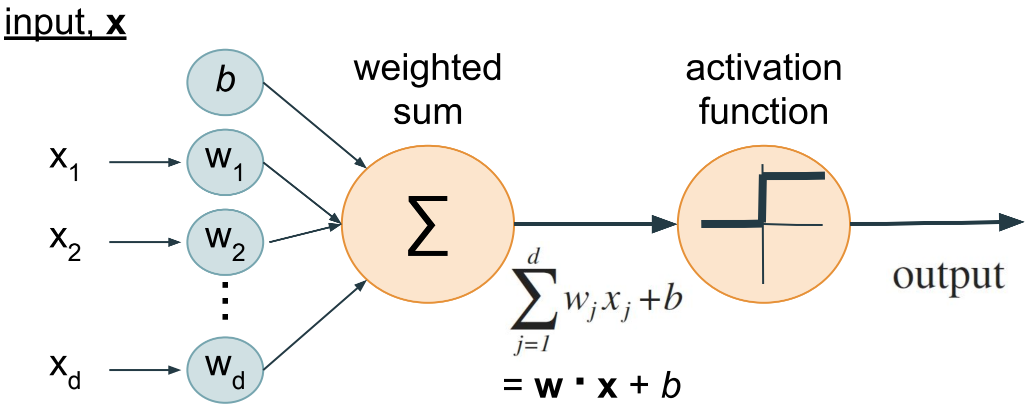
<u>input, **x**</u>



$$\text{output} = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} > 0 \\ 0 & \text{otherwise} \end{cases}$$

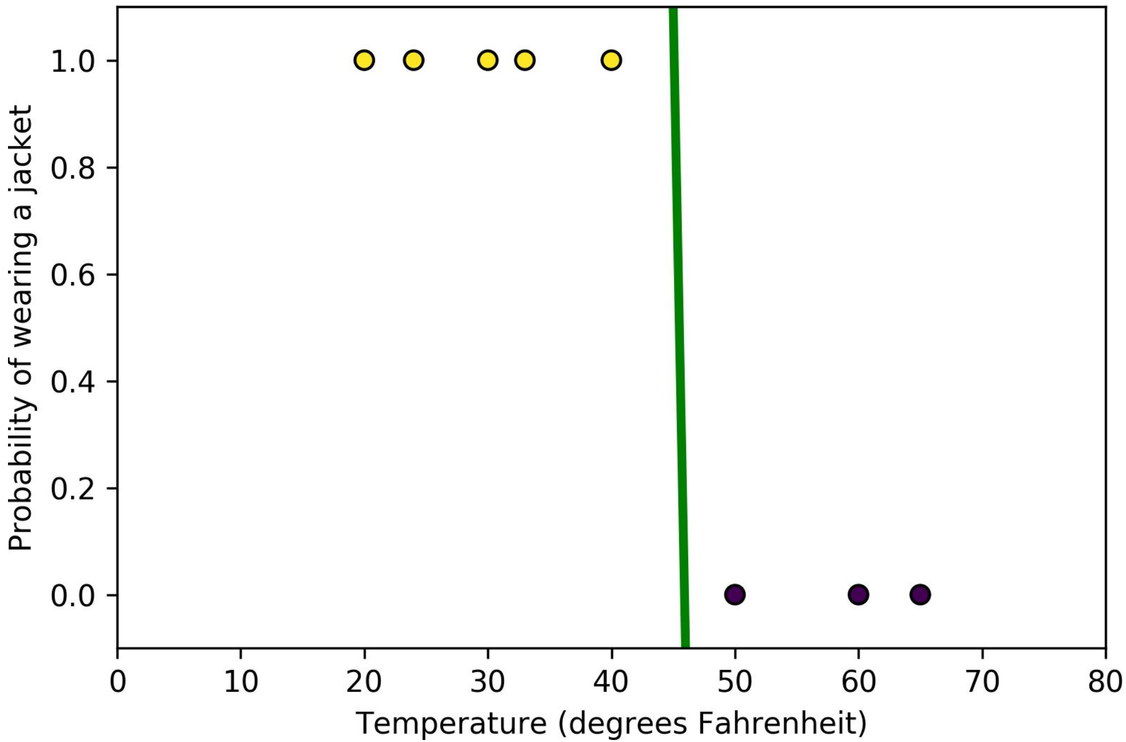
Neural Inspiration



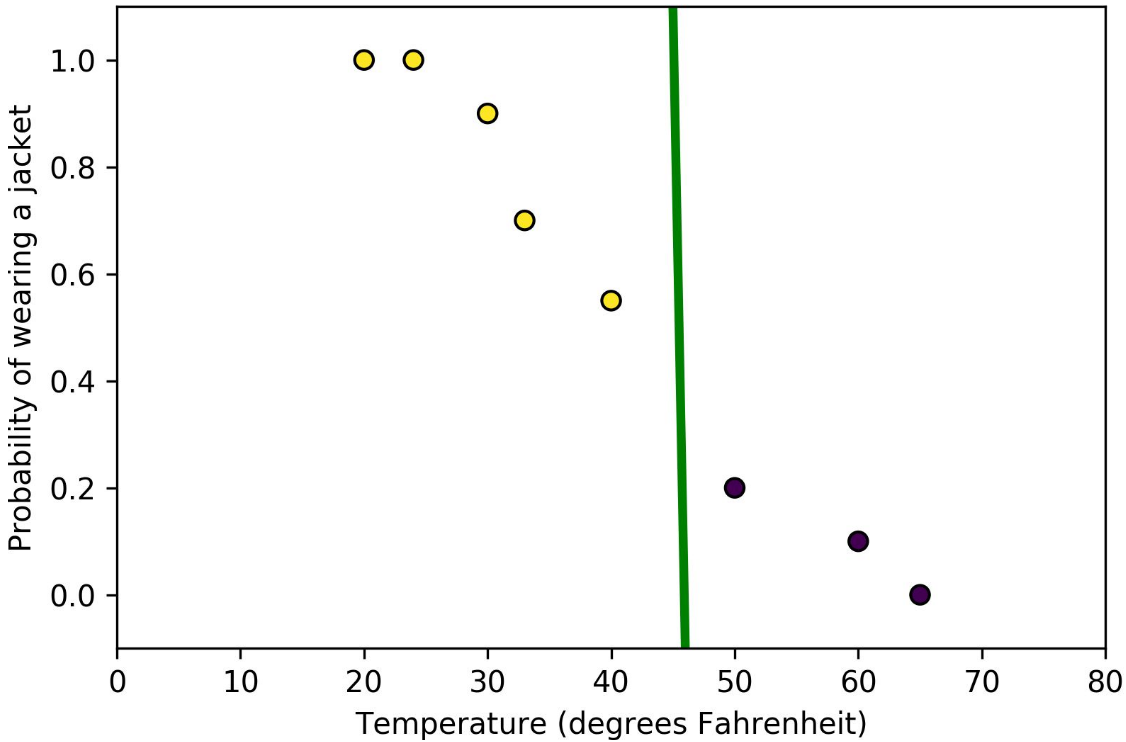


$$\text{output} = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

Should I wear a jacket?

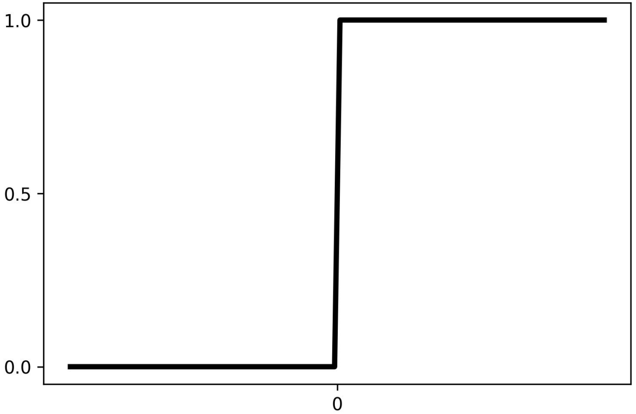


Should I wear a jacket?



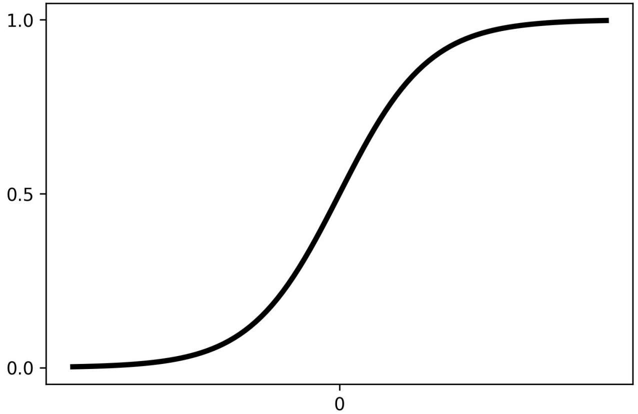
Hard Threshold vs. Sigmoid Function

Returns either 0 or 1



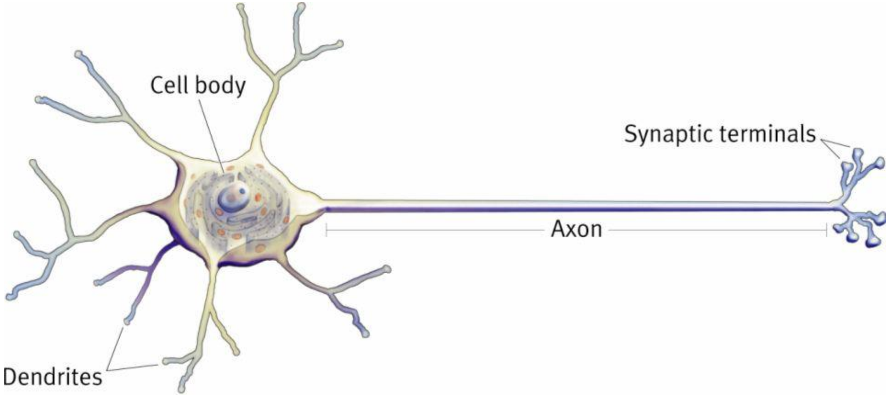
$$g(z) = \begin{cases} 0 \text{ if } & z<0 \\ 1 \text{ if } & z\geq 0 \end{cases}$$

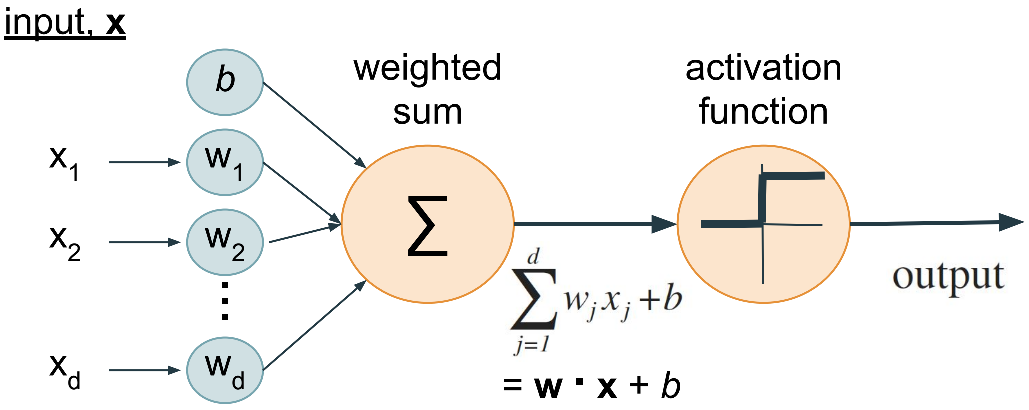
Returns a number between 0.0 and 1.0 that can be interpreted as a probability



$$g(z) = \frac{1}{1+e^{-z}}$$

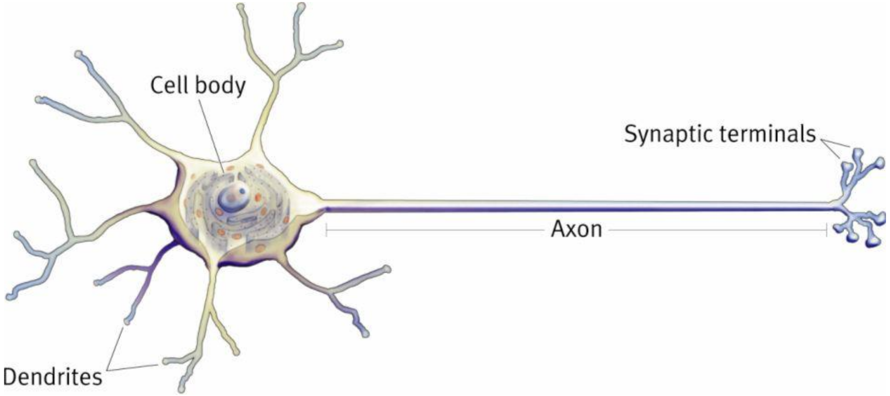
Neural Inspiration

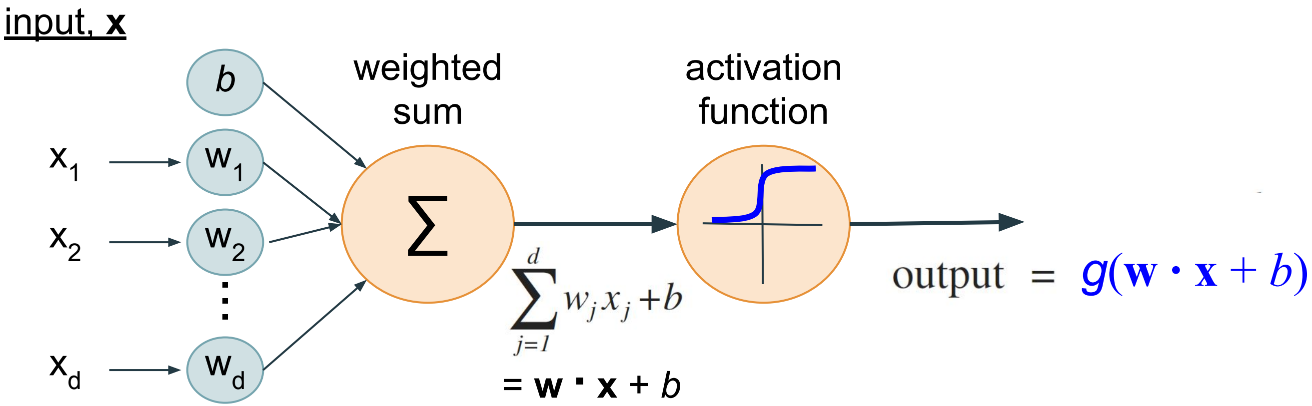




$$\text{output} = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

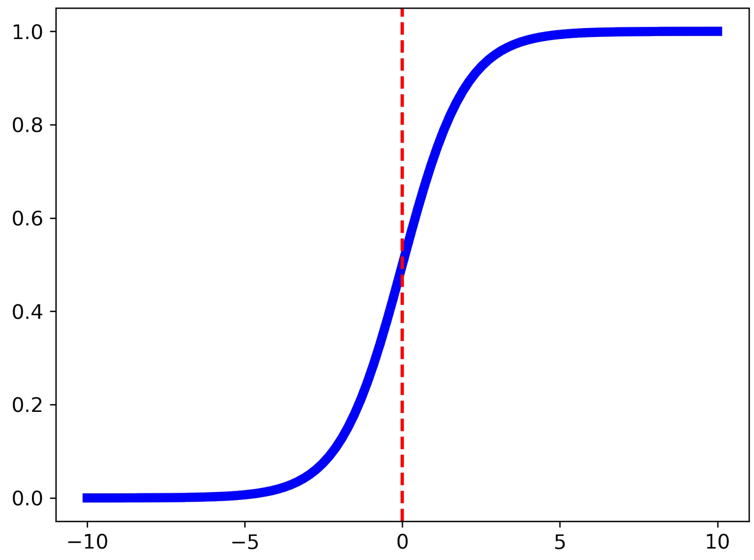
Neural Inspiration

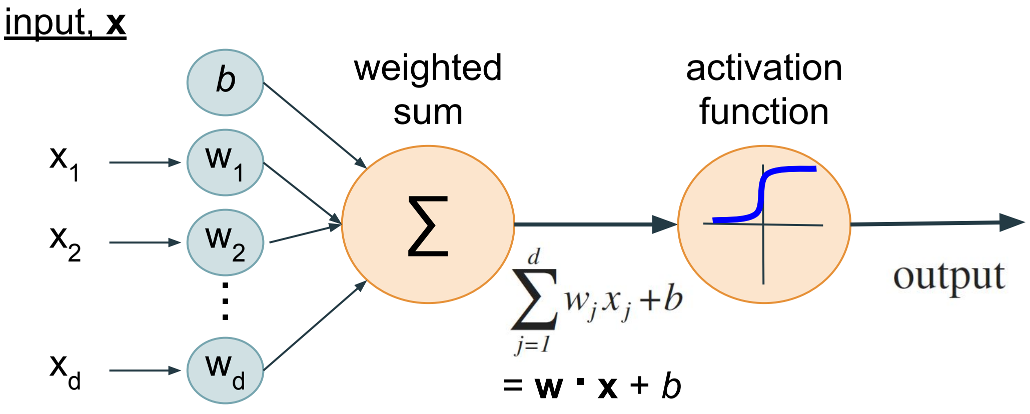




Sigmoid Function

$$g(z) = \frac{1}{1+e^{-z}}$$





$$\text{output} = g(\mathbf{w} \cdot \mathbf{x} + b)$$

$$= \frac{1}{1+e^{-(\mathbf{w}\cdot\mathbf{x}+b)}}$$

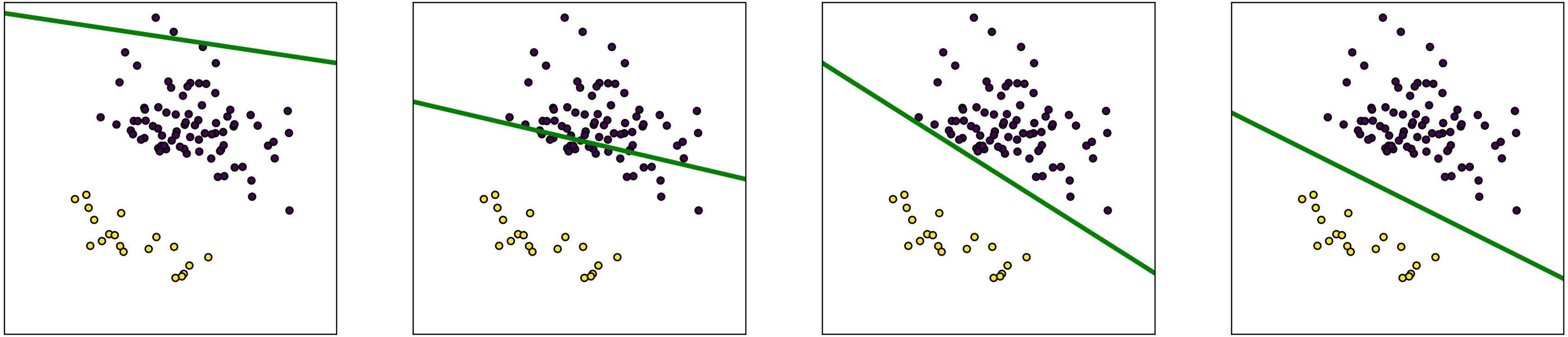
Forward Propagation

$$\hat{y} = g(\mathbf{w} \cdot \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w} \cdot \mathbf{x} + b)}}$$

- $\hat{y}$ is interpreted as the probability that $y = 1$ for input $\mathbf{x}$
- For example, what is the probability that some email message $\mathbf{x}$ is spam (1) as opposed to ham (0)?
  - If $\hat{y}$ is 0.25, the probability that the message is spam is 25% and we classify the message as ham (0)
  - If $\hat{y}$ is 0.75, the probability that the message is spam is 75% and we classify the message as spam (1)

Parameters w and *b*

Different values for parameters **w** and *b* lead to different decision boundaries

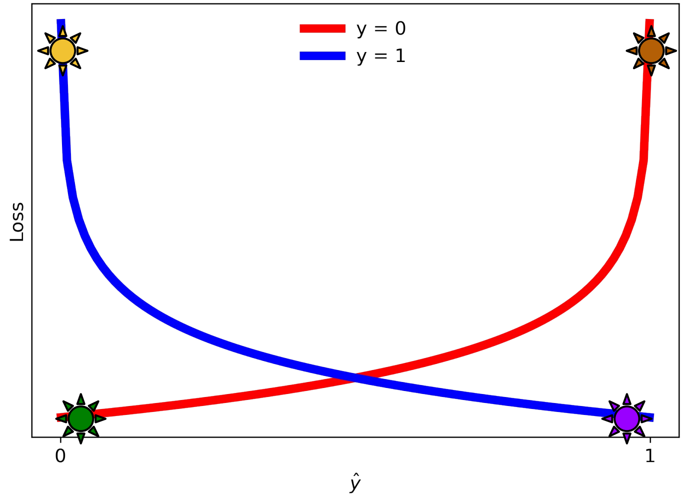


We want to quantify the *cost* associated with a given boundary (value settings for **w** and *b*) for our data

Then we can find the values of **w** and *b* that have the lowest *cost*

Loss

The *loss function*, L, quantifies the error, i.e., how far our prediction $\hat{y}$ is from the true label y



$$L = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

True label y	Prediction $\hat{y}$	Loss L
0	0.001	Close to 0
0	0.999	Large
1	0.999	Close to 0
1	0.001	Large

Cost

$$L = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

The *cost function*, J, is the average loss (error) of all *m* data points

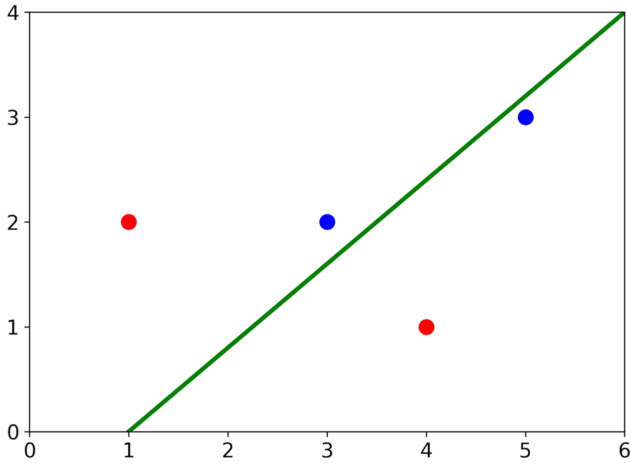
$$J = \frac{1}{m}\sum_{i=1}^{m} L$$

$$= \frac{1}{m}\sum_{i=1}^{m} -y_i\log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)$$

$$\hat{y} = g(\mathbf{w}\cdot\mathbf{x}+b) = \frac{1}{1+e^{-(\mathbf{w}\cdot\mathbf{x}+b)}}$$

Cost Examples

w = (4, -5)
b = -4



(4, 1)	(1, 2)	(3, 2)	(5, 3)	$\mathbf{x}$
7	-10	-2	1	$z = \mathbf{w} \cdot \mathbf{x} + b$
0.999	0.0004	0.12	0.73	$\hat{y} = g(z) = \dfrac{1}{1+e^{-z}}$
7.0009	0.0004	2.13	0.313	$L = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$

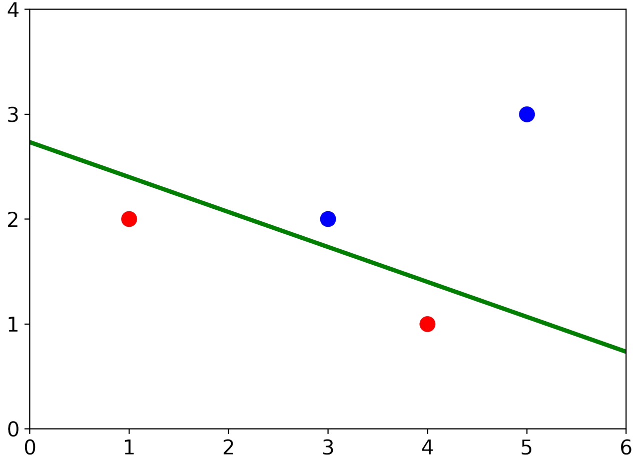
Average

2.36

$$J = \frac{1}{m}\sum_{i=1}^{m} L$$

Cost Examples

w = (1, 3)
b = -8.2



(4, 1)	(1, 2)	(3, 2)	(5, 3)	$\mathbf{x}$
-1.2	-1.2	-0.8	5.8	$z = \mathbf{w} \cdot \mathbf{x} + b$
0.23	0.23	0.69	0.997	$\hat{y} = g(z) = \dfrac{1}{1 + e^{-z}}$
0.26	0.26	0.37	0.003	$L = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$

Average

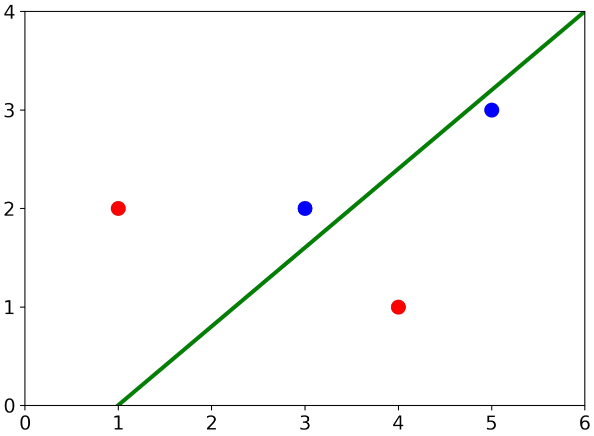
0.225

$$J = \frac{1}{m}\sum_{i=1}^{m} L$$

Cost Examples

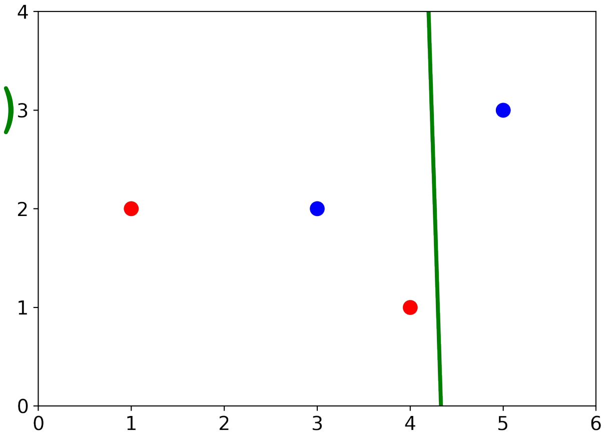
$w = (4, -5)$
$b = -4$

<u>Cost, J</u>
2.36



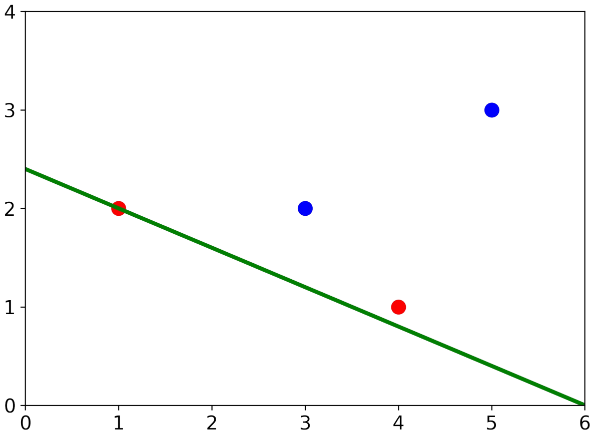
$w = (3, 0.1)$
$b = -13$

<u>Cost, J</u>
1.06



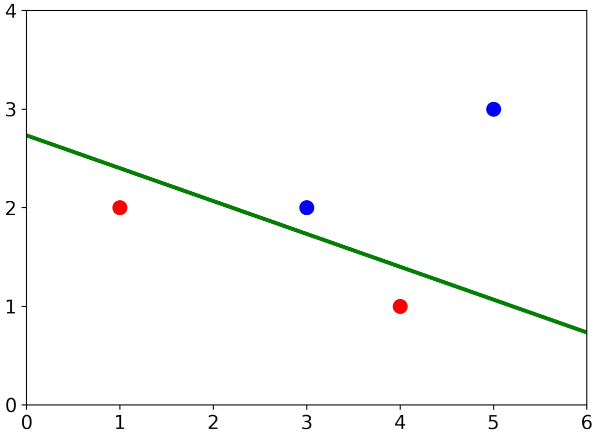
$w = (2, 5)$
$b = -12$

<u>Cost, J</u>
0.506

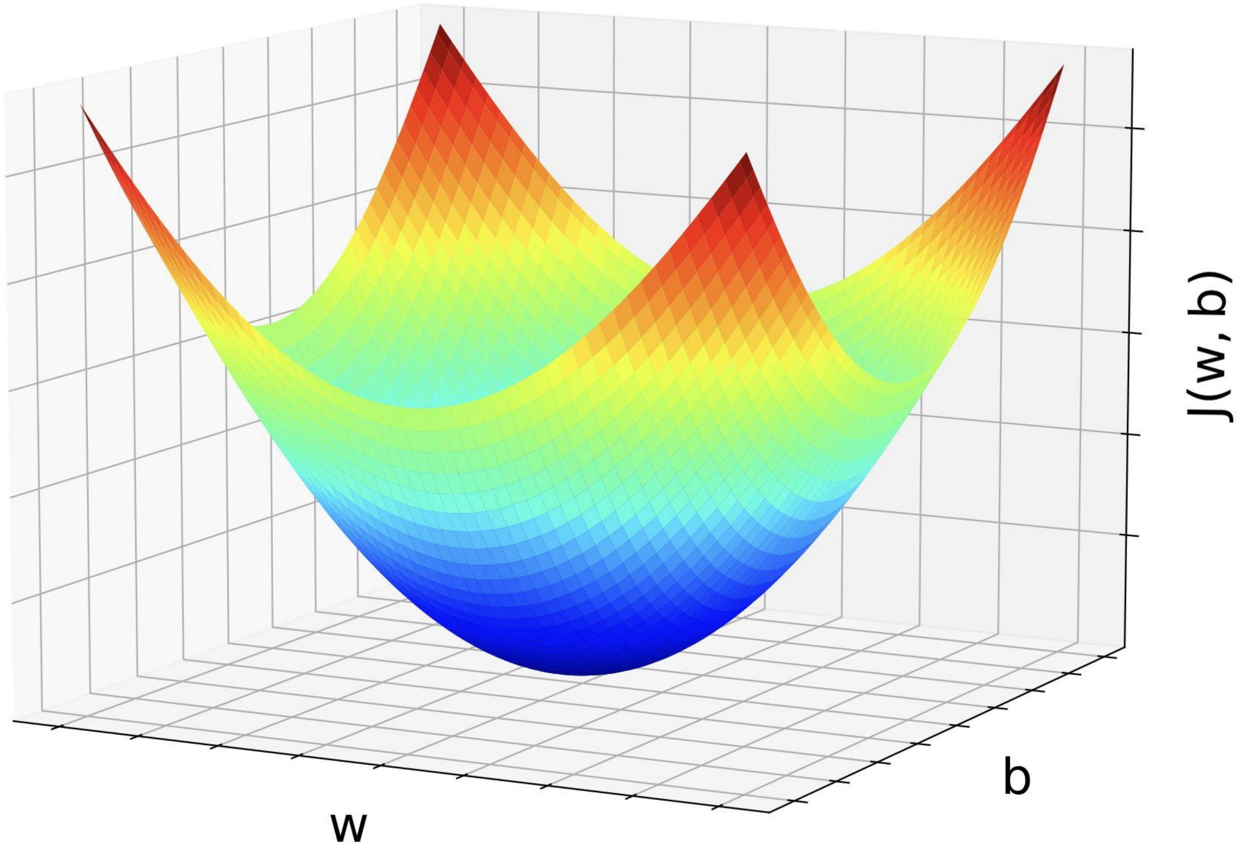


$w = (1, 3)$
$b = -8.2$

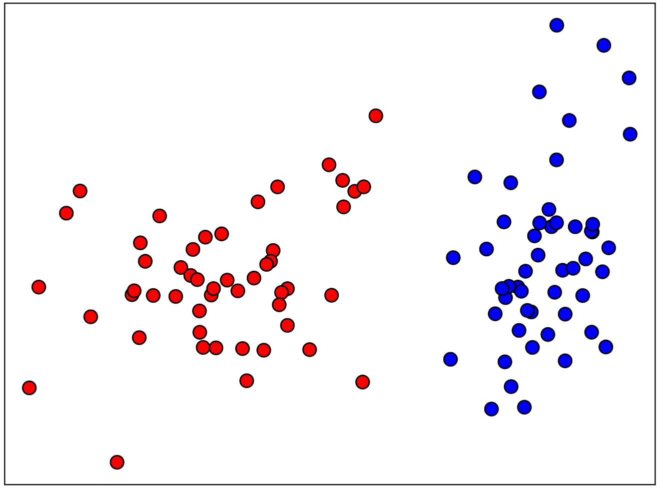
<u>Cost, J</u>
0.225



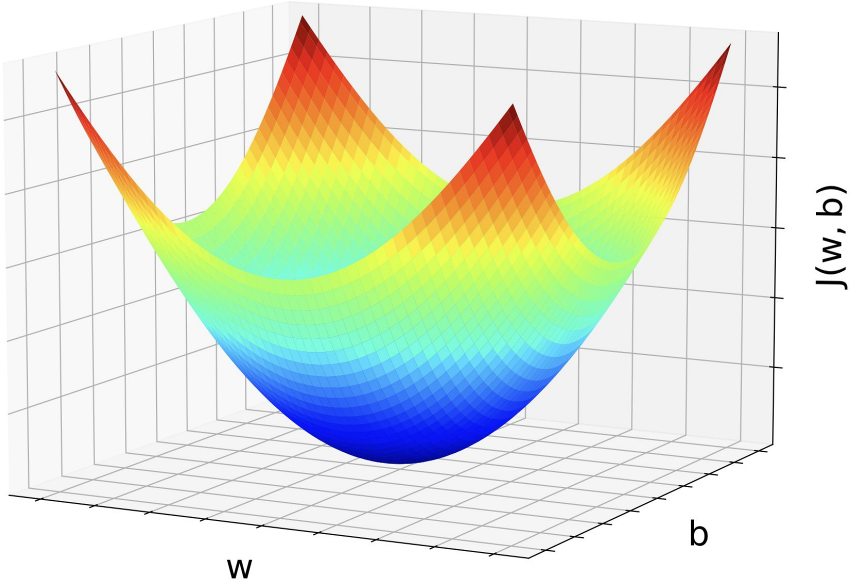
Cost Function



Cost Function

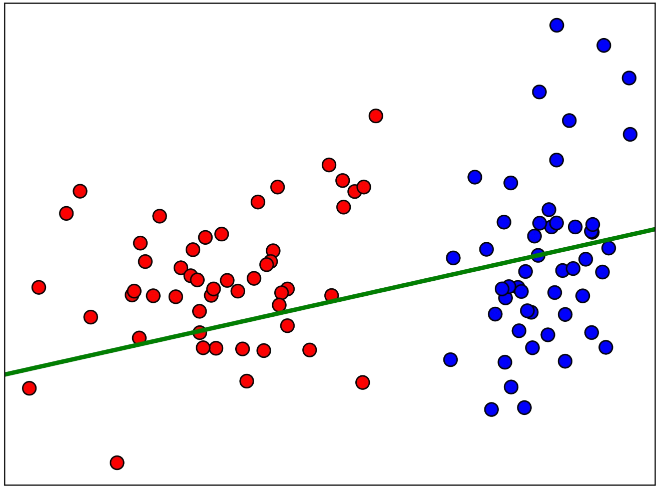


Data and Decision Boundary

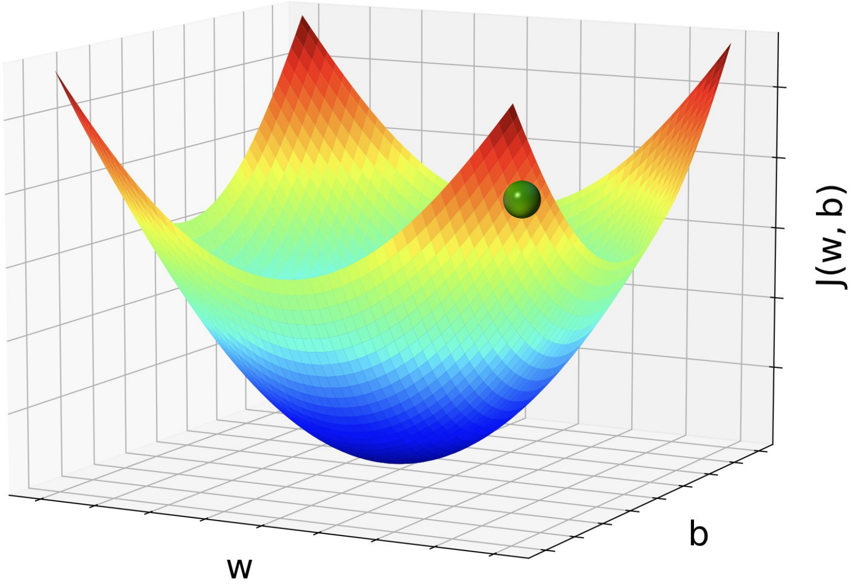


Cost Function

Cost Function

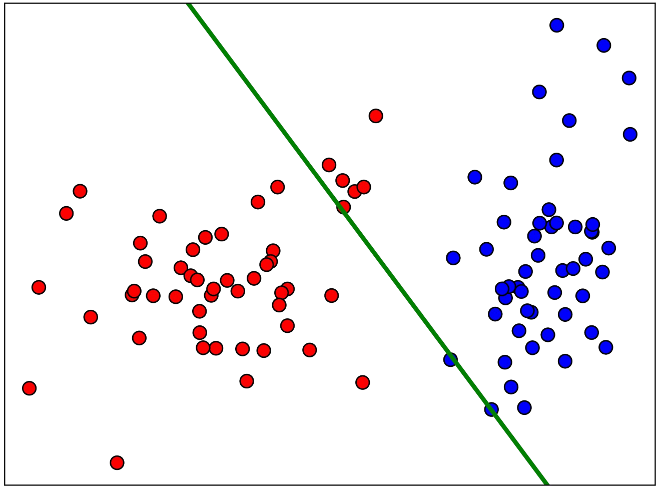


Data and Decision Boundary

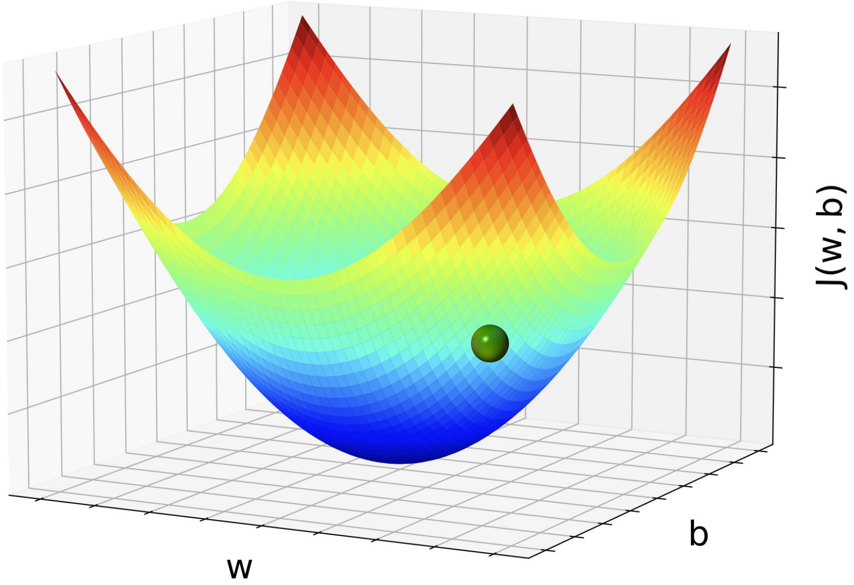


Cost Function

Cost Function

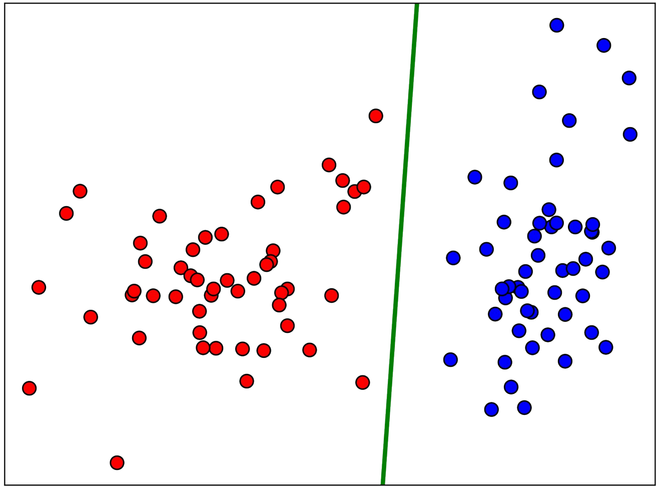


Data and Decision Boundary

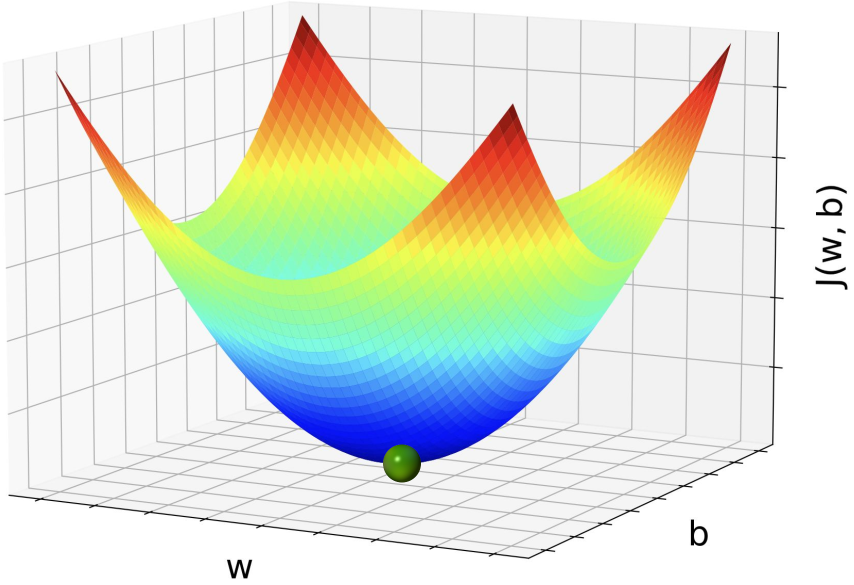


Cost Function

Cost Function



Data and Decision Boundary



Cost Function