Northeastern University
Khoury College of
Computer Sciences

Logistic Regression

DS 4400 | Machine Learning and Data Mining I

Zohair Shafi

Spring 2026

Monday | February 9, 2026

Updates

- Homework 1 Discussion
- Homework 3 Out - Due March 6th

Updates

- 17th Feb - Tuesday - Wanrou
- 1:30 PM - 3:00 PM | Location: Richards Hall 243
  - Linear algebra
    - Vectors
    - Matrices
    - Vector and Matrix operations
  - Probabilities
    - Bayes’ rule and conditional probability
    - Distributions
    - CDFs and PDFs

- 18th Feb - Wednesday - Zaiba
- 1:00 PM - 2:30 PM | Location: EL 311
  - Derivatives
    - Gradients
    - Derivatives of some common functions
    - Chain Rule, Product Rule, Quotient Rule

Today’s Outline

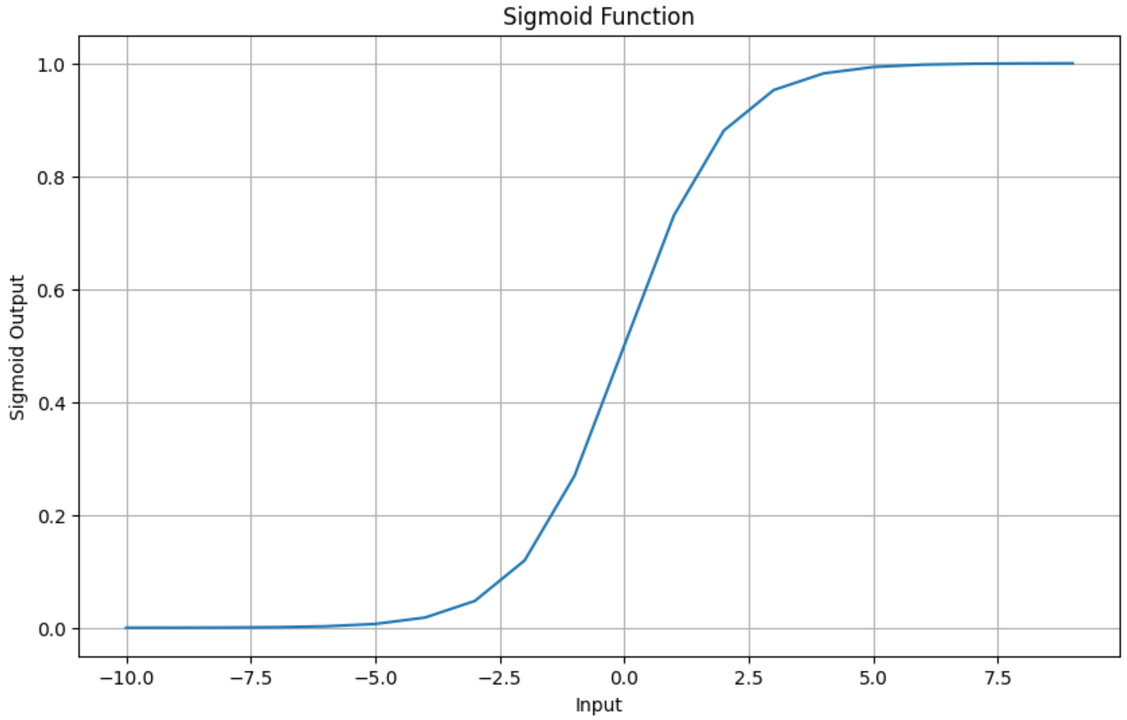
- Logistic Regression

Logistic Regression

Logistic Regression

- Despite its name, logistic regression is a **classification** algorithm.
- It models the probability of class membership using a logistic (sigmoid) function.

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

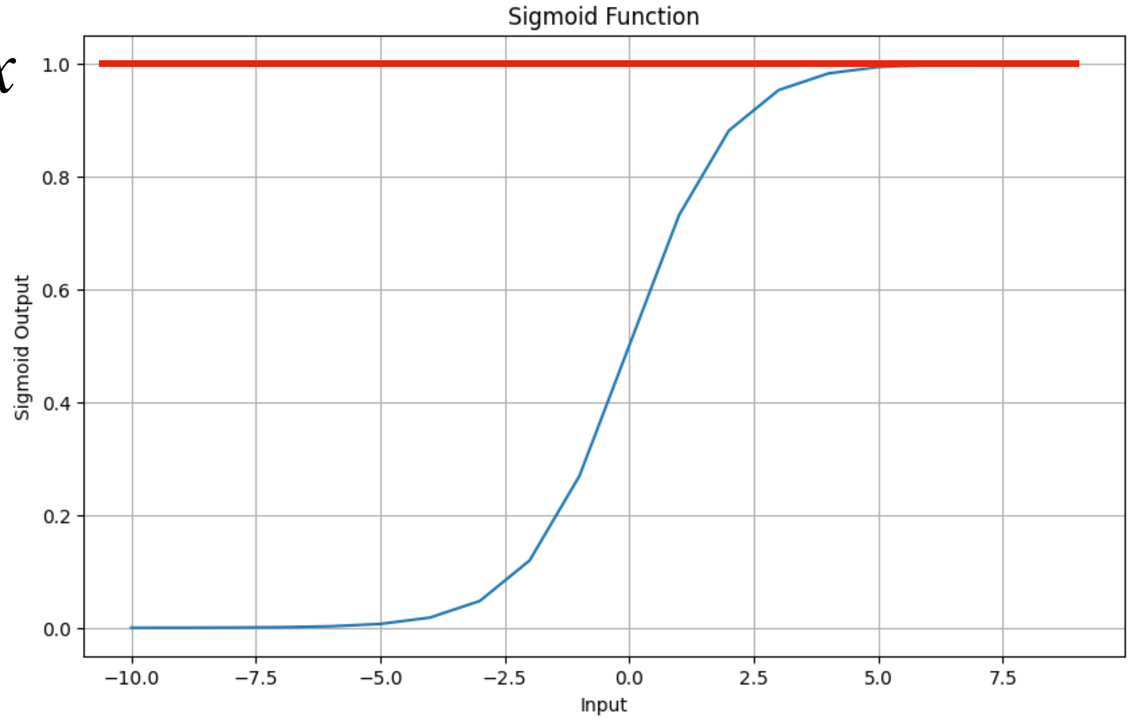


Logistic Regression

- Despite its name, logistic regression is a **classification** algorithm.
- It models the probability of class membership using a logistic (sigmoid) function.

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

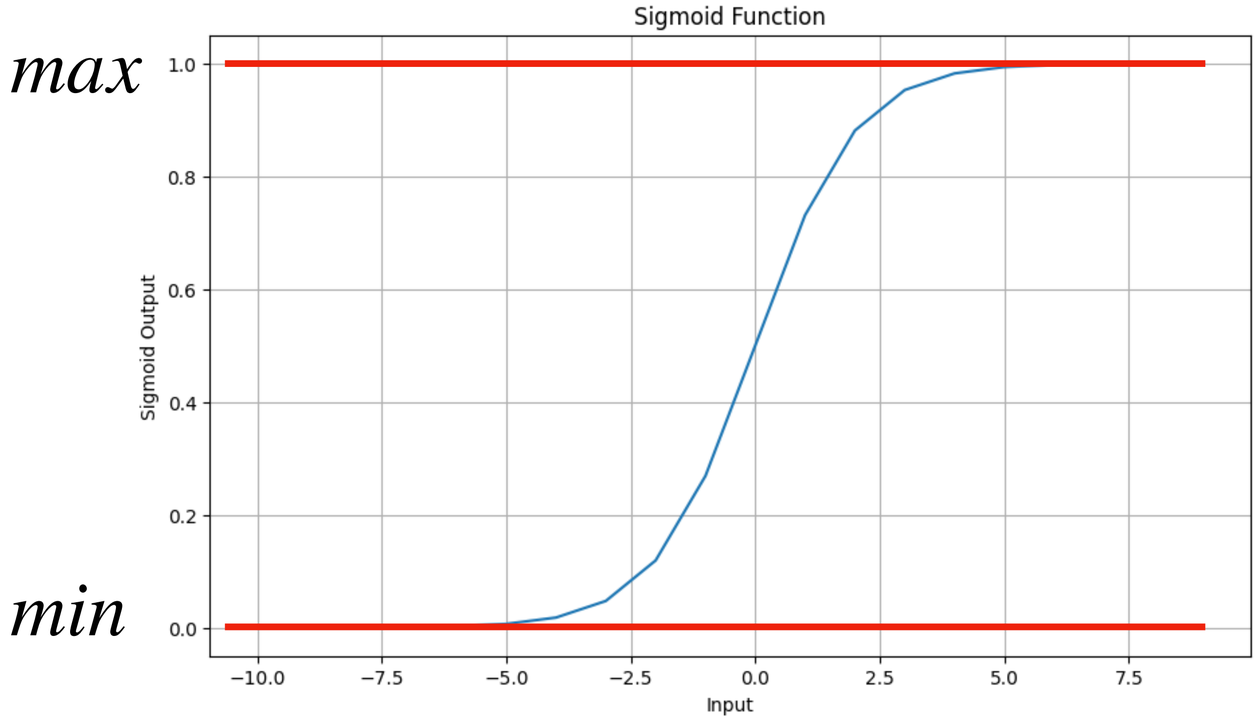
max



Logistic Regression

- Despite its name, logistic regression is a **classification** algorithm.
- It models the probability of class membership using a logistic (sigmoid) function.

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

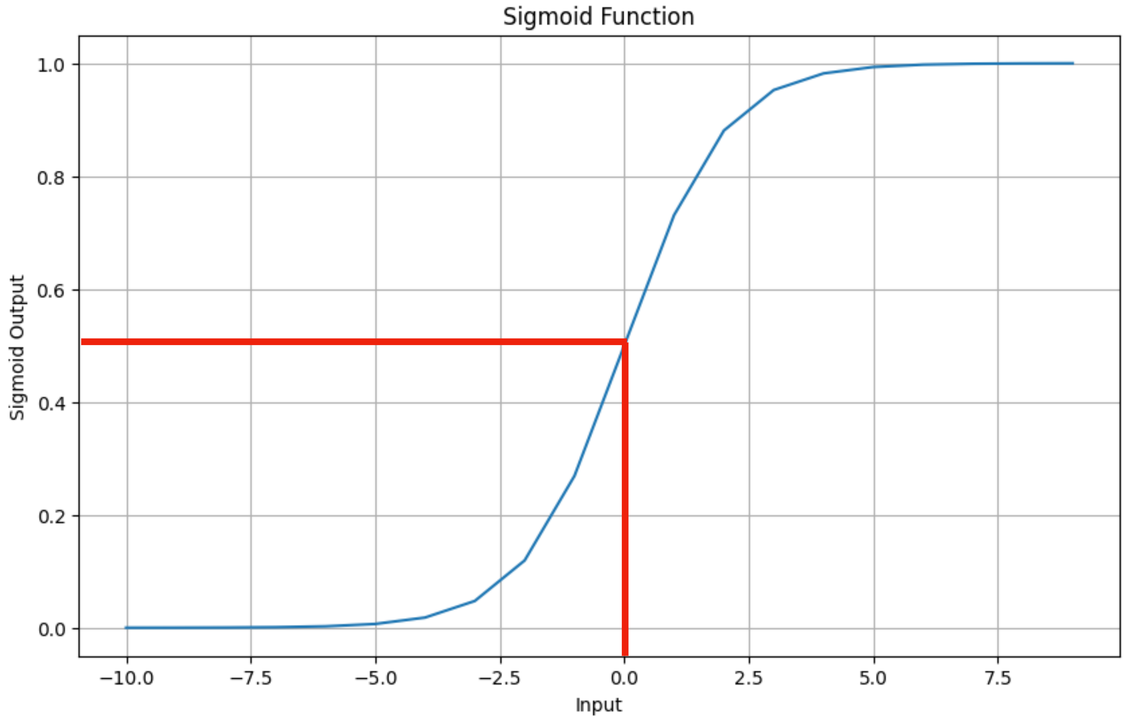


Logistic Regression

- Despite its name, logistic regression is a **classification** algorithm.
- It models the probability of class membership using a logistic (sigmoid) function.

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

$$mid = 0.5$$

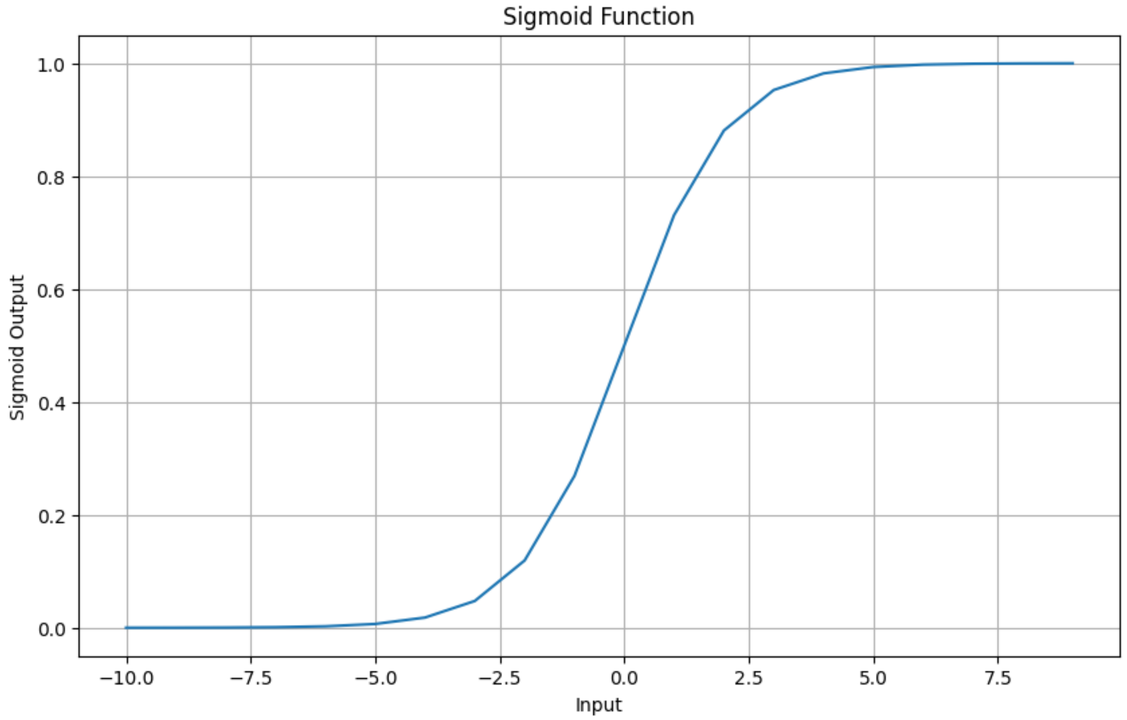


Logistic Regression

- Despite its name, logistic regression is a **classification** algorithm.
- It models the probability of class membership using a logistic (sigmoid) function.

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

- Linear regression predicts **unbounded** real values as $\hat{y} = \theta_0 + \theta_1 \cdot x$
- But we need probabilities in $[0, 1]$



Logistic Regression

- Wrap the linear regression equation in a Sigmoid function
- Logistics regression models the probability of the positive class

$$\mathbb{P}(Y = 1 \,|\, X = x) = \sigma(\theta_0 + \theta_1 \cdot x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 \cdot x)}}$$

The decision boundary is the hyperplane where $\mathbb{P}(Y = 1 \,|\, X = x) = 0.5$, which occurs when $\theta_0 + \theta_1 \cdot x = 0$

Logistic Regression

The decision boundary is the hyperplane where $\mathbb{P}(Y = 1 \mid X = x) = 0.5$, which occurs when $\theta_0 + \theta_1 \cdot x = 0$

Assume that threshold = 0.5

If $\theta_0 + \theta_1 \cdot x \geq 0$, classify as “positive class”

Why?

Logistic Regression

The decision boundary is the hyperplane where $\mathbb{P}(Y = 1 \mid X = x) = 0.5$, which occurs when $\theta_0 + \theta_1 \cdot x = 0$

Assume that threshold = 0.5

If $\theta_0 + \theta_1 \cdot x \geq 0$, classify as “positive class”

Why?

Because $\sigma(k \geq 0) \geq 0.5$

Logistic Regression

The decision boundary is the hyperplane where $\mathbb{P}(Y = 1 \mid X = x) = 0.5$, which occurs when $\theta_0 + \theta_1 \cdot x = 0$

Assume that threshold = 0.5

If $\theta_0 + \theta_1 \cdot x \geq 0$, classify as “positive class”

Why?

Because $\sigma(k \geq 0) \geq 0.5$

If $\theta_0 + \theta_1 \cdot x \leq 0$, classify as “negative class”

Why?

Logistic Regression

The decision boundary is the hyperplane where $\mathbb{P}(Y = 1 \mid X = x) = 0.5$, which occurs when $\theta_0 + \theta_1 \cdot x = 0$

Assume that threshold = 0.5

If $\theta_0 + \theta_1 \cdot x \geq 0$, classify as “positive class”

Why?

Because $\sigma(k \geq 0) \geq 0.5$

If $\theta_0 + \theta_1 \cdot x \leq 0$, classify as “negative class”

Why?

Because $\sigma(k \leq 0) \leq 0.5$

Logistic Regression

Model:

$$\hat{y} = \sigma(\theta_0 + \theta_1 \cdot x)$$

Loss:

$$\ell(\theta) = \frac{1}{m} \sum_{i=1}^{m} y^{(i)} log(\hat{y^{(i)}}) + (1 - y^{(i)}) log(1 - \hat{y^{(i)}})$$

Logistic Regression

How do we train this?

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a principled method for **estimating the parameters of a statistical model**.

Key Idea - Choose parameters that make the observed data **most probable**.

Logistic Regression

How do we train this?

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a principled method for **estimating the parameters of a statistical model**.

Key Idea - Choose parameters that make the observed data **most probable**.

Given some dataset D and a model with parameters θ

$$\hat{\theta}_{MLE} = arg\ max_{\theta} \mathbb{P}(D \mid \theta)$$

Logistic Regression

How do we train this?

Maximum Likelihood Estimation

Key Idea - Choose parameters that make the observed data most probable.

Given some dataset D and a model with parameters θ

$$\hat{\theta}_{MLE} = arg\ max_{\theta} \mathbb{P}(D \mid \theta)$$

Probability that we observe training dataset D, given that the model has parameters θ

Logistic Regression

How do we train this?

Maximum Likelihood Estimation

Key Idea - Choose parameters that make the observed data most probable.

Given some dataset D and a model with parameters θ

$$\hat{\theta}_{MLE} = arg\ max_{\theta} \mathbb{P}(D \mid \theta)$$

Find θ such that this probability is maximized

Logistic Regression

How do we train this?

Maximum Likelihood Estimation

Key Idea - Choose parameters that make the observed data most probable.

Given some dataset D and a model with parameters θ

$$\hat{\theta}_{MLE} = arg\ max_{\theta} \mathbb{P}(D \mid \theta)$$

Under what parameter values would we have been **most likely to observe exactly the data we did observe**?

Logistic Regression

How do we train this?

Maximum Likelihood Estimation

What we want to find:

$$\hat{\theta}_{MLE} = arg\ max_{\theta} \mathbb{P}(D \mid \theta)$$

Probability:

$\mathbb{P}(D \mid \theta)$ - Given **fixed parameters** θ**,**
what is the probability of observing data D?
This is a function of D with θ fixed.

Logistic Regression

How do we train this?

Maximum Likelihood Estimation

What we want to find:

$$\hat{\theta}_{MLE} = arg\ max_{\theta} \mathbb{P}(D \mid \theta)$$

Probability:

$\mathbb{P}(D \mid \theta)$ - Given **fixed parameters** θ**,**
what is the probability of observing data D?
This is a function of D with θ fixed.

Likelihood:

$L(\theta \mid D) = \mathbb{P}(D \mid \theta)$ **-** Given fixed observed data D, how **likely** are different parameter values θ?
This is a function of θ with D fixed.

Logistic Regression

Probability vs Likelihood

Coin Flips

Suppose you flip a coin **10 times** and get **7 heads**.

Logistic Regression

Probability vs Likelihood

Coin Flips

Suppose you flip a coin **10 times** and get **7 heads**.

Probability Perspective: If $\theta = 0.5$ (fair coin), what's $\mathbb{P}$(7 heads in 10 flips $\mid \theta$)?

Answer: $\mathbb{P}(X = 7 \mid \theta = 0.5) = \binom{10}{7} \cdot 0.5^7 \cdot (1 - 0.5)^{10-7} \approx 0.117$

Logistic Regression

Probability vs Likelihood

Coin Flips

Suppose you flip a coin **10 times** and get **7 heads**.

Probability Perspective: If $\theta = 0.5$ (fair coin), what's $\mathbb{P}$(7 heads in 10 flips $\mid \theta$)?

Answer: $\mathbb{P}(X = 7 \mid \theta = 0.5) = \binom{10}{7} \cdot 0.5^7 \cdot (1 - 0.5)^{10-7} \approx 0.117$

Likelihood Perspective: Given we observed 7 heads, which θ value makes this outcome most plausible?

Logistic Regression

Probability vs Likelihood

Coin Flips

Suppose you flip a coin **10 times** and get **7 heads**.

Probability Perspective: If $\theta = 0.5$ (fair coin), what's $\mathbb{P}$(7 heads in 10 flips $\mid \theta$)?

Answer: $\mathbb{P}(X = 7 \mid \theta = 0.5) = \binom{10}{7} \cdot 0.5^7 \cdot (1 - 0.5)^{10-7} \approx 0.117$

Likelihood Perspective: Given we observed 7 heads, which θ value makes this outcome most plausible?

$$L(\theta = 0.5 \mid X = 7) = 0.117$$

Logistic Regression

Probability vs Likelihood

Coin Flips

Suppose you flip a coin **10 times** and get **7 heads**.

Probability Perspective: If $\theta = 0.5$ (fair coin), what's $\mathbb{P}$(7 heads in 10 flips $\mid \theta$)?

Answer: $\mathbb{P}(X = 7 \mid \theta = 0.5) = \binom{10}{7} \cdot 0.5^7 \cdot (1 - 0.5)^{10-7} \approx 0.117$

Likelihood Perspective: Given we observed 7 heads, which θ value makes this outcome most plausible?

$$L(\theta = 0.5 \mid X = 7) = 0.117$$

$$L(\theta = 0.7 \mid X = 7) = 0.267 \text{ (higher)}$$

$$L(\theta = 0.3 \mid X = 7) = 0.009 \text{ (lower)}$$

Logistic Regression

Logistic Regression

Logistic Regression

Logistic Regression

Likelihood Function

For **independent** observations (rows of data) $D = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$, the likelihood is the **product** of individual probabilities

$$L(\theta \,|\, D) = \mathbb{P}(D \,|\, \theta) = \Pi_{i=1}^{m} \; \mathbb{P}(x^{(i)} \,|\, \theta)$$

Logistic Regression

Likelihood Function

For **independent** observations (rows of data) $D = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$, the likelihood is the **product** of individual probabilities

$$L(\theta \,|\, D) = \mathbb{P}(D \,|\, \theta) = \Pi_{i=1}^{m} \; \mathbb{P}(x^{(i)} \,|\, \theta)$$

But, products are numerically **unstable** and difficult to differentiate

So, we take log on both sides to convert products to sums

Logistic Regression

Likelihood Function

For **independent** observations (rows of data) $D = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$, the likelihood is the **product** of individual probabilities

$$L(\theta \mid D) = \mathbb{P}(D \mid \theta) = \Pi_{i=1}^{m} \ \mathbb{P}(x^{(i)} \mid \theta)$$

$$log(L(\theta \mid D)) = \sum_{i=1}^{m} log(\mathbb{P}(x^{(i)}) \mid \theta)$$

Using properties of log:

$$log(a^b) = b \cdot log(a)$$
$$log(ab) = log(a) + log(b))$$

Logistic Regression

Likelihood Function

For **independent** observations (rows of data) $D = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$, the likelihood is the **product** of individual probabilities

$$L(\theta \mid D) = \mathbb{P}(D \mid \theta) = \Pi_{i=1}^{m} \mathbb{P}(x^{(i)} \mid \theta)$$

$$log(L(\theta \mid D)) = \sum_{i=1}^{m} log(\mathbb{P}(x^{(i)}) \mid \theta)$$

For logistic regression

Input Features: $x \in \mathbb{R}^m$

Binary Labels: $y \in \{0,1\}$

Training Data: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

Logistic Regression

$$\mathbb{P}(Y = 1 \mid X = x; \theta) = \sigma(\theta_0 + \theta_1 \cdot x)$$

Logistic Regression

$$\mathbb{P}(Y = 1 \mid X = x; \theta) = \sigma(\theta_0 + \theta_1 \cdot x)$$

Each label y_i follows a **Bernoulli Distribution** with parameter

$$p_i = \mathbb{P}(Y = 1 \mid x_i)$$

Logistic Regression

Quick Aside: Bernoulli Distribution

Bernoulli Distribution models a single binary outcome

Is $\mathbb{P}(X = success) = p$ and
$\mathbb{P}(X = failure) = q = (1 - p)$

Then probability mass function P is

$$P(X = x) = p^x \cdot (1 - p)^{1-x}$$

Logistic Regression

$$\mathbb{P}(Y = 1 \mid X = x; \theta) = \sigma(\theta_0 + \theta_1 \cdot x)$$

Each label y_i follows a **Bernoulli Distribution** with parameter

$$p_i = \mathbb{P}(Y = 1 \mid x_i)$$

$$\mathbb{P}(Y = y \mid X = x) = p^y(1 - p)^{1-y}$$

Logistic Regression

$$\mathbb{P}(Y = 1 \mid X = x; \theta) = \sigma(\theta_0 + \theta_1 \cdot x)$$

Each label y_i follows a **Bernoulli Distribution** with parameter

$$p_i = \mathbb{P}(Y = 1 \mid x_i)$$

$$\mathbb{P}(Y = y \mid X = x) = p^y(1 - p)^{1-y}$$

When $y = 1 \rightarrow p^1(1 - p)^0 = p$

When $y = 0 \rightarrow p^0(1 - p)^1 = (1 - p)$

Logistic Regression

$$\mathbb{P}(Y = 1 \mid X = x; \theta) = \sigma(\theta_0 + \theta_1 \cdot x)$$

Each label y_i follows a **Bernoulli Distribution** with parameter

$$p_i = \mathbb{P}(Y = 1 \mid x_i)$$

$$\mathbb{P}(Y = y \mid X = x) = p^y (1 - p)^{1-y}$$

When $y = 1 \rightarrow p^1 (1 - p)^0 = p$

When $y = 0 \rightarrow p^0 (1 - p)^1 = (1 - p)$

Logistic Regression

For a **single** observation $(x^{(i)}, y^{(i)})$

Probability of observing $y^{(i)}$ given you have seen input data $x^{(i)}$ and θ

$$\mathbb{P}(y^{(i)} \mid x^{(i)}; \theta) = p_i^{y^{(i)}}(1 - p_i)^{1-y^{(i)}}$$

Where $p_i = \sigma(\theta_0 + \theta_1 \cdot x)$

Logistic Regression

For the **entire dataset** $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots (x^{(m)}, y^{(m)})\}$

Assuming observations are **independent**

Likelihood is the product of all individual probabilities

$$L(\theta \mid D) = \Pi_{i=1}^{m} \mathbb{P}(y^{(i)} \mid x^{(i)}; \theta) = \Pi_{i=1}^{m} p_i^{y^{(i)}} (1 - p_i)^{1 - y^{(i)}}$$

Logistic Regression

For the **entire dataset** $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots (x^{(m)}, y^{(m)})\}$

Assuming observations are **independent**

Likelihood is the product of all individual probabilities

$$L(\theta \mid D) = \Pi_{i=1}^{m} \mathbb{P}(y^{(i)} \mid x^{(i)}; \theta) = \Pi_{i=1}^{m} p_i^{y^{(i)}} (1 - p_i)^{1 - y^{(i)}}$$

We want to **maximize** likelihood

$$\hat{\theta}_{MLE} = arg\ max_{\theta} \mathbb{P}(D \mid \theta)$$

Logistic Regression

$$L(\theta \mid D) = \Pi_{i=1}^{m} p_i^{y^{(i)}} (1 - p_i)^{1 - y^{(i)}}$$

Logistic Regression

$$L(\theta|D) = \Pi_{i=1}^{m} p_i^{y^{(i)}}(1-p_i)^{1-y^{(i)}}$$

$$log(L(\theta)) = log(\Pi_{i=1}^{m} p_i^{y^{(i)}}(1-p_i)^{1-y^{(i)}})$$

Using properties of log:

$log(a^b) = b \cdot log(a)$
$log(ab) = log(a) + log(b))$

$$log(L(\theta)) = \sum_{i=1}^{m} log(p_i^{y^{(i)}}(1-p_i)^{1-y^{(i)}})$$

$$log(L(\theta)) = \sum_{i=1}^{m} y^{(i)} log(p_i) + (1-y^{(i)}) log(1-p_i)$$

Logistic Regression

$$L(\theta \mid D) = \Pi_{i=1}^{m} p_i^{y^{(i)}} (1 - p_i)^{1 - y^{(i)}}$$

$$log(L(\theta)) = log(\Pi_{i=1}^{m} p_i^{y^{(i)}} (1 - p_i)^{1 - y^{(i)}})$$

Using properties of log:

$$log(a^b) = b \cdot log(a)$$
$$log(ab) = log(a) + log(b))$$

$$log(L(\theta)) = \sum_{i=1}^{m} log(p_i^{y^{(i)}} (1 - p_i)^{1 - y^{(i)}})$$

$$log(L(\theta)) = \sum_{i=1}^{m} y^{(i)} log(p_i) + (1 - y^{(i)}) log(1 - p_i)$$

This is called the **log-likelihood** function for logistic regression

Logistic Regression

$$log(L(\theta)) = \sum_{i=1}^{m} y^{(i)} log(p_i) + (1 - y^{(i)}) log(1 - p_i)$$

This is called the **log-likelihood** function for logistic regression

Remember we want to **maximize** likelihood

But when we deal with “loss” functions and gradient descent, we want to **minimize** the loss

Logistic Regression

$$\ell(\theta) = -\sum_{i=1}^{m} y^{(i)} log(p_i) + (1 - y^{(i)}) log(1 - p_i)$$

Solution: Minimize **negative** likelihood

Logistic Regression

$$\ell(\theta) = -\sum_{i=1}^{m} y^{(i)} log(p_i) + (1 - y^{(i)}) log(1 - p_i)$$

Solution: Minimize **negative** likelihood

Remember that p_i is the predicted output where

$$p_i = \sigma(\theta_0 + \theta_1 \cdot x)$$

Logistic Regression

$$\ell(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} log(\hat{y}^{(i)}) + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})$$

Binary Cross Entropy Loss

Logistic Regression

$$\ell(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} log(\hat{y}^{(i)}) + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})$$

When $y^{(i)} = 1$, i.e., actual positive

$$\ell(\theta) = - log(\hat{y}^{(i)})$$

When $y^{(i)} = 0$, i.e., actual negative

$$\ell(\theta) = - log(1 - \hat{y}^{(i)})$$

Logistic Regression

$$\ell(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} log(\hat{y}^{(i)}) + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})$$

When $y^{(i)} = 1$, i.e., actual positive

$$\ell(\theta) = -\, log(\hat{y}^{(i)})$$

If $\hat{y}^{(i)} = 1$, Loss = 0

If $\hat{y}^{(i)} = 0$, Loss = $+\infty$

When $y^{(i)} = 0$, i.e., actual negative

$$\ell(\theta) = -\, log(1 - \hat{y}^{(i)})$$

If $\hat{y}^{(i)} = 0$, Loss = 0

If $\hat{y}^{(i)} = 1$, Loss = $+\infty$

Logistic Regression

Finding θ

$$\ell(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} log(\hat{y}^{(i)}) + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})$$

Find partial derivative

To simplify, lets find the derivative for a **single** sample

Logistic Regression

Finding θ

$$\ell(\theta) = ylog(\hat{y}) + (1-y)log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = \theta_0 + \theta_1 x$$

Want to find $\frac{\partial \ell}{\partial \theta}$

Using Chain Rule

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial \theta}$$

Logistic Regression

Finding θ

Summing over all samples

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} \cdot (\hat{y}^{(i)} - y^{(i)})$$

In matrix form

$$\nabla_{\theta}(\ell(\theta)) = \frac{1}{m} X^{T}(\hat{Y} - Y)$$

Logistic Regression

Summary

Model:

$$\hat{y} = \sigma(\theta_0 + \theta_1 x)$$

Loss:

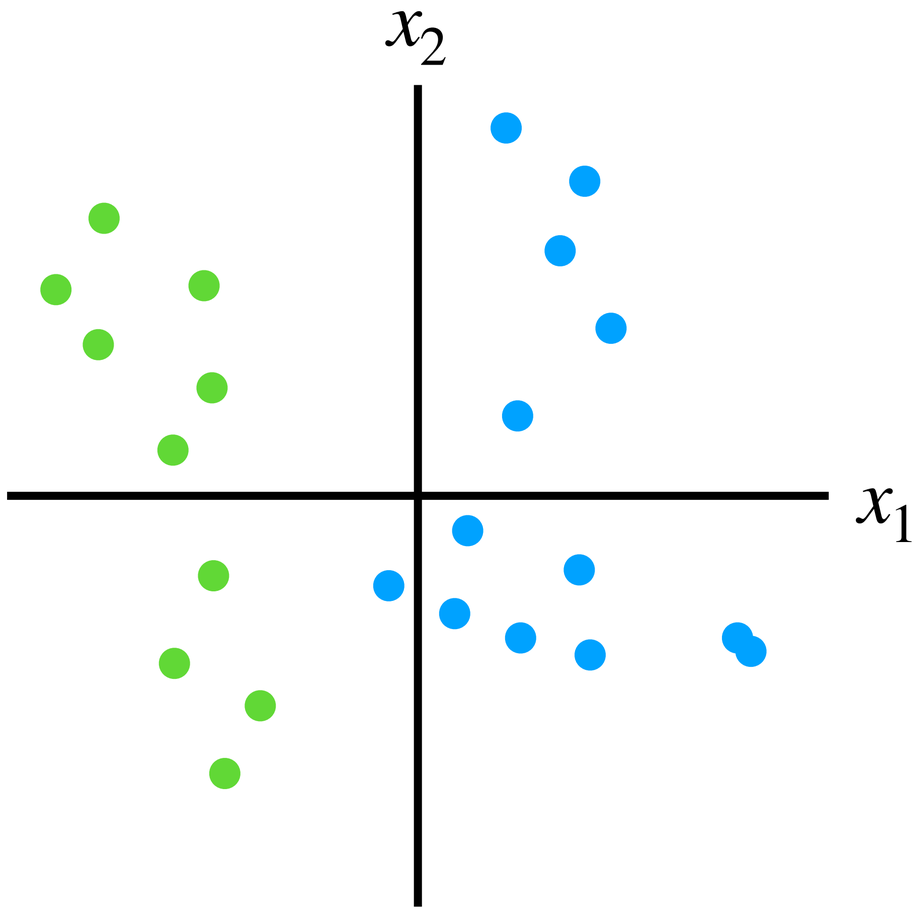
$$\ell(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} log(\hat{y}^{(i)}) + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})$$

Gradient:

$$\nabla_\theta(\ell(\theta)) = \frac{1}{m} X^T(\hat{Y} - Y)$$

Logistic Regression

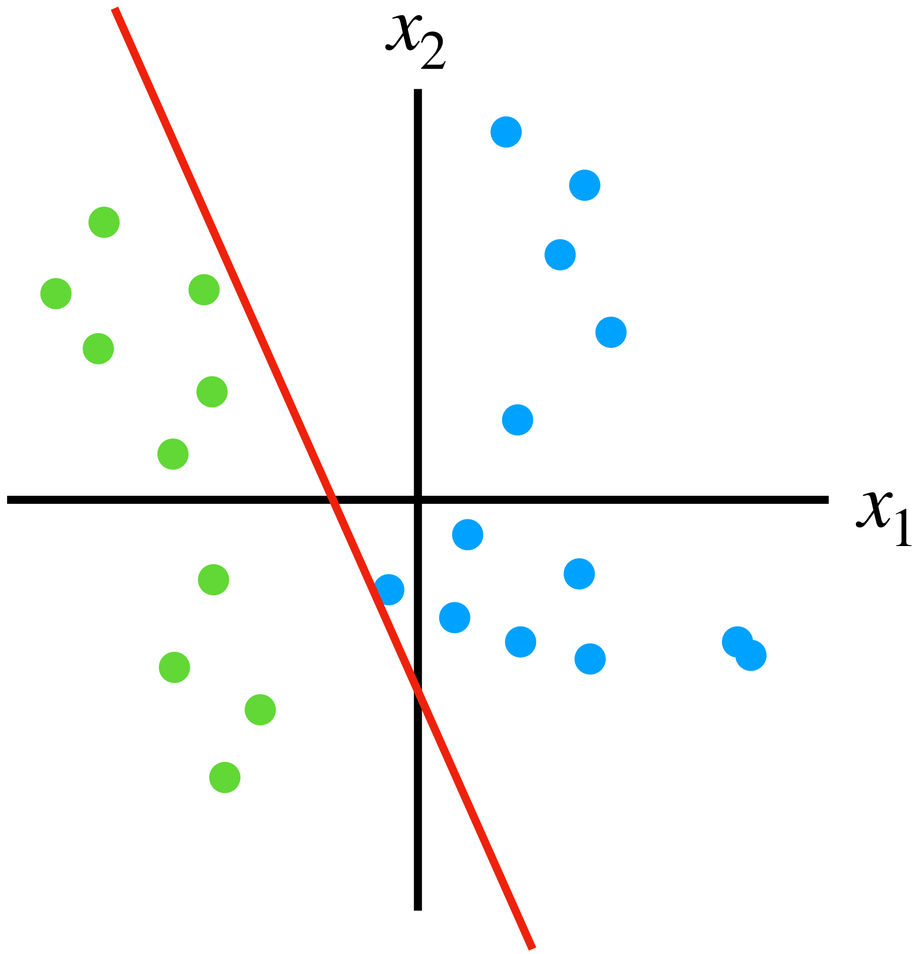
Summary



Logistic Regression

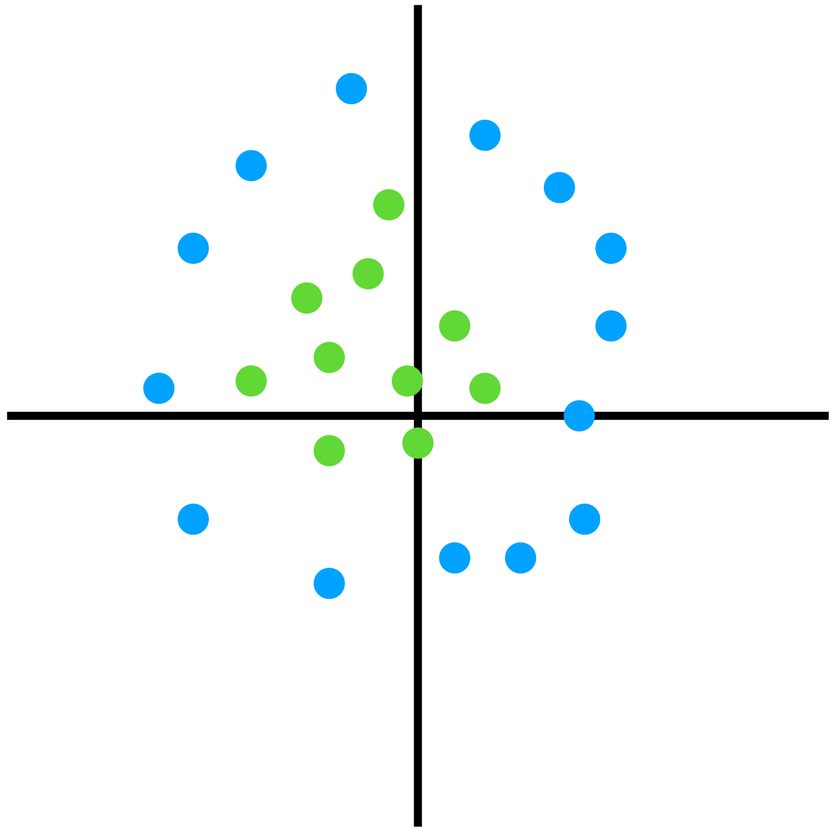
Summary

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$



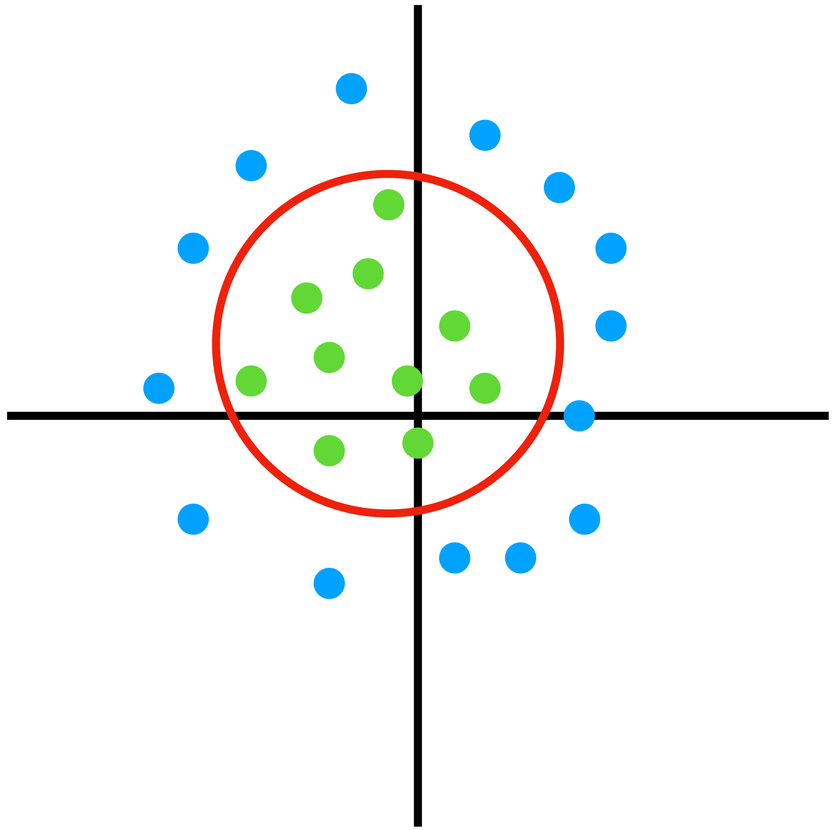
Logistic Regression

Summary



Logistic Regression

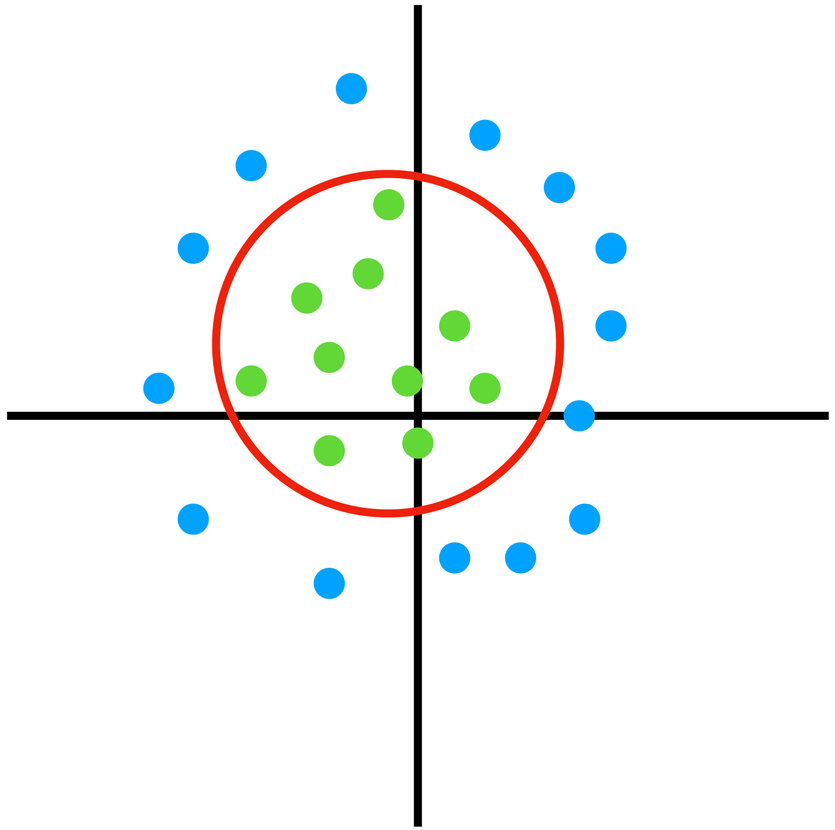
Summary



Logistic Regression

Summary

$$x_1^2 + x_2^2 = r^2$$

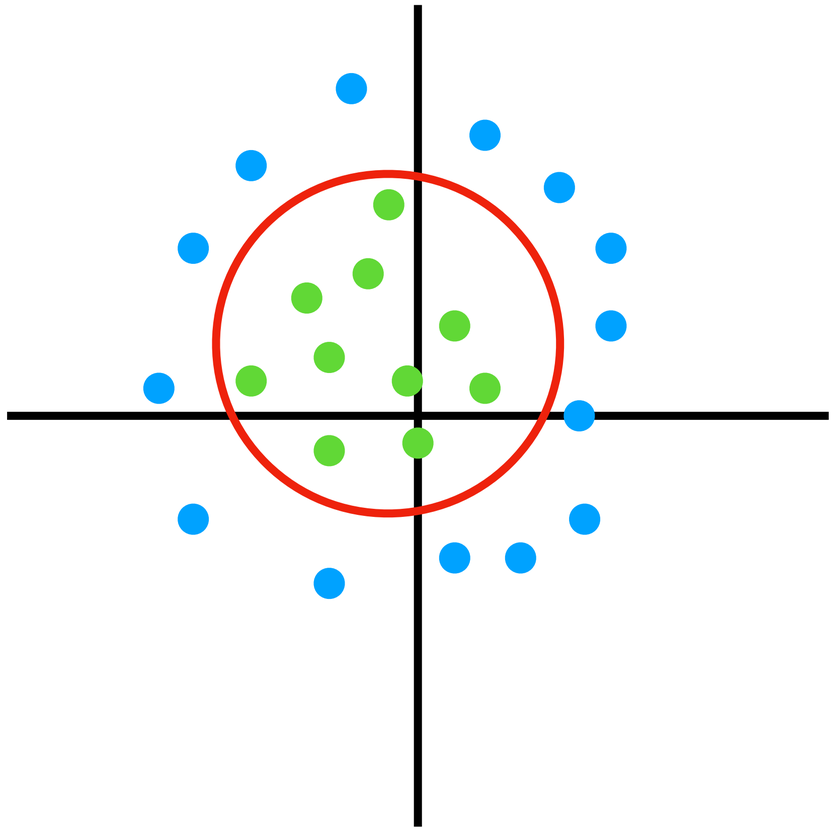


Logistic Regression

Summary

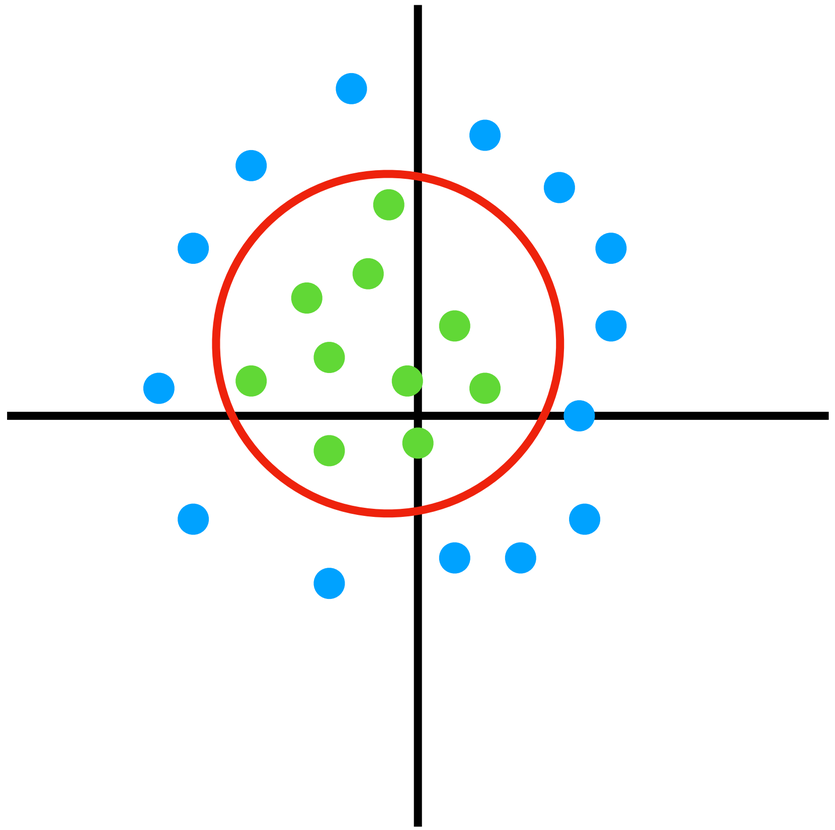
$$x_1^2 + x_2^2 = r^2$$

$$\theta_1^2(x_1^2 + x_2^2) = \theta_0^2$$



Logistic Regression

Summary



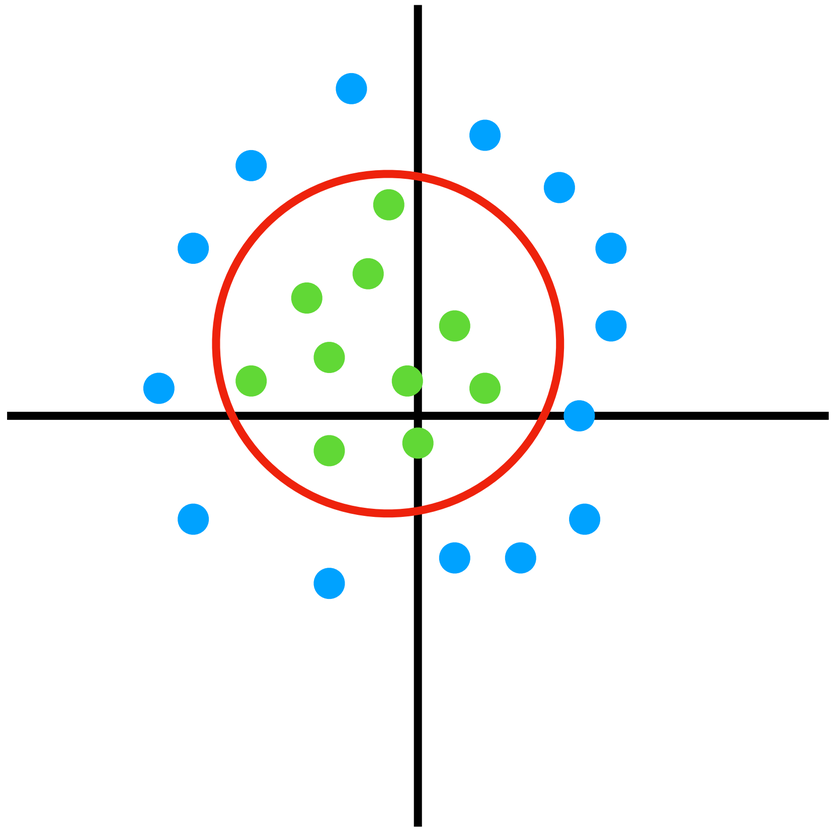
$$x_1^2 + x_2^2 = r^2$$

$$\theta_1^2(x_1^2 + x_2^2) = \theta_0^2$$

$$\sqrt{\theta_1^2(x_1^2 + x_2^2)} = \sqrt{\theta_0^2}$$

Logistic Regression

Summary



$$x_1^2 + x_2^2 = r^2$$

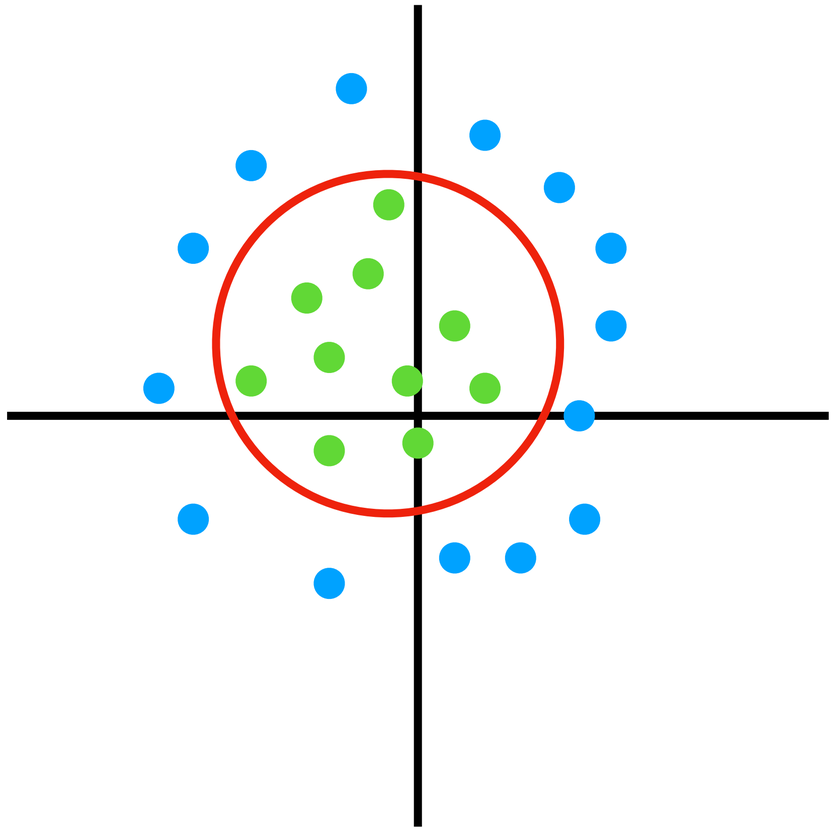
$$\theta_1^2(x_1^2 + x_2^2) = \theta_0^2$$

$$\sqrt{\theta_1^2(x_1^2 + x_2^2)} = \sqrt{\theta_0^2}$$

$$\theta_1\sqrt{(x_1^2 + x_2^2)} = \theta_0$$

Logistic Regression

Summary



$$x_1^2 + x_2^2 = r^2$$

$$\theta_1^2(x_1^2 + x_2^2) = \theta_0^2$$

$$\sqrt{\theta_1^2(x_1^2 + x_2^2)} = \sqrt{\theta_0^2}$$

$$\theta_1\sqrt{(x_1^2 + x_2^2)} = \theta_0$$

$$\hat{y} = \theta_1\sqrt{(x_1^2 + x_2^2)} - \theta_0$$

Next Class

- More classification algorithms