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 ~~6~~9th

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

① Model: $\hat{y} : \sigma(\underbrace{\theta_0 + \theta_1 x}_{\text{Unbounded}})$

$\sigma \rightarrow$ 0 - 1

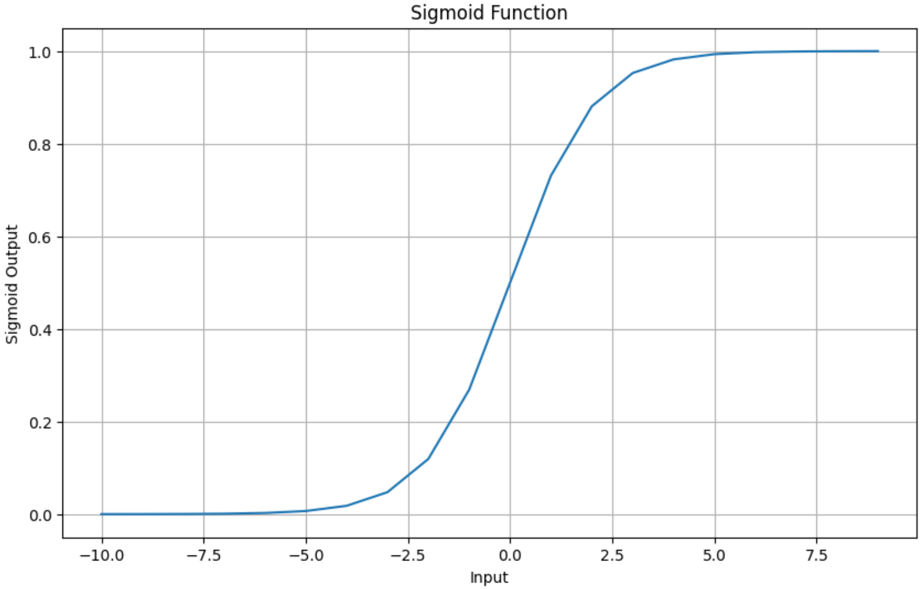
② Loss Function: Negative log likelihood loss.

③ Derivative. $\rightarrow$ 0 $\rightarrow$ solve for theta.

Logistic Regression

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

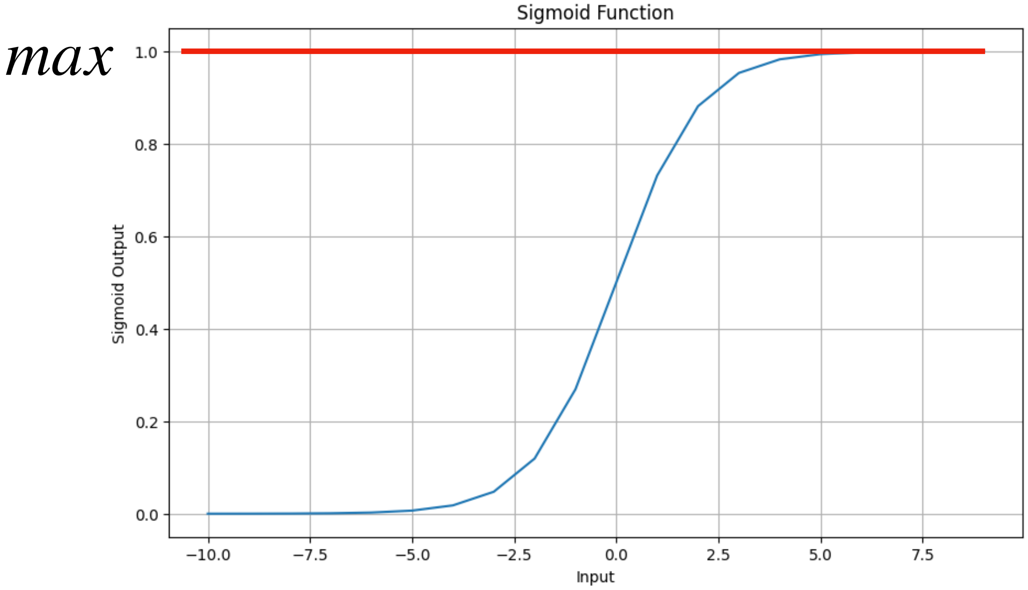
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Logistic Regression

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

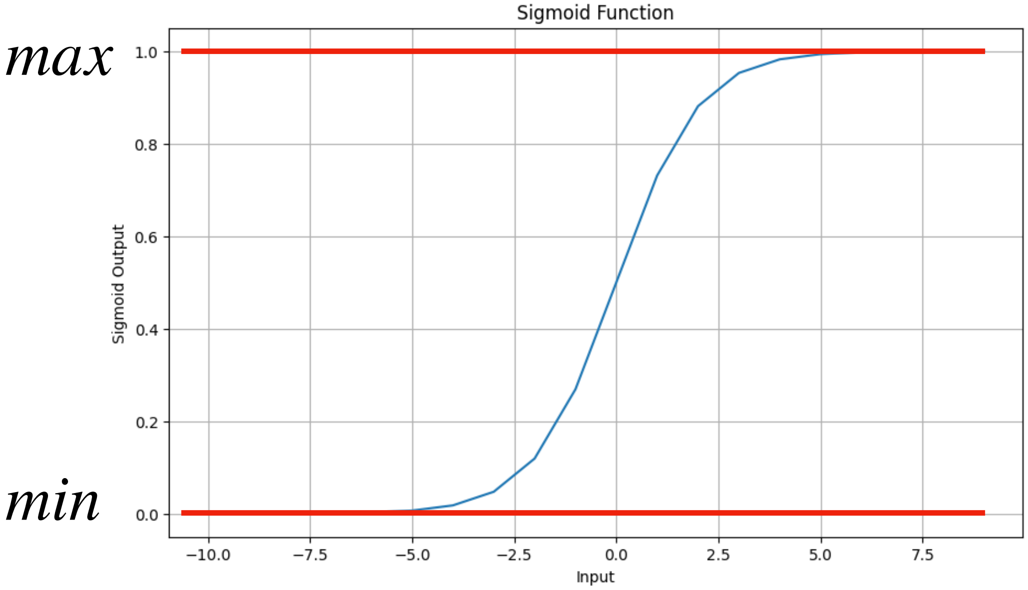
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Logistic Regression

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

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



Logistic Regression

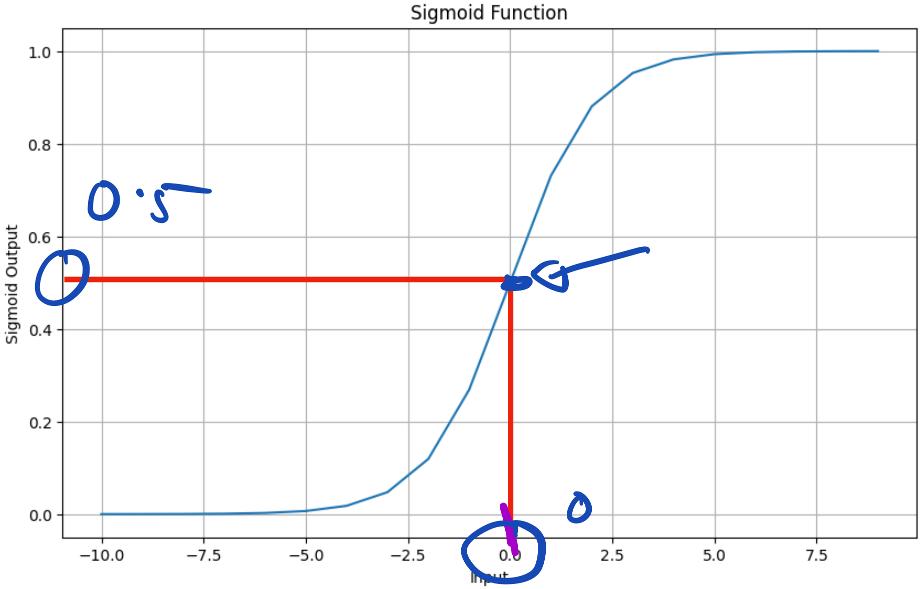
- Despite its name, logistic regression is a **classification** algorithm.
- It models the probability of class membership using a logistic (sigmoid) function.

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

$= 0.5$

$$\hat{y} = \sigma(\underbrace{\theta_0 + \theta_1 x_1 + \theta_2 x_2}_{=0}).$$

$$mid = 0.5$$

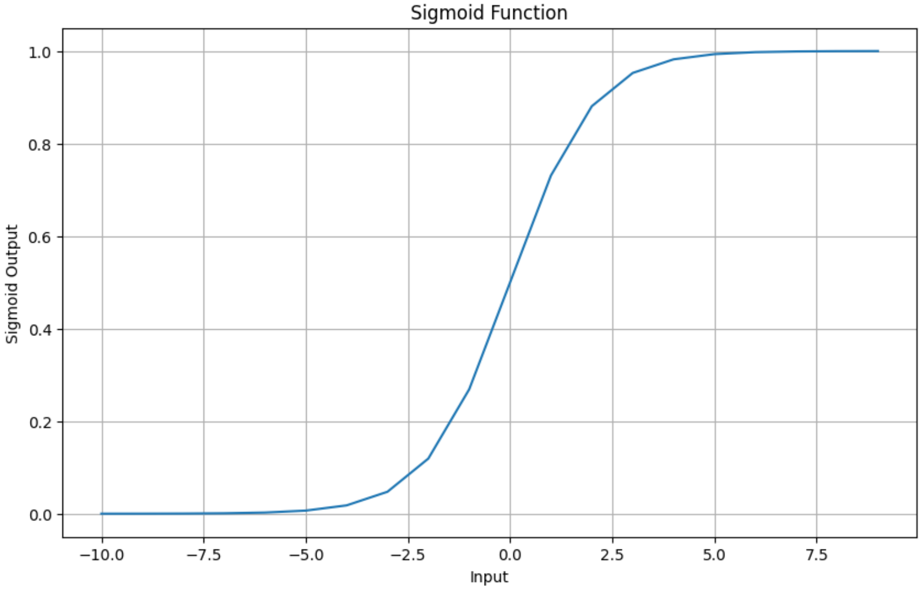


Logistic Regression

- 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

given

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

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

Logistic Regression

The decision boundary is the hyperplane where $\mathbb{P}(Y = 1 \,|\, X = x) = 0.5$, 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 = $\boxed{0.5}$

If $\underbrace{\theta_0 + \theta_1 \cdot x \geq}$ 0, classify as "positive class"

Why?

Because $\underbrace{\sigma(k \geq 0) \geq 0.5}$ $\longrightarrow$ +ve

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 $\underbrace{\boxed{\theta_0 + \theta_1 \cdot x}}_{k} \geq 0$, 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:

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

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 $\boxed{\theta}$

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

fixed

$\boxed{\theta_A}$

θ_B

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

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

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

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

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:

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

Probability:

$\mathbb{P}(D\,|\,\theta)$ - 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:

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

Probability:

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

Likelihood:

$L(\theta\,|\,\boxed{D}) = \mathbb{P}(\boxed{D}\,|\,\theta)$ - 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**.

Binomial Theorem:

$$P(7 \text{ heads} \mid 10 \text{ flips}) = P(\text{head})^{7} \cdot (1 - P(\text{head}))^{10-7} \cdot \binom{10}{7}$$

Logistic Regression

Probability vs Likelihood

D = 7 heads out of 10 trials.

Coin Flips

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

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

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

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 $|\,\theta$)?

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

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 $| \theta$)?

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

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

$$L(\theta = 0.5 \,|\, X = 7) = 0.117$$

Logistic Regression

Probability vs Likelihood

$\rightarrow D \Rightarrow 7$ heads / 10 trials.

$\mathbb{P}(D|\theta) = 0.117$

Coin Flips

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

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

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

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

$$\left.\begin{aligned} &L(\theta = 0.5 \mid X = 7) = 0.117 \\ &L(\theta = 0.7 \mid X = 7) = 0.267 \text{ (higher)} \\ &L(\theta = 0.3 \mid X = 7) = 0.009 \text{ (lower)} \end{aligned}\right\}$$

$\mathcal{L}(\theta \mid (x^{(1)}, y^{(1)}) \ldots (x^{(m)}, y^{(m)}))$

Logistic Regression

At this point, threshold doesn't matter.

① Model $\rightarrow \hat{y} = \sigma(\theta_0 + \theta_1 x) \Rightarrow \mathbb{P}(y=1 \mid x, \theta) \rightarrow \mathbf{p}$

② Bernoulli distribution $\rightarrow$ $\mathbb{P}(y) = \begin{cases} p & \text{if } y=1 \\ 1-p & \text{if } y=0 \end{cases}$ $\rightarrow$ step function piece wise

$$\boxed{p^{y} \cdot (1-p)^{(1-y)}}$$

If $y = 1 \rightarrow p \cdot (1-p)^{(0)} = p$

$y = 0 \rightarrow p^{0} \cdot (1-p)^{(1)} = (1-p)$

$D = \{ (x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)}) \}$

$x : \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix}$

③ MLE

Product

$$\underbrace{\mathcal{L}(\theta \mid D)}_{\text{Maximise}} = \mathbb{P}(\boxed{D} \mid \theta) = \prod \mathbb{P}(y_i \mid x_i, \theta)$$

Logistic Regression

$$\textcircled{1}\ \log(ab) = \log a + \log b$$

$$\textcircled{2}\ \log(a^b) = b \cdot \log(a)$$

$$\mathcal{L}(\theta|D) = \mathbb{P}(D|\theta) = \prod_{i=1}^{m} \mathbb{P}(y_i | x_i, \theta)$$

$$p = \mathbb{P}(y=1 | x, \theta)$$

$$\log(\mathcal{L}(\theta|D)) = \sum_{i=1}^{m} \log(\mathbb{P}(y_i | x_i, \theta))$$

$$= \sum_{i=1}^{m} \log\left(p_i^{y_i} \cdot (1-p_i)^{(1-y_i)}\right)$$

$$\log(\mathcal{L}(\theta|D)) = \sum_{i=1}^{m} y_i \log p_i + (1-y_i)\log(1-p_i) \rightarrow \text{log likelihood}$$

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

↳ Negative log likelihood.

Binary Cross Entropy.

$$\max f(x) = \min -f(x)$$

Logistic Regression

$$\mathcal{L}(\theta) = -y \log p + (1-y)\log(1-p)]$$

$$p = \sigma(z)$$

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

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_1} = \underbrace{\frac{\partial \mathcal{L}(\theta)}{\partial p}}_{\frac{-y}{p} + \frac{(1-y)}{(1-p)}} \cdot \underbrace{\frac{\partial p}{\partial z}}_{\sigma(z)\cdot(1-\sigma(z))} \cdot \underbrace{\frac{\partial z}{\partial \theta}}$$

$$\frac{-y}{p} + \frac{(1-y)}{(1-p)} \cdot \sigma(z)\cdot(1-\sigma(z))$$

if $\theta_0 \rightarrow 1$

$\theta_1 \rightarrow x$

① $\boxed{\frac{\partial(\log x)}{\partial x} = \frac{1}{x}}$

② $\frac{\partial(\sigma(x))}{\partial x} = \sigma(x)\cdot(1-\sigma(x))$

Logistic Regression

$$L(\theta) = -[\, y \log p + (1-y)\log(1-p)\,]$$

$$p = \sigma(z)$$

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

① $\boxed{\dfrac{\partial(\log x)}{\partial x} = \dfrac{1}{x}}$

② $\dfrac{\partial(\sigma(x))}{\partial x} = \sigma(x)\cdot(1-\sigma(x))$

$$\underbrace{\frac{-y}{p} + \frac{(1-y)}{(1-p)}}_{\frac{-y(1-p)+p(1-y)}{p(1-p)}} \cdot \underbrace{\sigma(z)\cdot(1-\sigma(z))}_{p\cdot(1-p)} \cdot x$$

$$\boxed{\frac{p-y}{p(1-p)}}$$

$$\frac{p-y}{\cancel{p(1-p)}} \cdot \cancel{p(1-p)} \cdot x$$

$$\frac{\partial L(\theta)}{\partial \theta_1} = x\cdot(p-y)$$

$$\frac{\partial L(\theta)}{\partial \theta_0} = (p-y)$$

Logistic Regression

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left[\underbrace{y\log p} + \underbrace{(1-y)}\,\underbrace{\log_e(1-p)}\right]$$

$$Derivative - (p-y)\cdot x$$

$$\underbrace{(\sigma(\theta_0+\theta_1 x) - y)}\cdot x = 0$$

1000

10 1 0 2

$$\left[\underbrace{\sigma(\boxed{\theta_0+\theta_1 x})}_{0-1} - \underbrace{y}_{0-1}\right]$$

$$y=1 \quad p=0 \rightarrow \log 0 \rightarrow \infty$$

$$p=1 \rightarrow \log 1 \rightarrow 0$$

$$y=0 \rightarrow p=0 \rightarrow \log 1 \rightarrow 0$$

$$p=1 \rightarrow \log 0 \rightarrow \infty$$

$$\sigma(\text{input}) = 0.7$$

0.6

0.5

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

$$\underbrace{L(\theta \,|\, D)} = \underbrace{\mathbb{P}(D \,|\, \theta)} = \underbrace{\Pi_{i=1}^{m} \; \mathbb{P}(x^{(i)} \,|\, \theta)}_{\text{assuming independence.}}$$

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

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

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

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

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

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

For logistic regression

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

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

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

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

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

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

$$\begin{cases} p & \text{if } y=1 \\ 1-p & \text{if } y=0 \end{cases}$$

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

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

Logistic Regression

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

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

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

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

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

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

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

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

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

Using properties of log:

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

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

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

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

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

Solution: Minimize **negative** likelihood

Remember that p_i is the predicted output where

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

Logistic Regression

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

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

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

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

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

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

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

$$\hat{y} = \sigma(z) \quad - 2$$

$$z = \theta_0 + \theta_1 x \quad - 3$$

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

Using Chain Rule

$$\frac{\partial \ell}{\partial \theta} = \underbrace{\frac{\partial \ell}{\partial \hat{y}}}_{1} \cdot \underbrace{\frac{\partial \hat{y}}{\partial z}}_{2} \cdot \underbrace{\frac{\partial z}{\partial \theta}}_{3}$$

Logistic Regression

Finding θ

Summing over all samples

$p - y$

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

\# examples

$x \in \mathbb{R}^{\boxed{m} \times \boxed{n}}$ — \# features.

$p, y \in \mathbb{R}^{m \times 1}$

In matrix form

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

$X^T(p - y]$

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

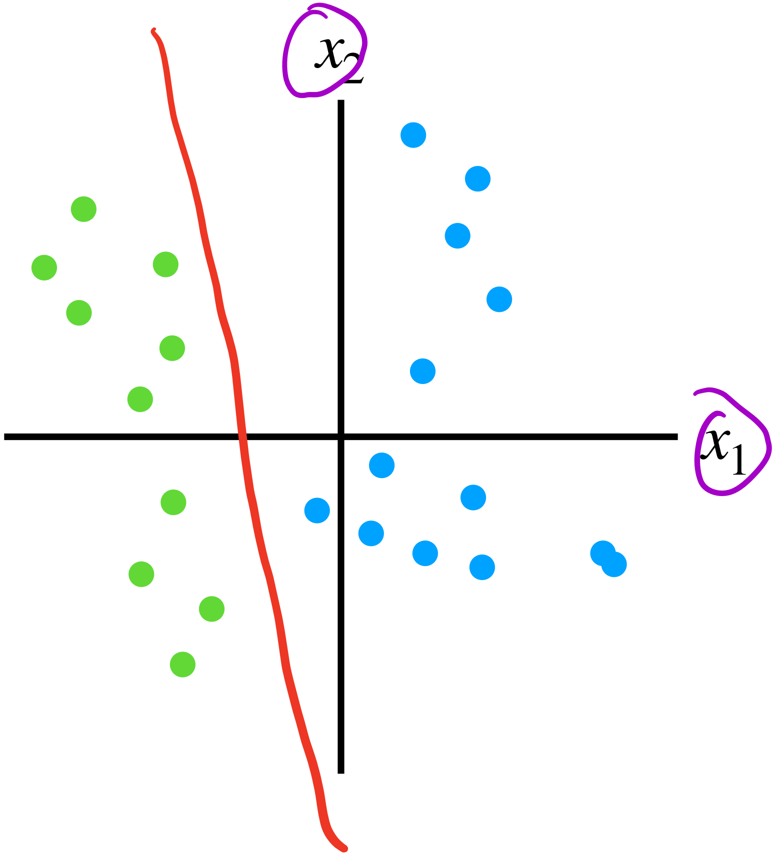
Why not MSE?

1. Not convex for logistic regression
2. Gradients are small

Logistic Regression

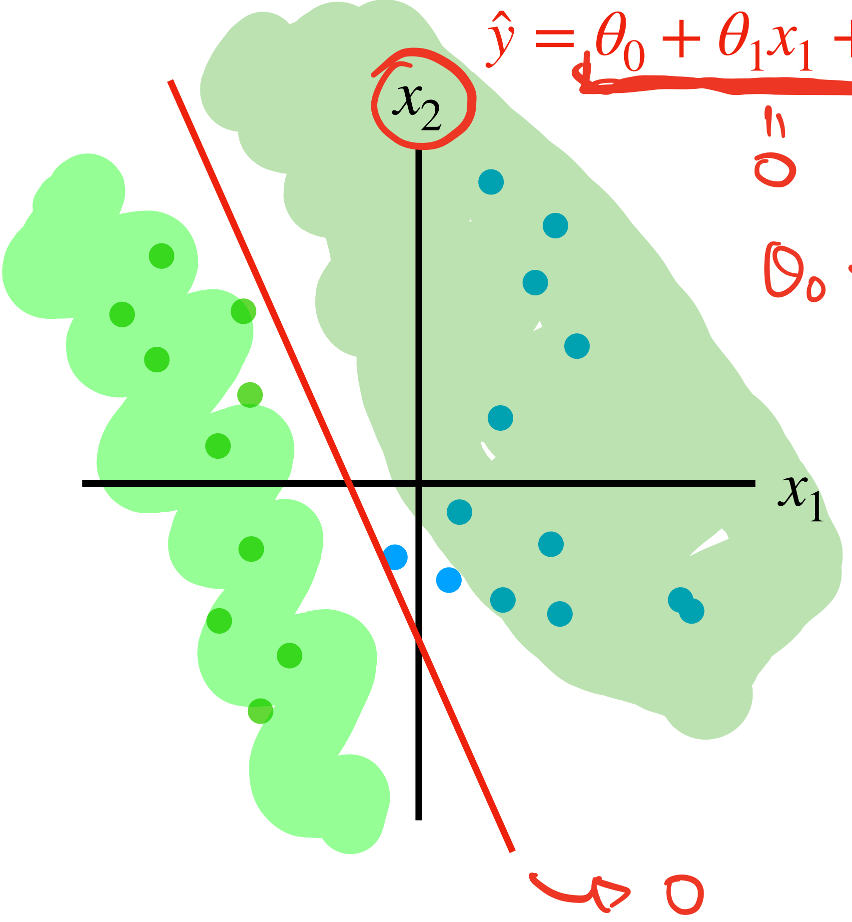
Summary

$\sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$



Logistic Regression

Summary



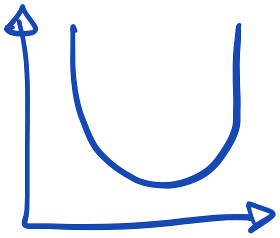
$$\underbrace{\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2}_{\shortparallel \atop 0}$$

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$$

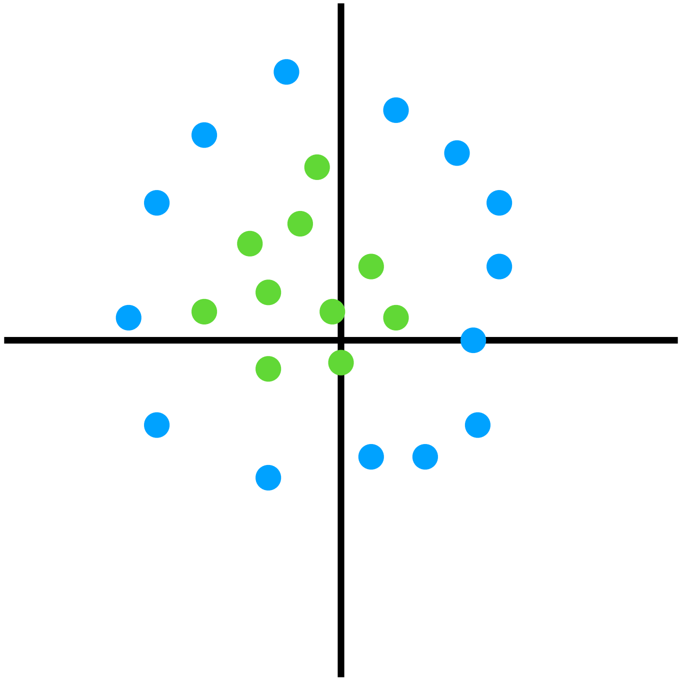
$$\underbrace{x_2} = \frac{-\theta_1 \boxed{x_1} - \theta_0}{\theta_2}$$

Logistic Regression

Summary

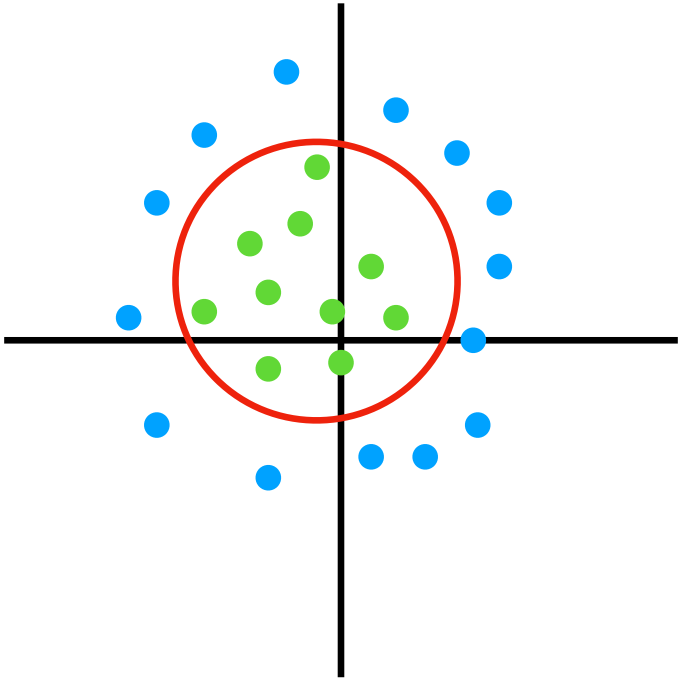


$\theta_0 + \theta_1 x^2$



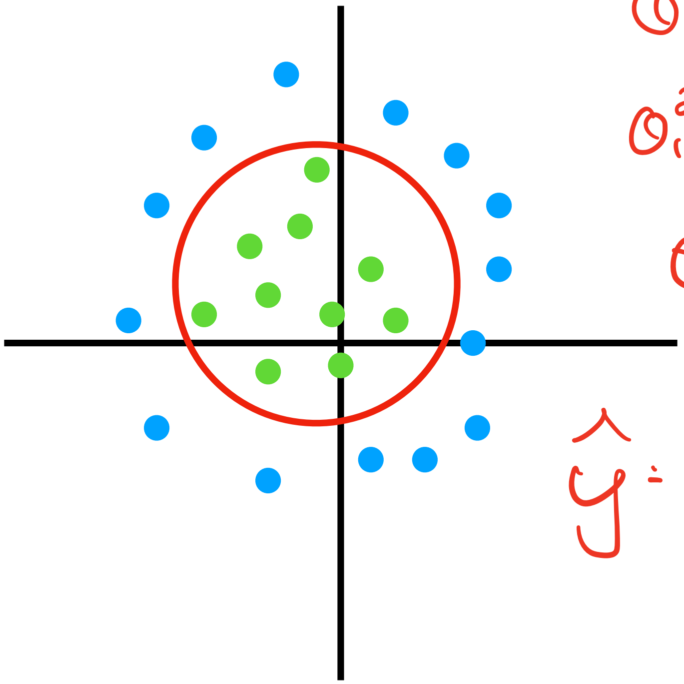
Logistic Regression

Summary



Logistic Regression

Summary



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

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

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

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

$$\hat{y} = \sigma\left(\theta_1 \underbrace{\sqrt{x_1^2 + x_2^2}}_{b} - \theta_0\right)$$

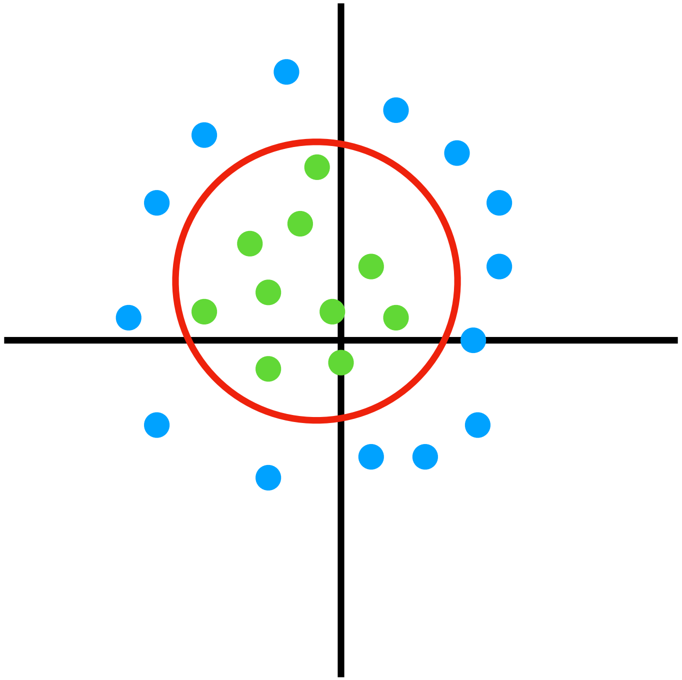
$$x_3$$

Logistic Regression

Summary

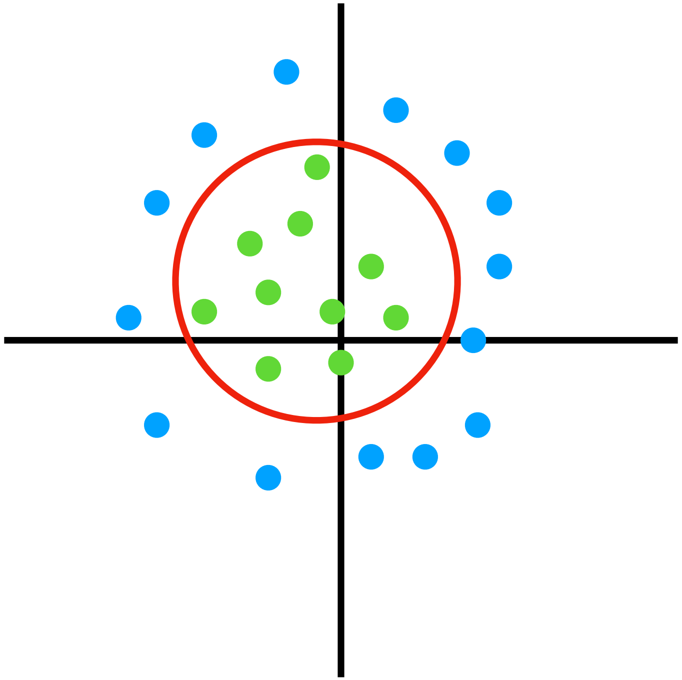
$$x_1^2 + x_2^2 = r^2$$

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



Logistic Regression

Summary



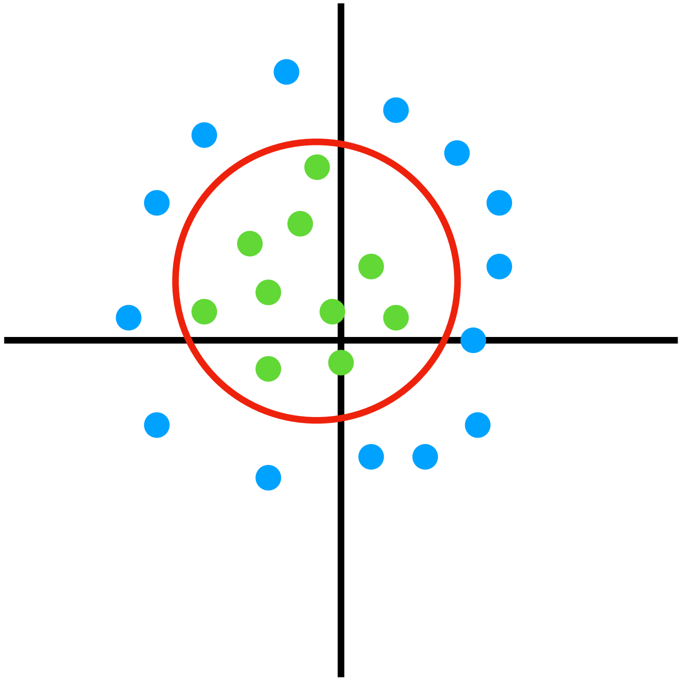
$$x_1^2 + x_2^2 = r^2$$

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

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

Logistic Regression

Summary



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

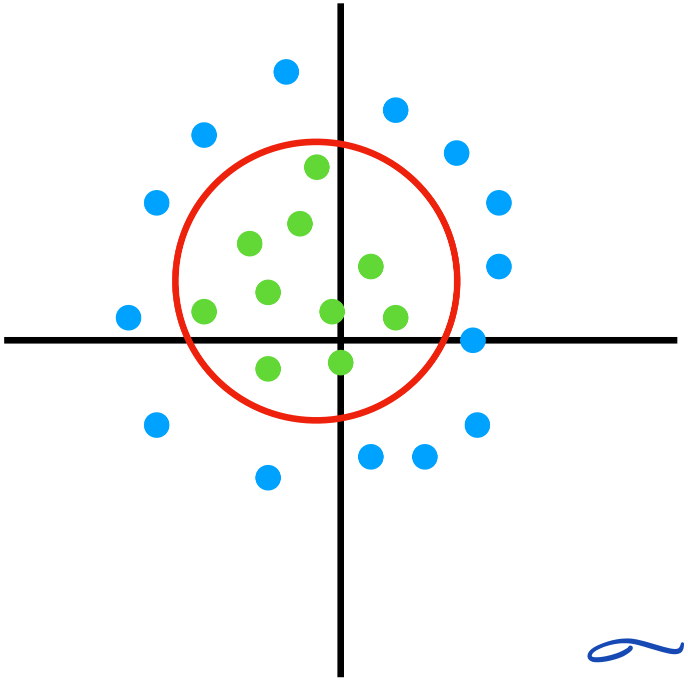
$$\theta_1^2(x_1^2 + x_2^2) = \theta_0^2$$

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

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

Logistic Regression

Summary



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

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

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

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

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

$$\sim \left(\theta_1 \sqrt{(x_1^2 + x_2^2} - \theta_0\right)$$

Next Class

- More classification algorithms