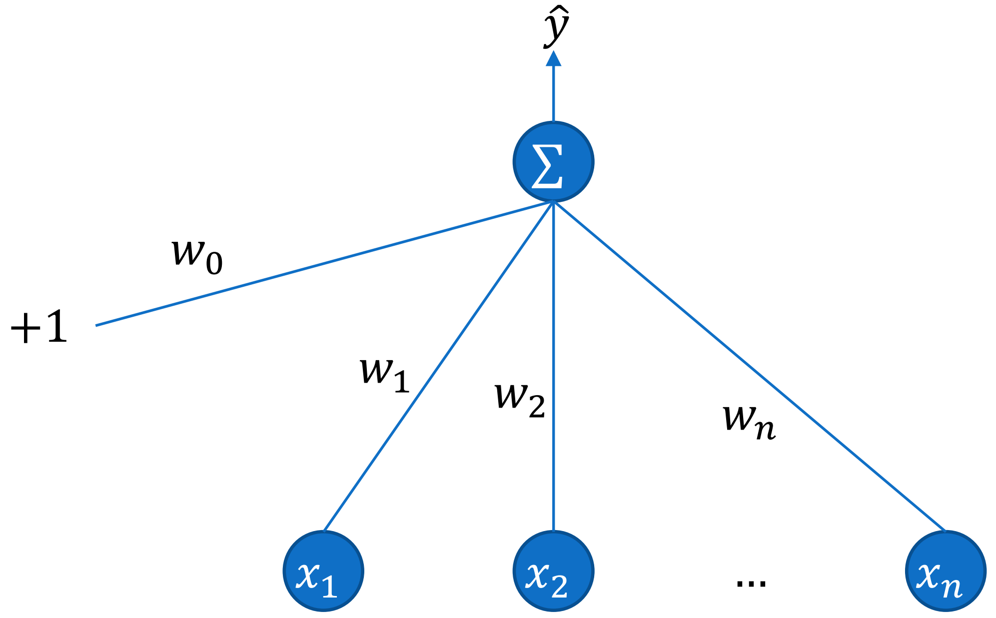
Review

LOGISTIC REGRESSION

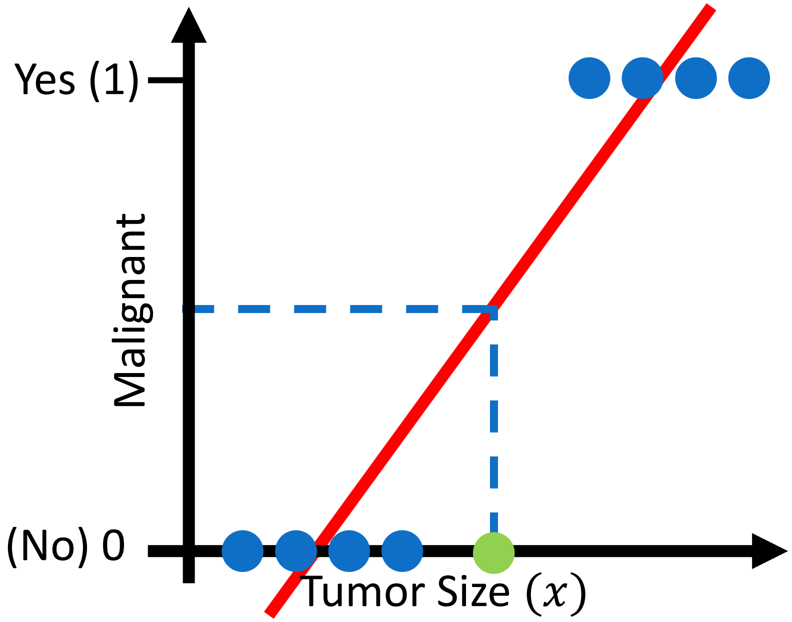
Linear Regression: A Visual Perspective

$$\boldsymbol{h(X) = W^T X = w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}$$

Compute Error: $y - \hat{y}$



Can we use Regression for Classification?



What will happen if we use Linear Regression?

$$h(X) = W^T X$$

What is the label for this data point?

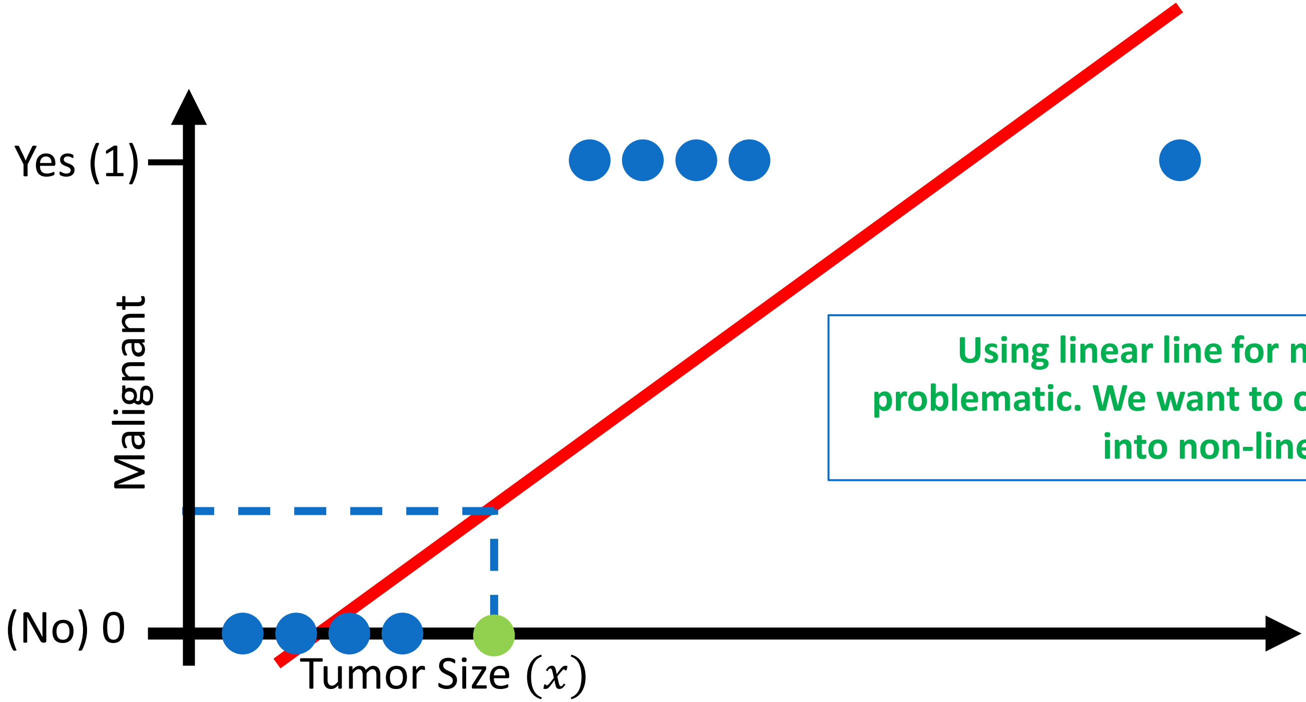
We need to define some threshold!

If $h(X) \geq 0.5$, predict $y = 1$
Else, predict $y = 0$

This also mean the output should be between 0-1 for this threshold to work!

A Threshold classifier $h(X)$ at 0.5

What about this case?



Using linear line for non-linear data is problematic. We want to convert this linear line into non-linear line!

This threshold does not work now!

A Threshold classifier $h(X)$ at 0.5

If $h(X) \geq 0.5$, predict $y = 1$
Else, predict $y = 0$

Online Demo

- https://www.desmos.com/calculator

$$\boldsymbol{h(x) = \sigma(z) = \frac{1}{1 + e^{-(w_0 + w_1 x_1)}}}$$

Or equally…

$$\boldsymbol{h(x) = \sigma(z) = \frac{1}{1 + e^{-(W^T X)}}}$$

Or equally…

$$\boldsymbol{z = w_0 + w_1 x_1}$$

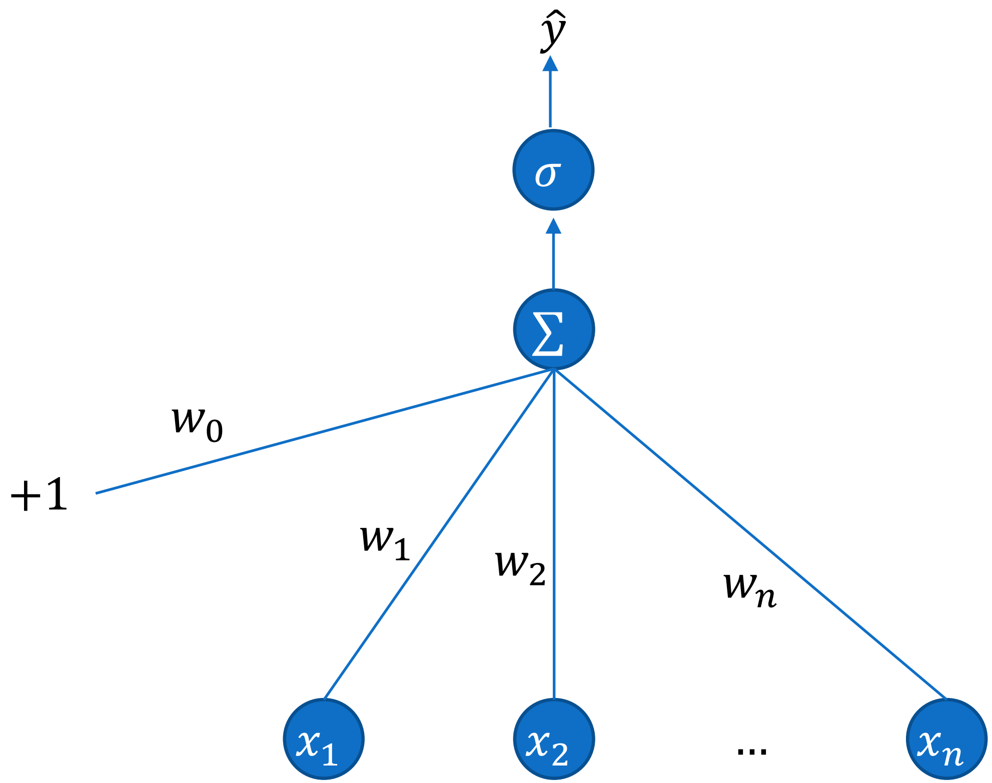
$$\boldsymbol{h(x) = \sigma(z) = \frac{1}{1 + e^{-z}}}$$

Just finding the S curve is not important. Bias (w_0) is also equally important that determines the location of the threshold 0.5.

Logistic Regression: A Visual Perspective

$$h(x) = \sigma(z) = \frac{1}{1 + e^{-(W^T X)}}$$

Compute Error: $y - \hat{y}$



Advantages of a Sigmoid

- Maps real-valued numbers ($\mathbb{R}$) into the range $[0,1]$
- Nearly linear around 0 but has a shar slope toward the ends
- It tends to squash outlier values toward 0 or 1
- It is differentiable, which is handy for learning
- To make it a probability:

$$\boldsymbol{P(y=1) = \sigma(W^T X) = \frac{1}{1+e^{-W^T X}}}$$

$$\boldsymbol{P(y=1) = 1 - \sigma(W^T X) = 1 - \frac{1}{1+e^{-W^T X}}}$$

- **How do we make decisions about label?**
  - For a test instance x_1, we say **yes** if the probability $P(y=1)$ is equal or greater than 0.5, and **no** otherwise.
  - We call **0.5 the decision boundary**

$$\mathbf{h(X) = \hat{y} = \begin{cases} 1 \text{ if } P(y=1|x) \geq 0.5 \\ 0 \text{ otherwise} \end{cases}}$$

Example: Sentiment Classification

- Binary sentiment classification on movie review text
  - 6 features $x_1, \ldots x_6$ of input
  - Learned weights for each of these features : $[2.5, -5.0, -1.2, 0.5, 2.0, 0.7]$, while $w_0 = 0.1$.
  - Note that $w_1 = 2.5$ is positive, while $w_2 = -5.0$ is negative, means:
    - Negative words are negatively associated with a positive sentiment decisions and are about twice as important as positive words.

Var	Definition
x_1	count(positive lexicon) $\in$ doc)
x_2	count(negative lexicon) $\in$ doc)
x_3	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$
x_4	count(1st and 2nd pronouns $\in$ doc)
x_5	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$
x_6	log(word count of doc)

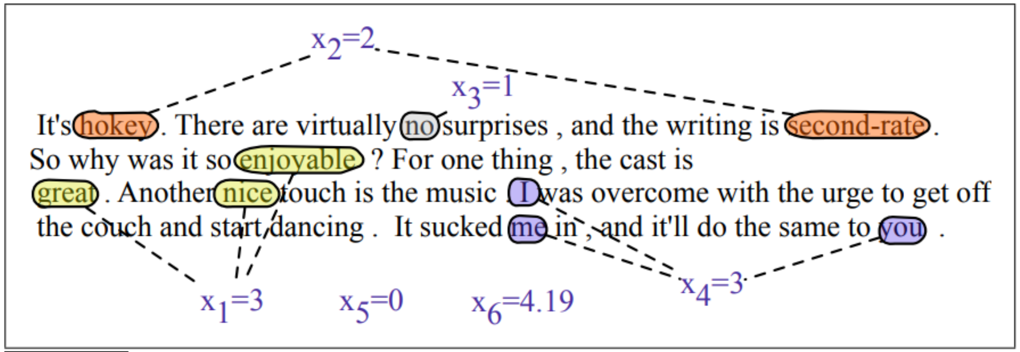


Figure 5.2 A sample mini test document showing the extracted features in the vector x.

x_1	x_2	x_3	x_4	x_5	x_6	y
3	2	1	3	0	4.19	?

Example: Sentiment Classification

- Learned weights for each of these features : $[2.5, -5.0, -1.2, 0.5, 2.0, 0.7]$, while $w_0 = 0.1$

x_1	x_2	x_3	x_4	x_5	x_6	y
3	2	1	3	0	4.19	?

$$P(+|x) = P(y=1|x) = \sigma(w.x + b)$$
$$= \sigma([2.5, -5.0, -1.2, 0.5, 2.0, 0.7].[3, 2, 1, 3, 0, 4.19] + 0.1)$$
$$= \sigma(.833)$$
$$= 0.7$$

Or equally…

$$P(+|X) = P(y=1|X) = \sigma(W.X)$$
$$= \sigma([0.1, 2.5, -5.0, -1.2, 0.5, 2.0, 0.7].[1, 3, 2, 1, 3, 0, 4.19])$$
$$= \sigma(.833)$$
$$= 0.7$$

Whats the probability of negative class?

$$P(-|X) = P(y=0|X) = 1 - \sigma(W.X)$$
$$= 1 - 0.7$$
$$= 0.3$$

Putting it all together…

- Use Sigmoid to squash the output in range 0-1
- Perform thresholding to convert the output probabilities into categorical labels
- That's how, we can use regression for classification!

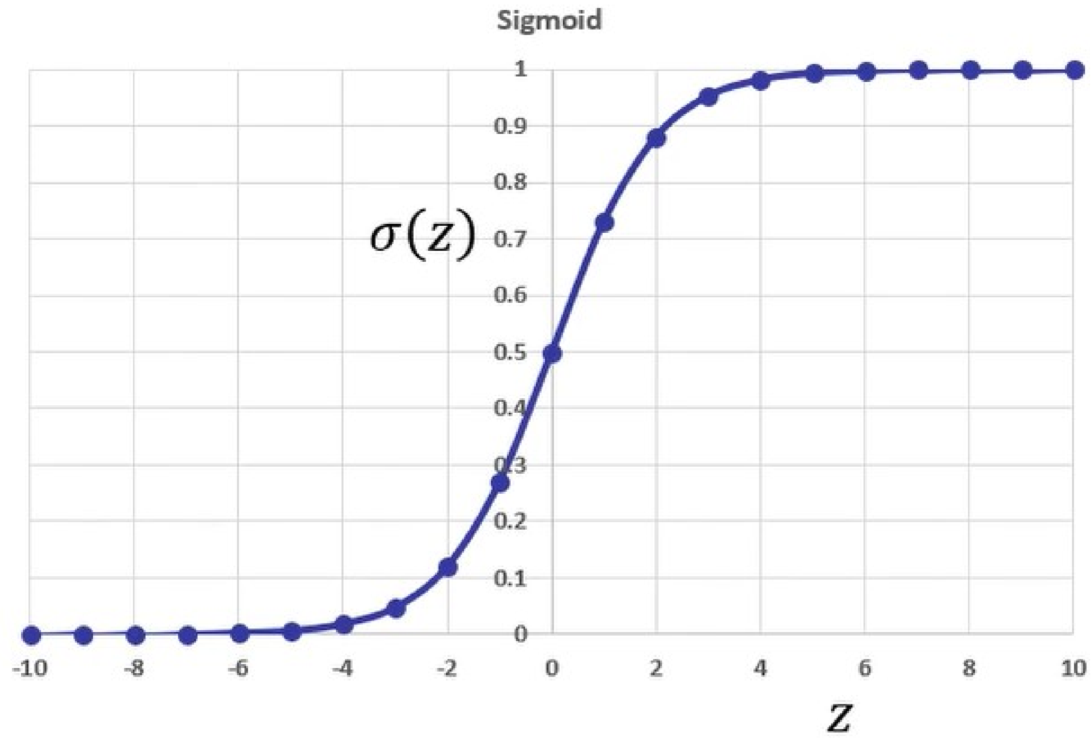
Now that the output is "activated" by sigmoid function, what will happen to the cost function?

Visualizing Decision Boundary in Logistic Regression

Logistic Regression

$$z = W^T X$$

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

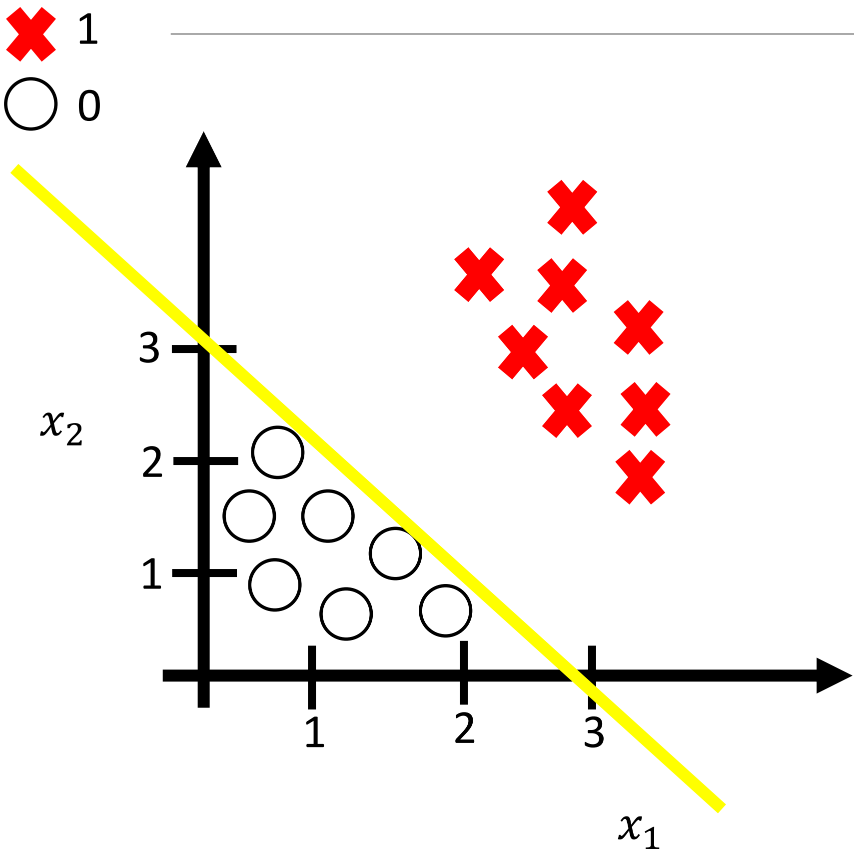


Predict "$y = 1$" if $\sigma(z) \geq 0.5$
i.e., $W^T X \geq 0$
Predict "$y = 0$" if $\sigma(z) < 0.5$
i.e., $W^T X < 0$

Min value is 0 and max is 1

Side Note: We are always interested in error and not accuracy. Why?

Decision Boundary



$$h(x) = (w_0 + w_1 x_1 + w_2 x_2)$$

Suppose we are able to train a model an get the following weights…

$$W = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

Predict $y = 1$, if $-3 + 1 \times x_1 + 1 \times x_2 \geq 0$

Predict $y = 1$, if $-3 + x_1 + x_2 \geq 0$

Predict $y = 1$, if $x_1 + x_2 \geq 3$

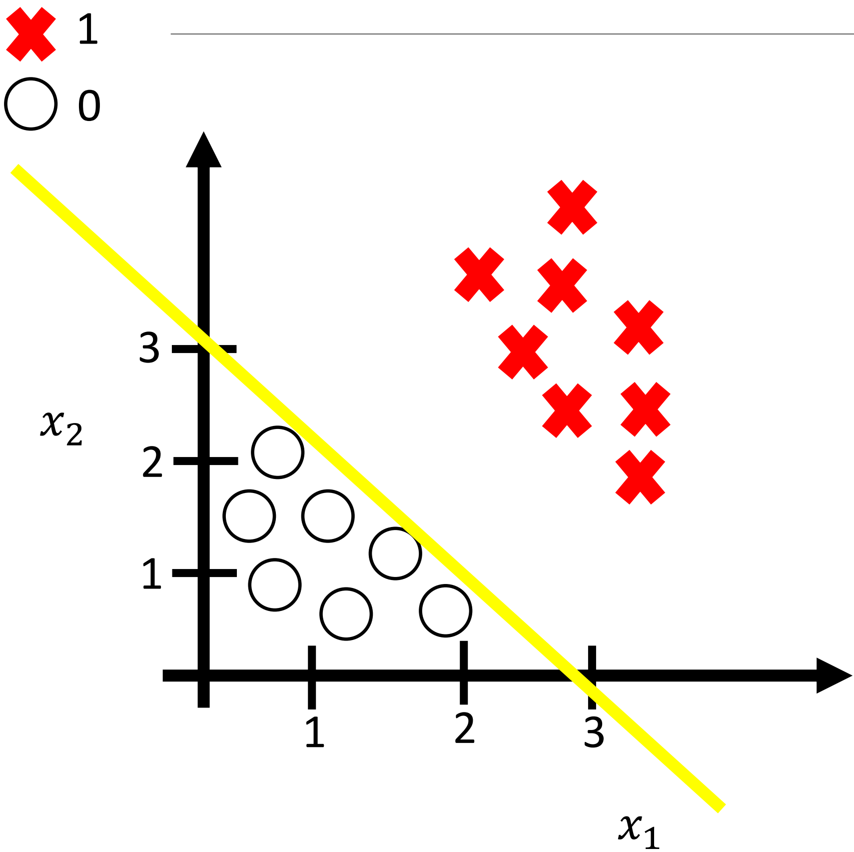
Side Note: How many decision boundaries are possible between these two classes?

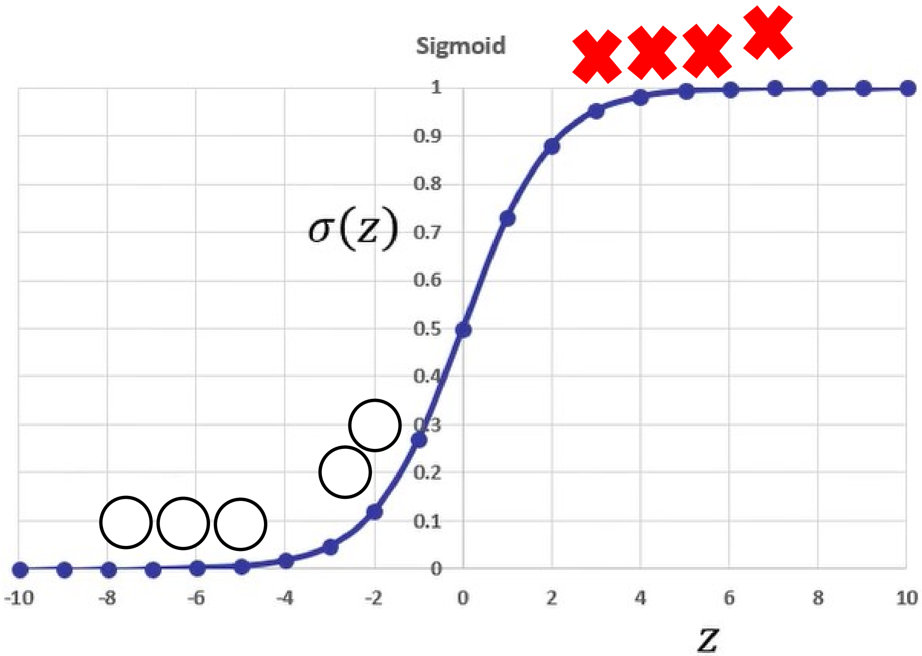
Infinite!

Where decision boundary represents: $x_1 + x_2 = 3$

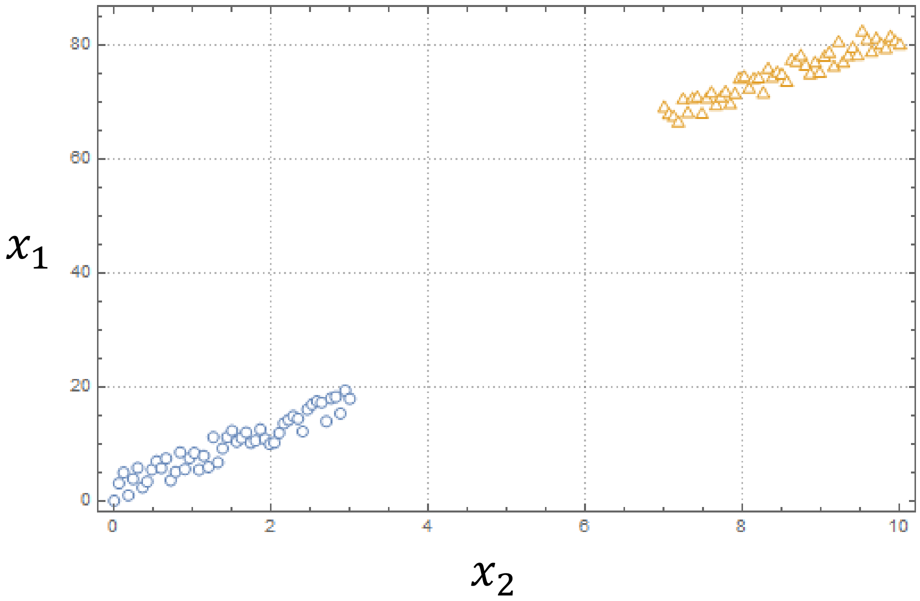
Where is sigmoid in this boundary?

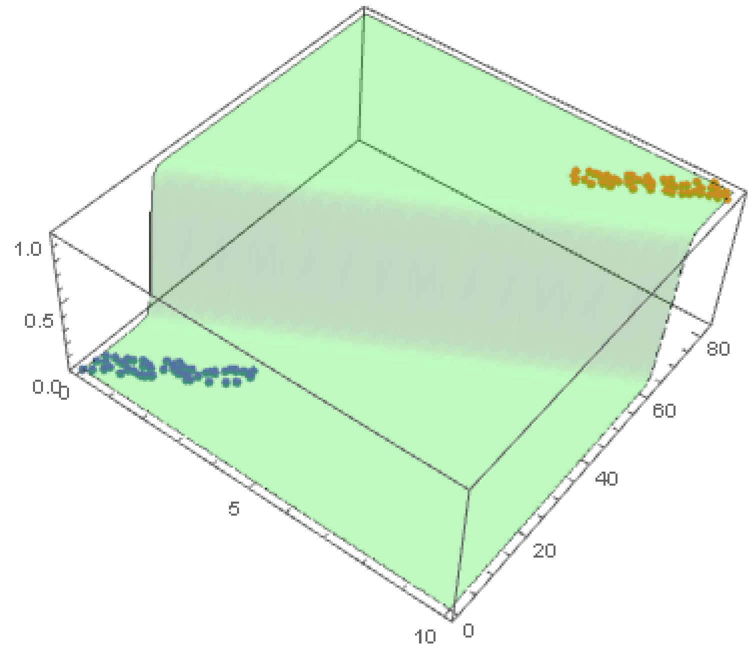
Decision Boundary

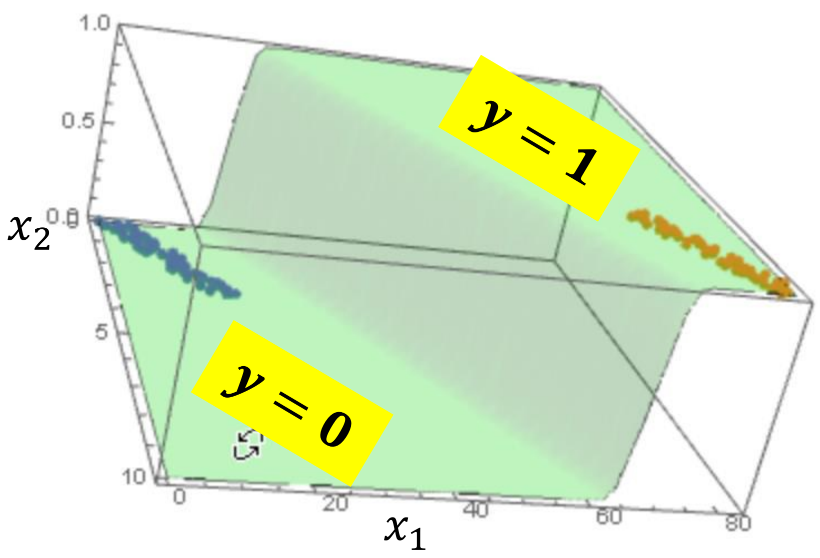




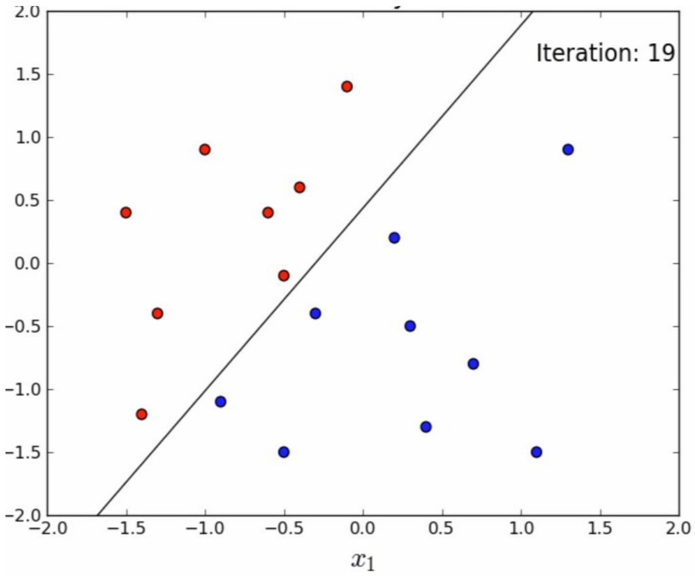
If decision boundary after sigmoid is 0.5, then…

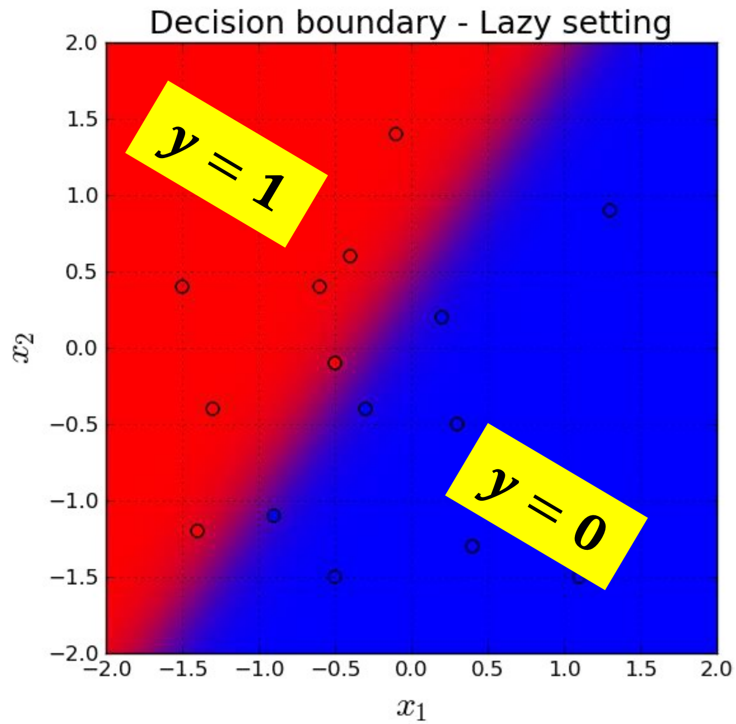


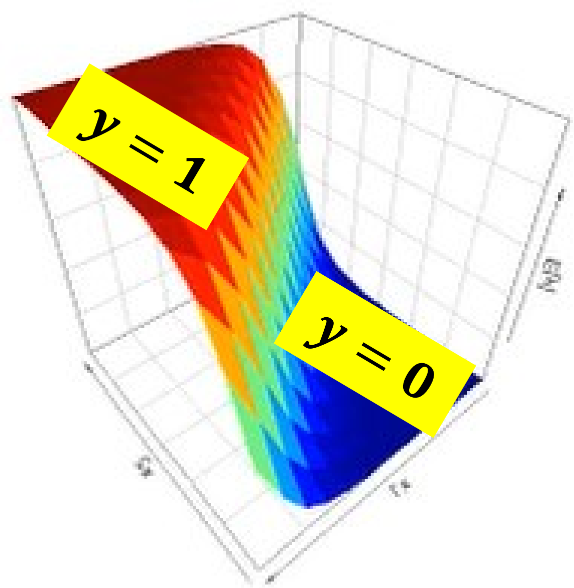




Credit: https://datascience.stackexchange.com/questions/49573/how-to-plot-logistic-regression-decision-boundary

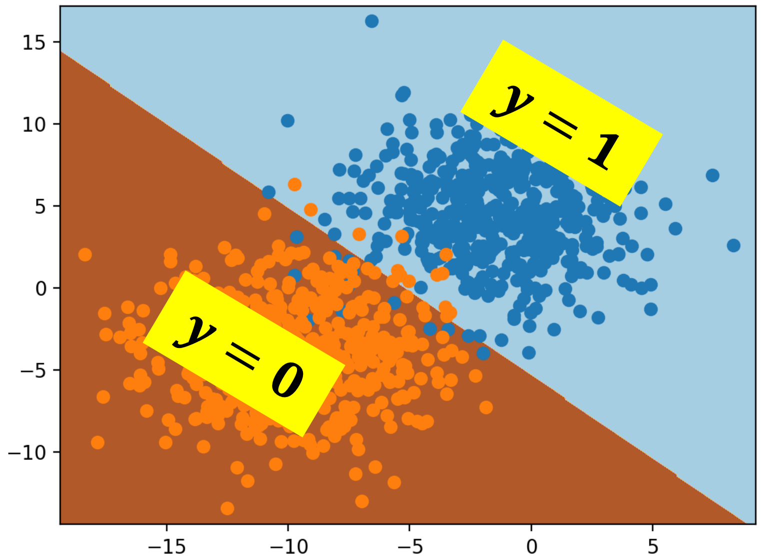


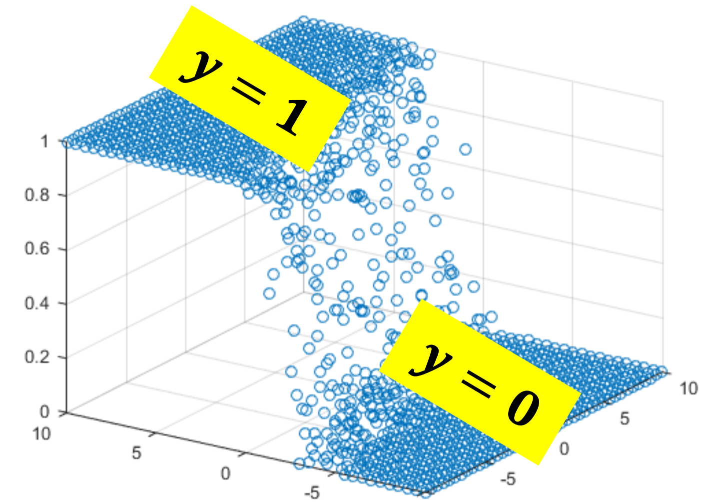


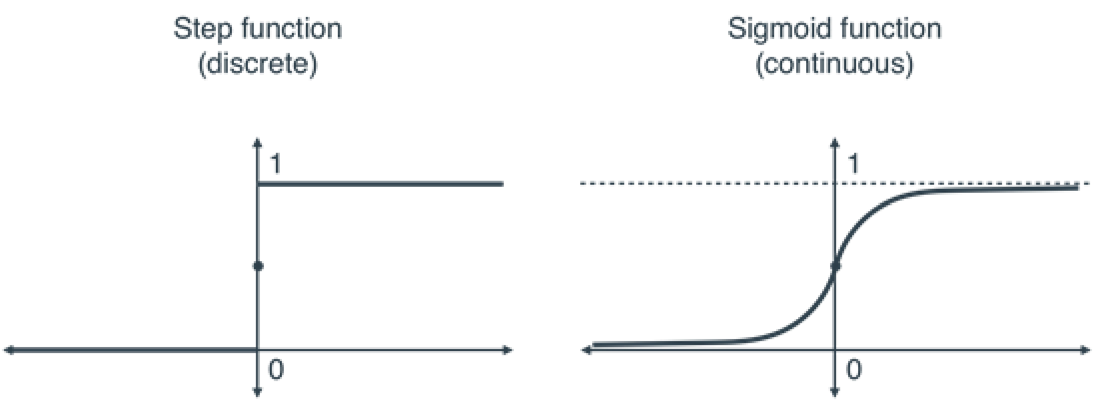


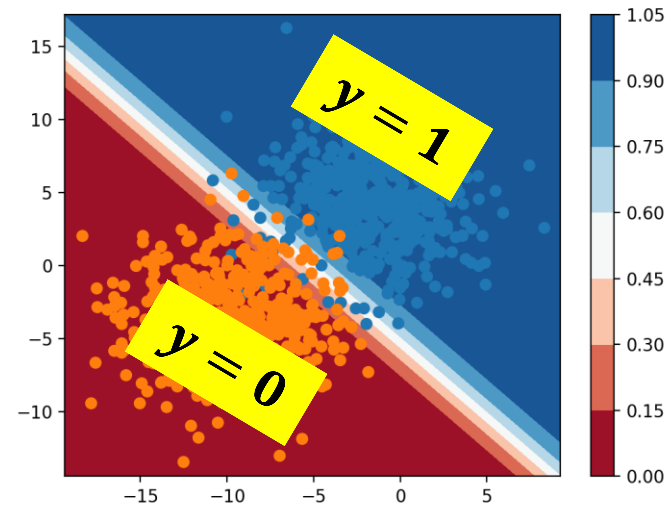
Credit: http://saitcelebi.com/tut/output/part1.html

Hard VS Soft Boundaries Classifiers



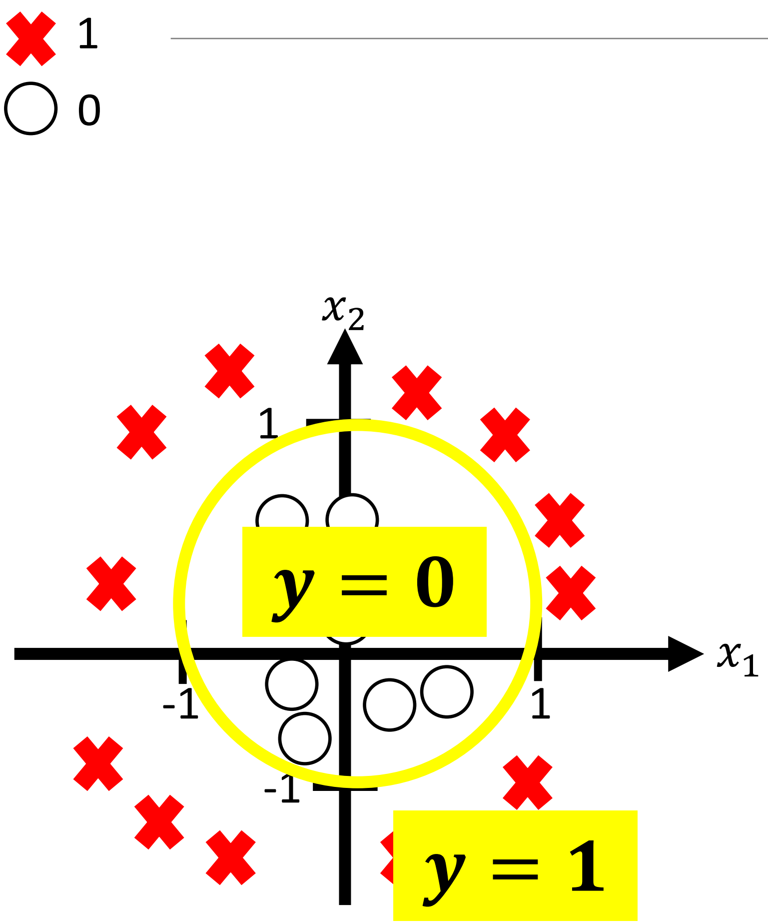






https://livebook.manning.com/concept/machine-learning/sigmoid-function
https://machinelearningmastery.com/plot-a-decision-surface-for-machine-learning/
https://stackoverflow.com/questions/29360872/fitting-3d-sigmoid-to-data

Side Note: Non-linear Decision Boundary



Suppose we use polynomial features…

$$h(x) = (w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2)$$

Suppose we are able to train a model an get the following weights…

$$W = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Predict $y = 1$, if $-1 + x_1^2 + x_2^2 \geq 0$

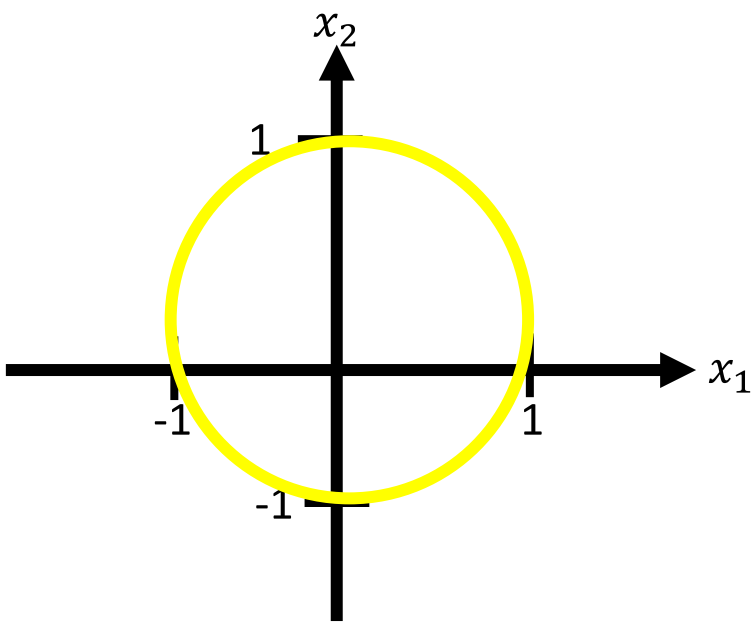
Predict $y = 1$, if $x_1^2 + x_2^2 \geq 1$

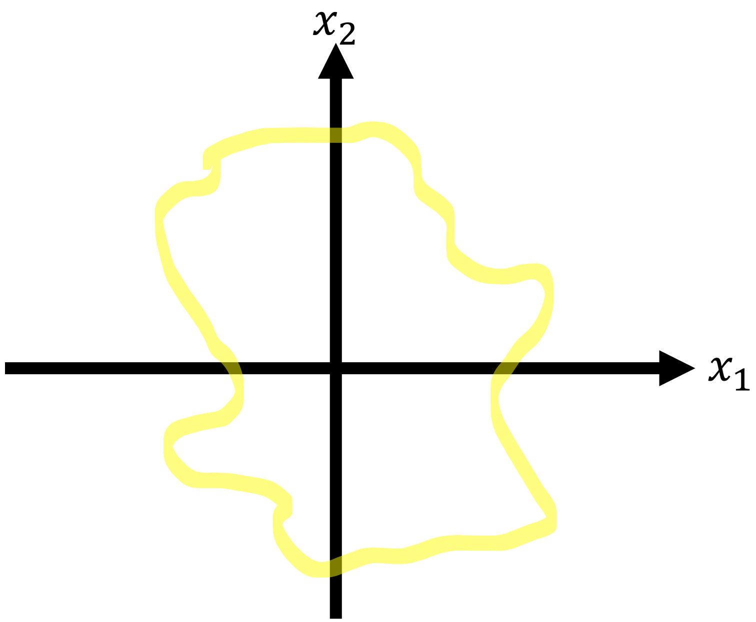
Equation of a circle

This means by controlling the weights, we can build complex decision boundaries!
Or simpler boundaries from complex features!

Side Note: Non-linear Decision Boundary

$$h(x) = \left(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1^2 x_2 + w_5 x_1^2 + x_2^2 + w_6 x_1^3 x_2 + \ \ldots\right)$$





This means by controlling the weights, we can build complex decision boundaries!
Or simpler boundaries from complex features!

Recall that more complex boundaries can cause overfitting (i.e., high variance)

Book Reading

- Murphy – Chapter 8
- Jurafsky – Chapter 5