CSE 4820/5819 Introduction to Logistic Regression

Suining He
Department of Computer Science and Engineering
University of Connecticut
suining.he@uconn.edu

Outline

- Background
- Gradient Descent
- Comparison with Linear Regression

Classification

- ***Data***: A set of data records (also called examples, instances or cases) described by
  - Each example is labelled with a pre-defined class.
- ***Goal***: To learn a classification model from the data that can be used to predict the classes of new (future, or test) cases/instances.

Example: Loan Application Binary Classification

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**

old	false	false	fair	?

Why Use Logistic Regression

- Why?
  - Estimation by maximum likelihood
  - Interpreting coefficients
  - Hypothesis testing
  - Evaluating the performance of the model
- There are many important research topics for which the dependent variable is "limited."
- For example: voting, morbidity or mortality, and participation data is not continuous or distributed normally.

Background

- Binary ***logistic*** regression is a type of regression analysis where the dependent variable is binary
  - 0/1, True/False, Yes/No
  - For example: loan application.
    - Coded No (not approved) or Yes (approved)
  - $y \in \{0, 1\}$

P → probability a data pt. belongs to 1

Logistic Regression Model: An Overview

- The logistic regression model solves these problems:
  $\ln[p/(1-p)] = \boldsymbol{x}^T\boldsymbol{w} + b$

logarithm
natural base
$\ln \frac{p}{1-p}$

$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$ $w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$

- p is the probability that an event occurs
  - $p/(1-p)$ is the **odds ratio**
  - $\ln[p/(1-p)]$ is the **log odds ratio**, or **logit**

Logistic Regression Model: An Overview

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability (using $\hat{y}$ in the place of p) is

$$p = \hat{y} = \frac{1}{1+\exp(-z)}$$

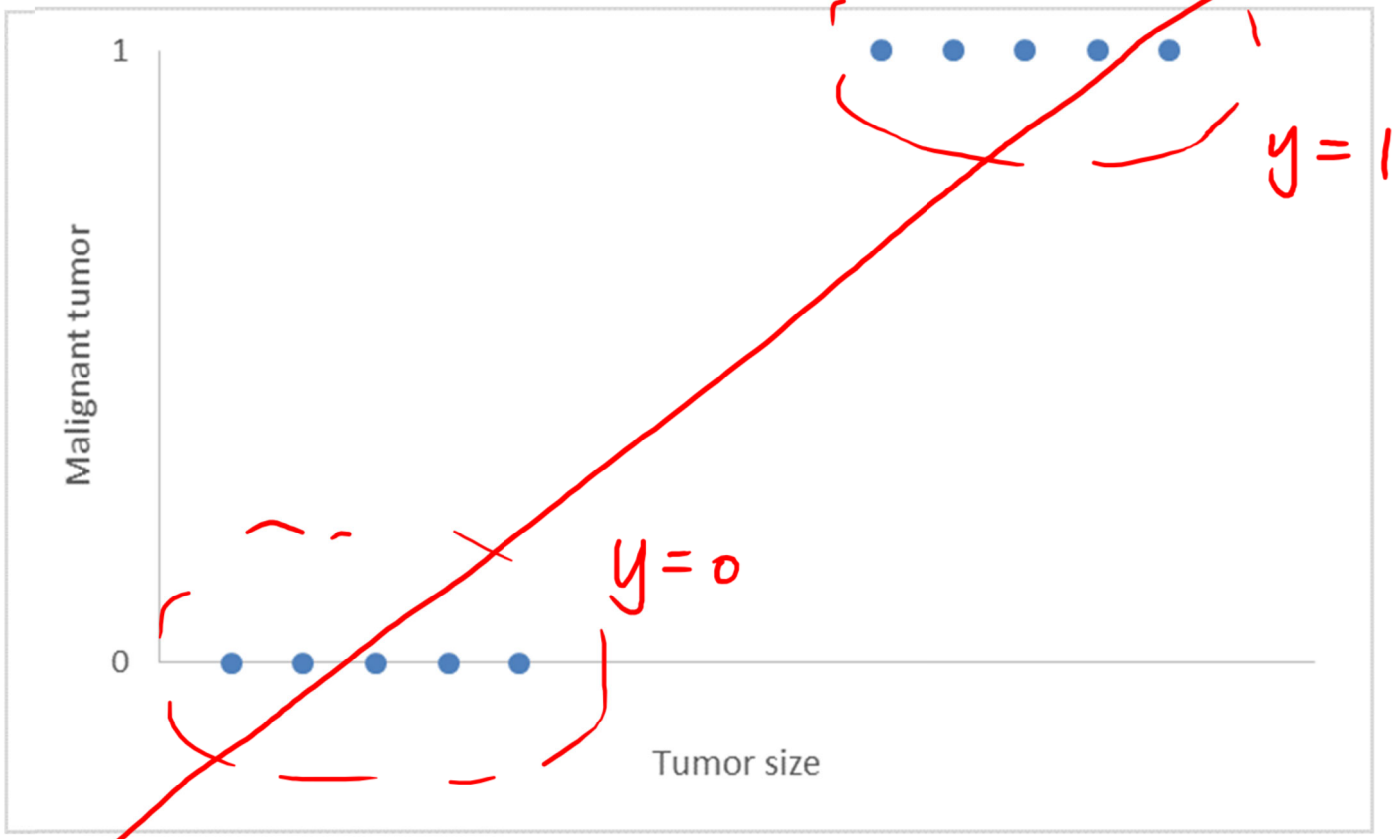
- where $z = \boldsymbol{x}^T\boldsymbol{w} + b$
- If you let b + $\boldsymbol{x}^T\boldsymbol{w}$ =0, then p = .50
  - As b + $\boldsymbol{x}^T\boldsymbol{w}$ gets really big, p approaches 1
  - As b + $\boldsymbol{x}^T\boldsymbol{w}$ gets really small, p approaches 0

$z = 0 \quad p = \frac{1}{1+\exp(0)}$

$z \to +\infty \quad p \to \frac{1}{1+0} = \frac{1}{2} = 0.5$

$z \to -\infty \quad p \to 0 = 1$

Example

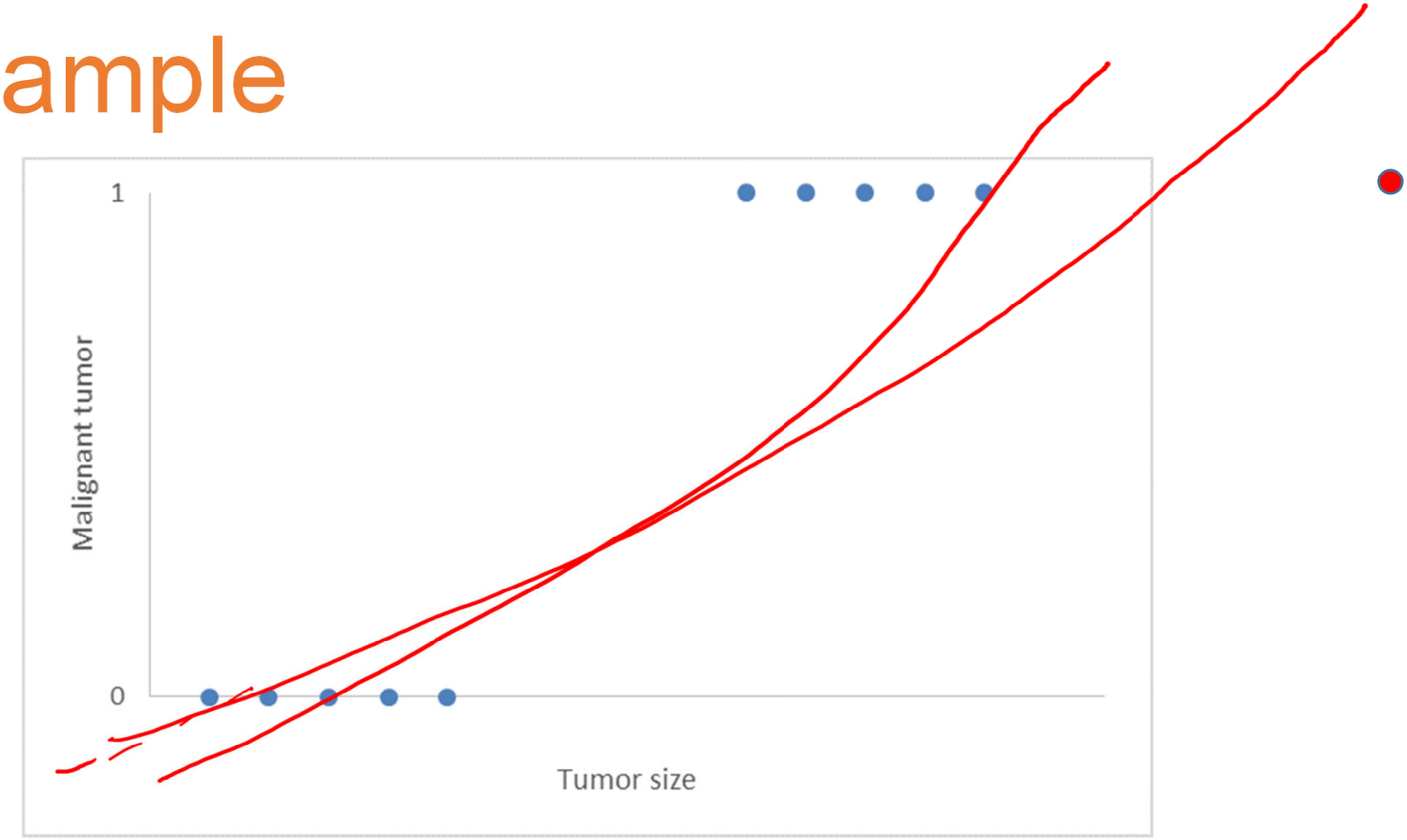


$y = x^T w + b$

linear model.

- Simplest way: thresholding

Example



- What about the linear model learned in linear regression?

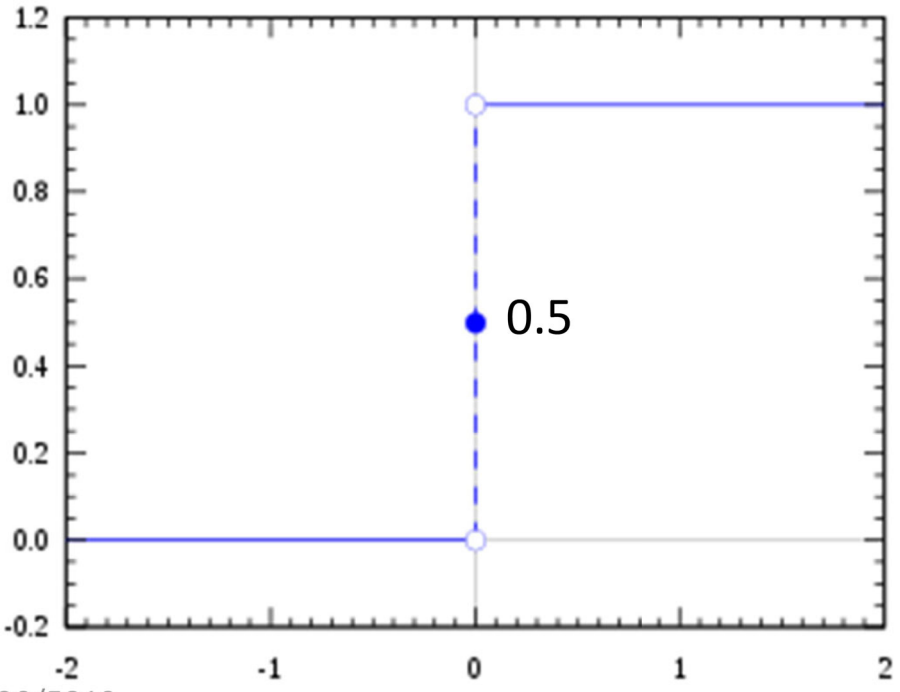
Origin and Connection with Linear Regression

- Recall $y = \boldsymbol{x}^T \boldsymbol{w} + b$ returns $y \in R$
  - It can be any real value
  - We need to turn it into the values of 0 and 1 for the *logistic* use
- One simple way:
  - Unit-step function

$z = x^T w + b$

$$\hat{y} = \begin{cases} 0, & \text{if } z < 0 \\ 0.5, & \text{if } z = 0 \\ 1, & \text{if } z > 0 \end{cases}$$

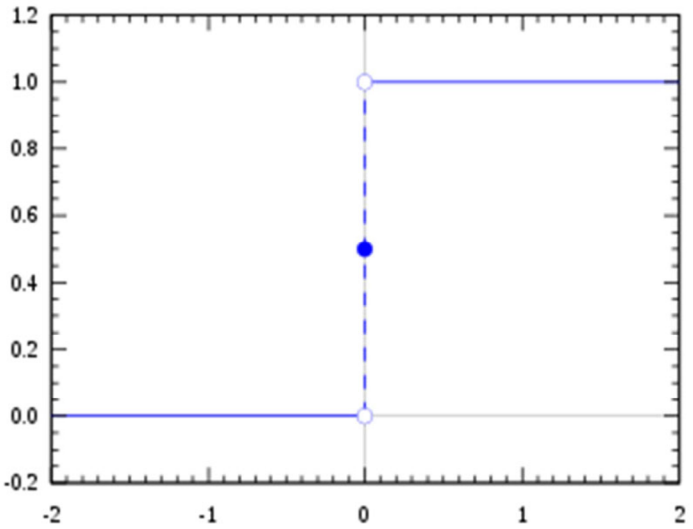
probability P



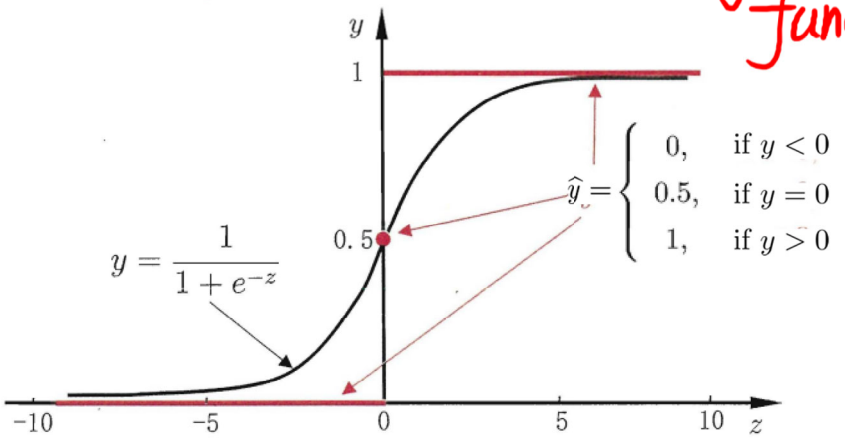

UNIVERSITY OF CONNECTICUT
1881

Normally, we consider Unit Step Function

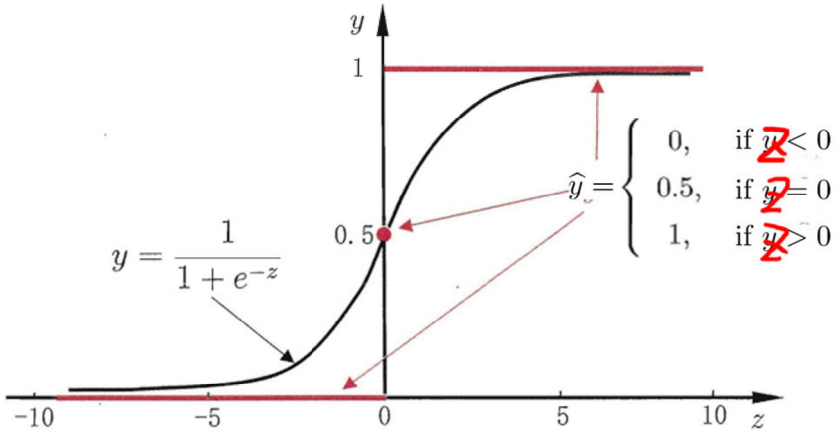
- Unit function is not ***continuous***
  - We need a *surrogate* function
  - A surrogate function is a function to approximate another



$y = \frac{1}{1+e^{-z}}$ sigmoid function



Surrogate Function



- Logistic function

$$\hat{y} = \frac{1}{1+\exp(-z)}$$

- With $z = \boldsymbol{x}^T \boldsymbol{w} + b$, then we can have

probility ⇐ $$\hat{y} = \frac{1}{1+\exp(-(\mathbf{x}^T \mathbf{w}+b))}$$

Logistic Regression

- Find a function f that best predicts the conditional probability directly?

  $p(y_i = \text{“yes” } or \text{ “no”} | \mathbf{x}_i)$ (binary classification)

  - Still find a linear function of $\mathbf{x}_i$ parameterized with $\mathbf{w}$, i.e., $\mathbf{x}_i^T\mathbf{w} + \mathrm{b}$
  - But $\mathbf{x}_i^T\mathbf{w} + \mathrm{b}$ is used to compute the probability (p)

- $P \in [0,1]$, however, $\mathbf{x}_i^T\mathbf{w} + \mathrm{b} \in (-\infty, +\infty)$

So, how to link $\mathbf{x}_i^T\mathbf{w} + b$ to p?

Further Transformation

probability.

- We can further convert $\hat{y} = \frac{1}{1+\exp(-(\mathbf{x}^T\mathbf{w}+b))}$ into

$$\ln \frac{\hat{y}}{1-\hat{y}} = \mathbf{x}^T\mathbf{w} + b$$

linear model.

- Recall that we are expecting y has value of either 1 or 0 (e.g., $y = 1$ for class 1 and $y = 0$ for class 2)
- The *odds* is defined as $\frac{\hat{y}}{1-\hat{y}}$
  - Log odds (logit): $\ln \frac{\hat{y}}{1-\hat{y}}$

⇒ output prob of data.

linear model.

$$\hat{y} = \frac{1}{1+\exp(-(\mathbf{x}^T\mathbf{w}+b))} \longrightarrow \ln\frac{\hat{y}}{1-\hat{y}} = \mathbf{x}^T\mathbf{w}+b$$

binary classification

toss a coin

p → probability head. ($\hat{y}$)

1-p → prob tail. ($1-\hat{y}$).

logit $\ln\frac{\hat{y}}{1-\hat{y}}$

$$1-\hat{y} = \frac{1+\exp(-(x^Tw+b))}{1+\exp(-(x^Tw+b))} - \frac{1}{1+\exp(-(x^Tw+b))}$$

$$= \frac{\exp(-(x^Tw+b))}{1+\exp(-(x^Tw+b))}$$

$$\frac{\hat{y}}{1-\hat{y}} = \frac{\frac{1}{1+\exp(-(x^Tw+b))}}{\frac{\exp(-(x^Tw+b))}{1+\exp(-(x^Tw+b))}} = \frac{1}{\exp(-(x^Tw+b))} = \exp(x^Tw+b)$$

$$\ln\frac{\hat{y}}{1-\hat{y}} = \ln(\exp(x^Tw+b)) = x^Tw+b.$$

Quick Example for Settings

$\vec{x} = \begin{bmatrix} x_w \\ x_h \end{bmatrix}$ $\vec{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ weight, height

$\ln \frac{\hat{y}}{1-\hat{y}} = x^T w + b \Leftrightarrow \hat{y} = \frac{1}{1+\exp(-(x^T w+b))}$

1st Pika $\vec{x}_1 = \begin{bmatrix} 4.79 \\ 0.38 \end{bmatrix}$

$\hat{y}_1 = \frac{1}{1+\exp(-(x^T w+b))}$

$= \frac{1}{1+\exp\left(-\left(\begin{bmatrix} 4.79 \\ 0.38 \end{bmatrix}^T \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + b\right)\right)}$

y 0 Not Evolved 1 Evolved → 0 ground truth (not evolved)

$p \rightarrow$ probability that pokemon evolved (prob($y=1$))

$\hat{y}_2 = \frac{1}{1+\exp\left(-\left(\begin{bmatrix} -31.42 \\ 0.82 \end{bmatrix}^T \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + b\right)\right)}$

→ 1 ground truth (evolved)

Connection Between Linear Regression and Logistic Regression

$$\text{Odds}(P) = \frac{P(y_i=1|\mathbf{x}_i)}{P(y_i=0|\mathbf{x}_i)} = \frac{P(y_i=1|\mathbf{x}_i)}{1-P(y_i=1|\mathbf{x}_i)} = \frac{p}{1-p}$$

p → probab belongs to 1

1-p → prob belongs to 0

Connection Between Linear Regression and Logistic Regression

$$\ln \frac{p}{1-p} = x^T w + b$$

$$\text{Odds}(P) = \frac{P(y_i=1|\mathbf{x}_i)}{P(y_i=0|\mathbf{x}_i)} = \frac{P(y_i=1|\mathbf{x}_i)}{1-P(y_i=1|\mathbf{x}_i)}$$

Probability $P(y_i = 1|\mathbf{x}_i)$	Odds
1.0	$+\infty$
0.99	99
0.75	3
0.5	1
0.25	0.3333
0.01	0.0101
0	0

$\frac{p}{1-p}$

Probability: $\mathrm{P} \in [0,1]$

Odd value: $\text{Odds}(P) \in (0, +\infty)$

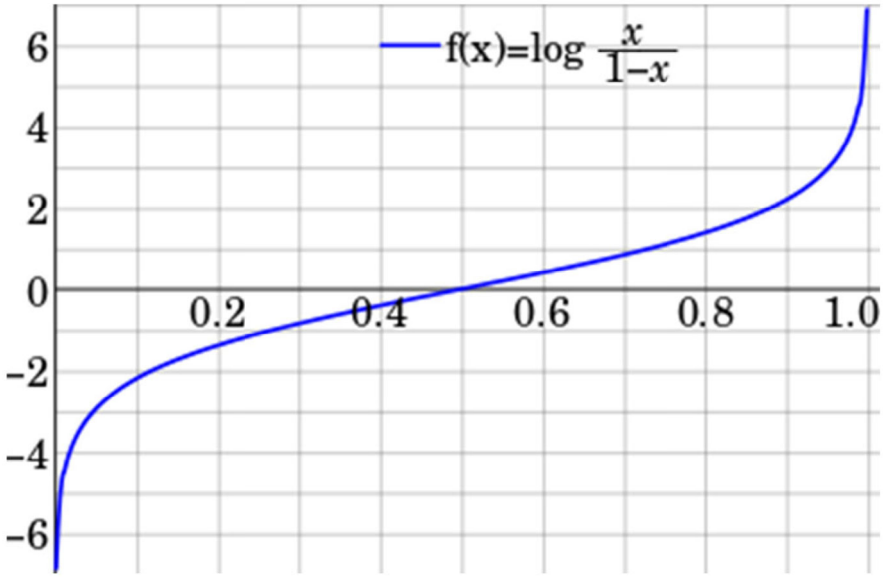
$\mathbf{x}^{\mathbf{T}}\mathbf{w} + b \in (-\infty, +\infty)$?

Logit Function

- Take the logarithm of odds

$$\ln(Odds) = \ln(\frac{P}{1-P})$$

logit.



We call this function the ***logit function*** of P, i.e.,

$$logit(P) = \ln(\frac{P}{1-P})$$

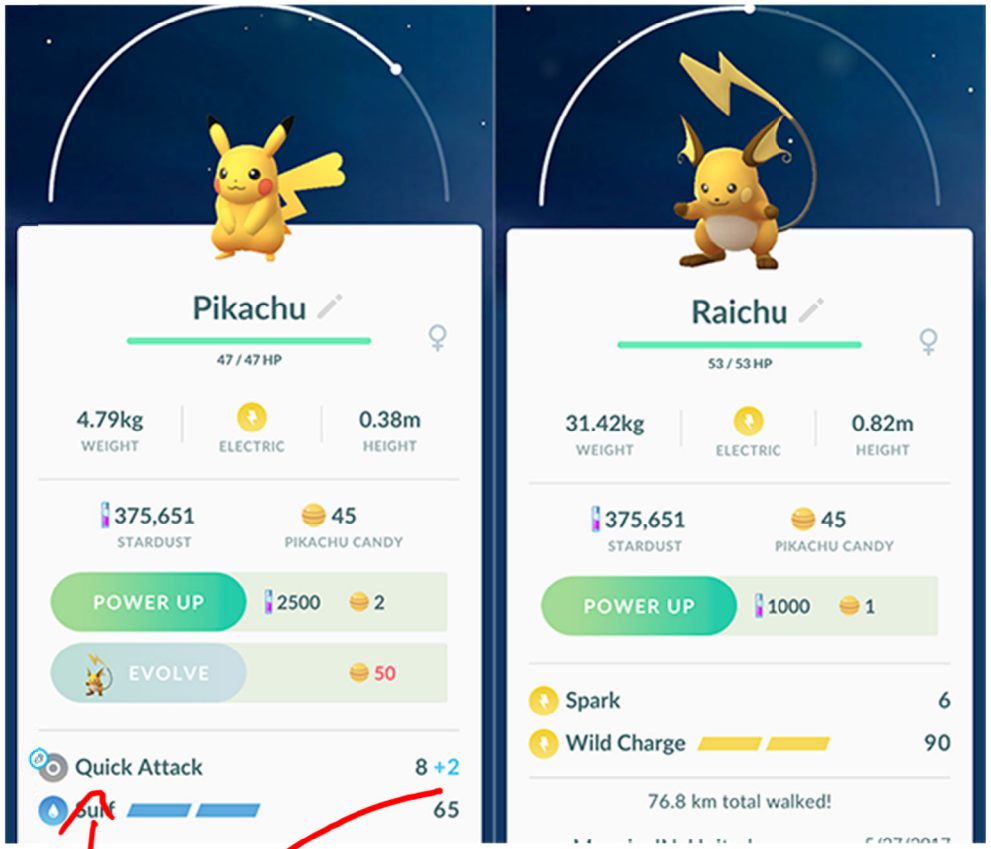
$$logit(P) \in (-\infty, +\infty)$$

$$\mathbf{x}^T\mathbf{w} + \mathbf{b} \in (-\infty, +\infty)$$

attribute of data pt.

Quick Example (Again!)

$$\log\left(\frac{y}{1-y}\right) = x^T w + b$$



① pika

$y_1 = 0$. $\hat{y}_1 \to 0$

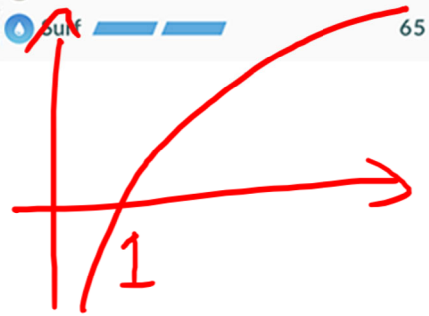
true label. output

odds ratio.

$$\frac{y_1}{1-y_1} = 0 \qquad \frac{\hat{y}_1}{1-\hat{y}_1} \to 0$$

logit.

$$\log\left(\frac{y_1}{1-y_1}\right) \to -\infty \qquad \log\left(\frac{\hat{y}_1}{1-\hat{y}_1}\right) \to -\infty$$ sign



$f(x) = \log x$

② $y_2 = 1$ $\hat{y}_2 \to 1$

$$\log\left(\frac{\hat{y}_2}{1-\hat{y}_2}\right) \to +\infty$$ sign

Interpretation of Model

$$f_{\mathbf{w},b}(x) = \hat{y} = \frac{1}{1+\exp(-(\mathbf{x}^T\mathbf{w}+b))}$$

- Interpreted as:
  - Estimated probability that $y = 1$ on the input x
- Tumor example x = size of tumor
  - $\hat{y} = 0.7$
  - Tell patient that 70% chance of tumor being malignant

$$f_{\mathbf{w},b}(x) = P(\hat{y} = 1|x; \mathbf{w}, b)$$

- "Probability that $\hat{y} = 1$, given x, parameterized by $\boldsymbol{w}, b$"
- $\mathrm{P}(\hat{y} = 1|x; \hat{w}) + \mathrm{P}(\hat{y} = 0|x; \hat{w}) = 1$
- $\mathrm{P}(\hat{y} = 1|x; \hat{w}) = 1 - \mathrm{P}(\hat{y} = 0|x; \hat{w})$

$\hat{\mathbf{w}} = [\boldsymbol{w}; \boldsymbol{b}]$

$\hat{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ b \end{bmatrix}$

Logistic Regression

- Assume log odds is a linear function of $\mathbf{x}$

logit

$$\ln\left(\frac{P(y_i = 1|\mathbf{x}_i)}{1 - P(y_i = 1|\mathbf{x}_i)}\right) = \mathbf{x}_i^T \hat{\mathbf{w}}$$

linear model.

if $\mathbf{x}_i^T \hat{w} > 0$ $p > 0.5$

When $\boldsymbol{x}_i^T \hat{\mathbf{w}} > 0$, $\mathrm{P}(y_i = 1|\boldsymbol{x}_i) > 0.5$

if $\mathbf{x}_i^T \hat{w} < 0$ $p < 0.5$

$$P(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \hat{\mathbf{w}})}{1 + \exp(\mathbf{x}_i^T \hat{\mathbf{w}})}$$

- ***Linear decision*** boundary, the line $\mathbf{x}_i^T \hat{\mathbf{w}} = 0$ $p = 0.5$
- ***Decision Rule***: For binary classification, if P(y=1|X)> P(y=0|X) (or P(y=1|X)>0.5), then Y=1; otherwise, Y=0

Quick Deduction

$$\hat{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \\ b \end{bmatrix}$$

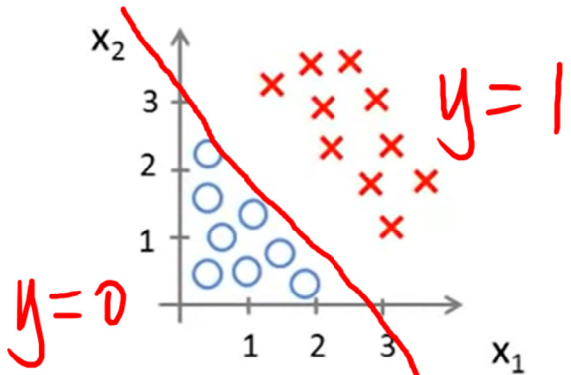
$$p_i = \frac{1}{1+\exp(-(x_i^T w + b))}$$

$$= \frac{1}{1+\exp(-\hat{x}_i^T \hat{w})}$$

$$= \frac{1 \cdot \exp(\hat{x}_i^T \hat{w})}{1 \cdot \exp(\hat{x}_i^T \hat{w}) + \exp(-\hat{x}_i^T \hat{w}) \cdot \exp(-\hat{x}_i^T \hat{w})}$$

$$= \frac{\exp(\hat{x}_i^T \hat{w})}{\exp(\hat{x}_i^T \hat{w}) + 1}$$

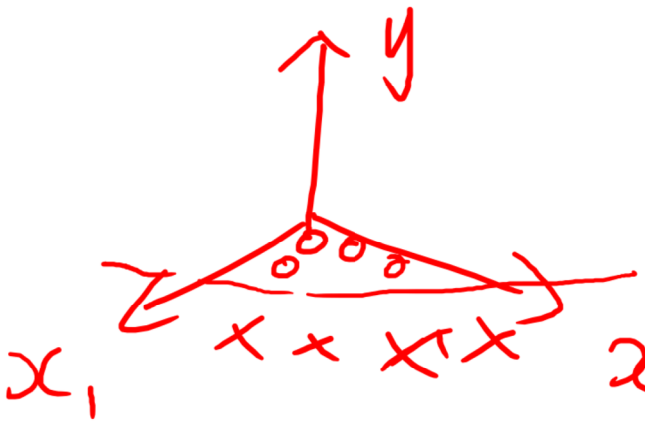
Example of Using Decision Boundary



$$f_{\mathbf{w},b}(x) = \hat{y} = \frac{1}{1+\exp(-(\mathbf{x}^T\mathbf{w}+b))} = \frac{1}{1+\exp(b+w_1x_1+w_2x_2)}$$

- Predict "$\hat{y} = 1$" if $-3 + x_1 + x_2 \geq 0$
  - Decision boundary $x_1 + x_2 = 3$

$\sum_i^T \hat{w} = 0$



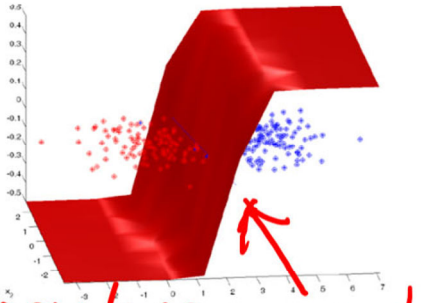
$$y = \frac{1}{1+\exp(-(x^Tw+b))}$$

Visualization

$p = \frac{1}{1+\exp(-3+x_1+x_2)}$

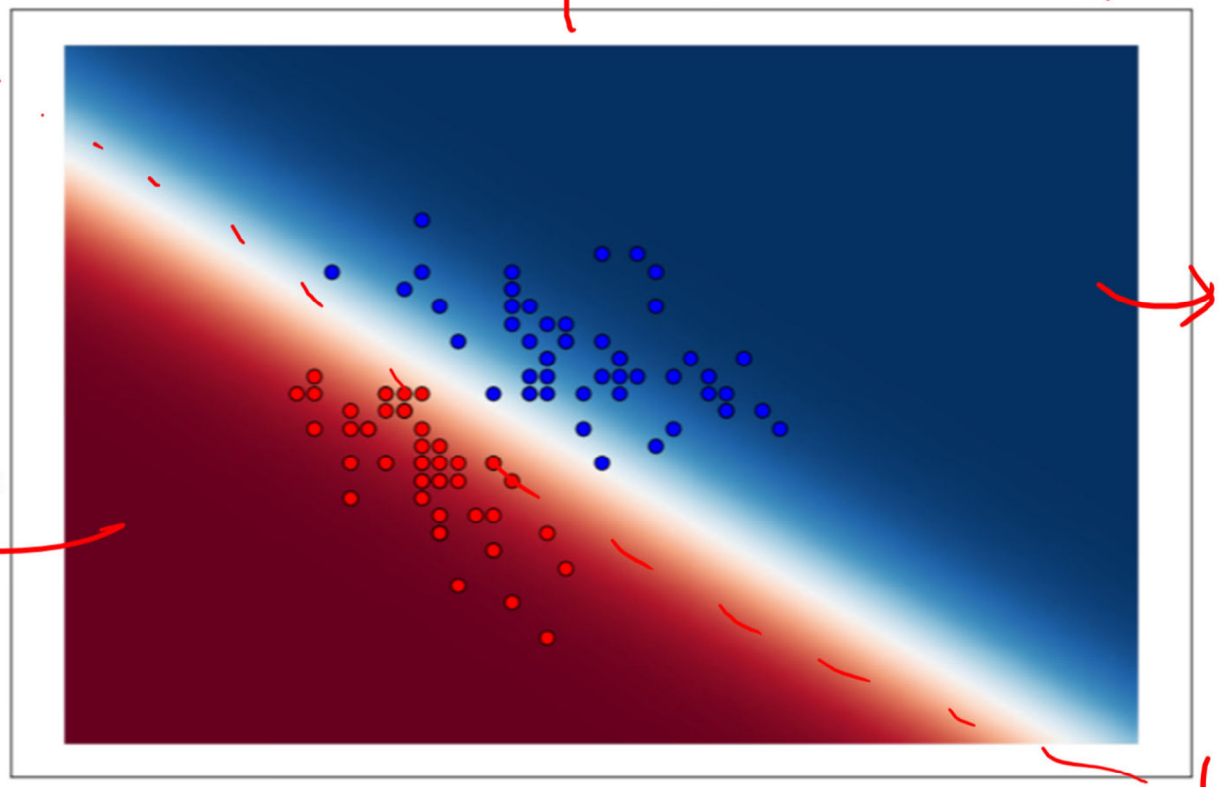
if $-3+x_1+x_2 = 0$

$p = 0.5$



hyperplane surrogate function

$p = 0.5$



$p(y=1|x) \geq 0$

$\downarrow$

$y \to 0$

$p(y=1|x)$

$\Downarrow$

$y = 1$

decision boundary

$-3+x_1+x_2=0$

More on Logistic Regression

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

$$\hat{y} = \frac{1}{1+\exp(-(\mathbf{x}^T\mathbf{w}+b))}$$

- If you let $x^T w + b = 0$, then $\hat{y} = 0.5$
  - As $x^T w + b$ gets really big, $\hat{y}$ approaches 1
  - As $x^T w + b$ gets really small, $\hat{y}$ approaches 0

Understanding with Maximum Likelihood

- When $y_i = 1$, find $\mathbf{w}$ that maximizes

$$P(y_i = 1|\mathbf{x}_i)$$

- When $y_i = 0$, find $\mathbf{w}$ that maximizes

$$P(y_i = 0|\mathbf{x}_i) = 1 - P(y_i = 1|\mathbf{x}_i)$$

- Overall

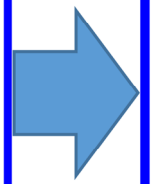
$$\max_{\mathbf{w},b} \left(\prod_{i:y_i=1} P(y_i = 1|\mathbf{x}_i) \prod_{i:y_i=0} \left(1 - P(y_i = 1|\mathbf{x}_i)\right) \right)$$

Example

Training Data

y_n: 1 for class 1 (C_1), 0 for class 2 (C_2)

x_1	x_2	x_3	……
C_1	C_1	C_2	



x_1	x_2	x_3	……
$y_1 = 1$	$y_2 = 1$	$y_3 = 0$	

Assume the data is generated based on

$$f_{\mathbf{w},b}(x) = P_{\mathbf{w},b}(C_1|x) = P(y_i = 1|\mathbf{x}_i)$$

Given a set of $\boldsymbol{w}$ and b, what is the joint probability of generating the data?

$$L(w,b) = f_{\mathbf{w},b}(x_1) f_{\mathbf{w},b}(x_2) \left(1 - f_{\mathbf{w},b}(x_3)\right) \cdots f_{\mathbf{w},b}(x_N)$$

The most likely w^* and b^* is the one with the largest $L(\boldsymbol{w}, b)$.

$$\boldsymbol{w}^*, b^* = arg \max_{\boldsymbol{w},b} L(w,b)$$

Maximum Likelihood Example

$y_1 = 0$ $y_2 = 1$ $y_3 = 0$ $y_4 = 1$

$\hat{y}_1 \to 0$ pika

$\hat{y}_2 \to 1$ raichu

$\hat{y}_3 \to 0$ mew

$\hat{y}_4 \to 1$ mew2

$$L(w,b) = (1-f_{w,b}(x_1)) \cdot f_{w,b}(x_2) \cdot (1-f_{w,b}(x_3)) \cdot f_{w,b}(x_4)$$

$$1-f_{w,b}(x_1) = \left(f_{w,b}(x_1)\right)^{y_1}\left(1-f_{w,b}(x_1)\right)^{1-y_1}$$

$y_1 = 0$.

$$f_{w,b}(x_2) = \left(f_{w,b}(x_2)\right)^{y_2}\left(1-f_{w,b}(x_2)\right)^{1-y_2}$$

$y_2 = 1$

Bernoulli Distribution

- We can write it in a way of **Bernoulli distribution**

$$\max_{\mathbf{w},\mathrm{b}} \left(\prod_i (f_{\mathbf{w},b}(x))^{y_i} (1 - f_{\mathbf{w},b}(x))^{(1-y_i)} \right)$$

objective

- $f_{\mathbf{w},b}(x)$ represent the probability that the y_i is 1 (belong to C_1)

Example


Pikachu
47 / 47 HP
4.79kg
WEIGHT
ELECTRIC
0.38m
HEIGHT
375,651
STARDUST
45
PIKACHU CANDY
POWER UP
2500
2
EVOLVE
50
Quick Attack
8 +2
Surf
65
Raichu
53 / 53 HP
31.42kg
WEIGHT
ELECTRIC
0.82m
HEIGHT
375,651
STARDUST
45
PIKACHU CANDY
POWER UP
1000
1
Spark
6
Wild Charge
90
76.8 km total walked!


CP 3927
Mew
HP 1000/ 1000
Psychic
Type
13.12 kg
Weight
0.78 m
Height
62655
STARDUST
28
MEW CANDY
POWER UP
1900
2
EVOLVE
25
CP 4629
Mewtwo
HP 1000/ 1000
Psychic
Type
13.12 kg
Weight
0.78 m
Height
62655
STARDUST
28
MEWTWO CANDY
POWER UP
1900
2
EVOLVE
25

x_1	x_2	x_3	……
C_1	C_1	C_2	

x_1	x_2	x_3	……
$y_1 = 1$	$y_2 = 1$	$y_3 = 0$	

y_n: 1 for class 1, 0 for class 2

$$L(w,b) = f_{\boldsymbol{w},b}(x_1) f_{\boldsymbol{w},b}(x_2) \left(1 - f_{\boldsymbol{w},b}(x_3)\right) \cdots$$

$$\boldsymbol{w}^*, b^* = arg \max_{\boldsymbol{w},b} L(\boldsymbol{w}, b) \quad = \quad \boldsymbol{w}^*, b^* = arg \min_{\boldsymbol{w},b} -\log(L(\boldsymbol{w}, b))$$

$$f_{w,b}(x_i) = \left(f_{(w,b)}(x)\right)^{y} \cdot \left(1 - f_{w,b}(x)\right)^{1-y}$$

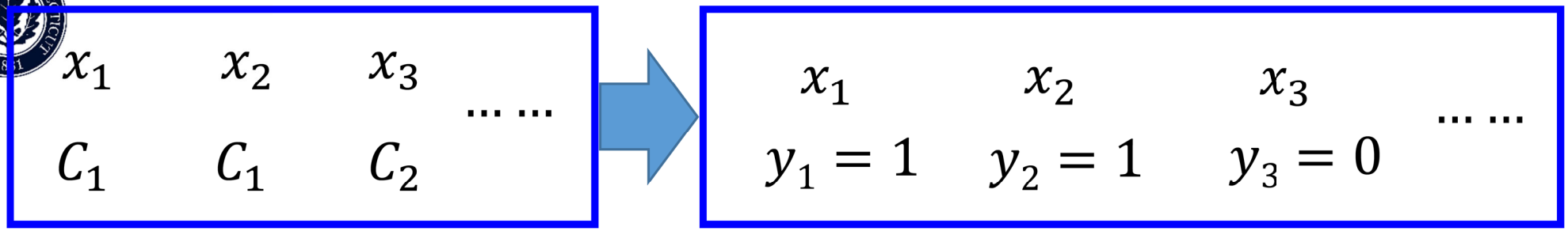
$$-\log L(\boldsymbol{w}, b)$$

$$= -\log f_{\boldsymbol{w},b}(x_1) \rightarrow -[1 \log f(x_1) + (1 - 0) \log(1 - f(x_1))]$$

$$-\log f_{\boldsymbol{w},b}(x_2) \rightarrow -[1 \log f(x_2) + (1 - 0) \log(1 - f(x_2))]$$

$$-\log\left(1 - f_{\boldsymbol{w},b}(x_3)\right) \rightarrow -[0 \log f(x_3) + (1 - 1) \log(1 - f(x_3))]$$

$$\vdots$$



y_n: 1 for class 1, 0 for class 2

Maximum Likelihood Estimation

- Parameter estimation of $\boldsymbol{w}$ and b

$$P(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\mathbf{w}+\mathrm{b})}{1+\exp(\mathbf{x}_i^T\mathbf{w}+\mathrm{b})}, \quad P(y_i = 0|\mathbf{x}_i) = \frac{1}{1+\exp(\mathbf{x}_i^T\mathbf{w}+\mathrm{b})}$$

$f_{w,b}(x)$

- Maximum likelihood: find the parameters that can maximize the log likelihood

$$L(\boldsymbol{w}, b) = \sum_{i=1}^{m} \log p(y_i|x_i; \boldsymbol{w}, b)$$

Preparing Maximum Likelihood

$$\max_{\mathbf{w},b}\left(\prod_{i:y_i=1} P(y_i=1|\mathbf{x}_i)\prod_{i:y_i=0}\left(1-P(y_i=1|\mathbf{x}_i)\right)\right)$$

Let $p_i = P(y_i = 1|\mathbf{x}_i)$

$$\max_{\mathbf{w},\mathrm{b}}\left(\prod_i {p_i}^{y_i}(1-p_i)^{(1-y_i)}\right)$$

Preparing Maximum Log-Likelihood

$$L(\boldsymbol{w}, b) = \sum_{i=1}^{m} \log p(y_i | x_i; \boldsymbol{w}, b)$$

$$\max_{\mathbf{w},\mathrm{b}} \left(\prod_i {p_i}^{y_i} (1 - p_i)^{(1-y_i)} \right)$$

Take log

$$\max_{\mathbf{w},\mathrm{b}} \left(\sum_i y_i \log p_i + (1 - y_i) \log(1 - p_i) \right)$$

$p_i = \frac{\exp(\mathbf{x}_i^T \widehat{\mathbf{w}})}{1+\exp(\mathbf{x}_i^T \widehat{\mathbf{w}})}$, where $\widehat{\boldsymbol{w}} = (\boldsymbol{w}; b)$

The loss function $L(w)$

Solved with gradient descent

Journey for Solving the Logistic Regression

- Loss function

$$L(\hat{\mathbf{w}}) = \max_{\hat{\mathbf{w}}} \left(\sum_i y_i \log p_i + (1 - y_i) \log(1 - p_i) \right)$$

$$p_i = \frac{\exp(\mathbf{x}_i^T \hat{\mathbf{w}})}{1 + \exp(\mathbf{x}_i^T \hat{\mathbf{w}})}$$

$$\max_{\hat{\mathbf{w}}} \left(\sum_i y_i \mathbf{x}_i^T \hat{\mathbf{w}} - \log\left(1 + \exp(\mathbf{x}_i^T \hat{\mathbf{w}})\right) \right)$$

$$\min_{\hat{\mathbf{w}}} \left(\sum_i \log\left(1 + \exp(\mathbf{x}_i^T \hat{\mathbf{w}})\right) - y_i \mathbf{x}_i^T \hat{\mathbf{w}} \right)$$

Solved with gradient descent

The loss function $L(\hat{w})$

Deduction (Again!)

Understanding the Loss Function: Cross Entropy

$$L(\mathbf{w}, b) = f_{\mathbf{w},b}(x_1) f_{\mathbf{w},b}(x_2) \left(1 - f_{\mathbf{w},b}(x_3)\right) \cdots f_{\mathbf{w},b}(x_N)$$

$$-\log\left(L(\mathbf{w}, b)\right) = -[\log(f_{\mathbf{w},b}(x_1)) + \log\left(f_{\mathbf{w},b}(x_2)\right) + \log\left(\left(1 - f_{\mathbf{w},b}(x_3)\right)\right) \cdots]$$

y_i: 1 for class 1, 0 for class 2

$$= \sum_i -\left[y_i \log f_{\mathbf{w},b}(x_i) + (1 - y_i) \log\left(1 - f_{\mathbf{w},b}(x_i)\right)\right]$$

Cross entropy between two Bernoulli distributions

Distribution p:

$\mathrm{p}(x = 1) = y_i$

$\mathrm{p}(x = 0) = 1 - y_i$

cross entropy

Distribution q:

$\mathrm{q}(x = 1) = f(x_i)$

$\mathrm{q}(x = 0) = 1 - f(x_i)$

$$H(p, q) = -\sum_x p(x) \log\left(q(x)\right)$$

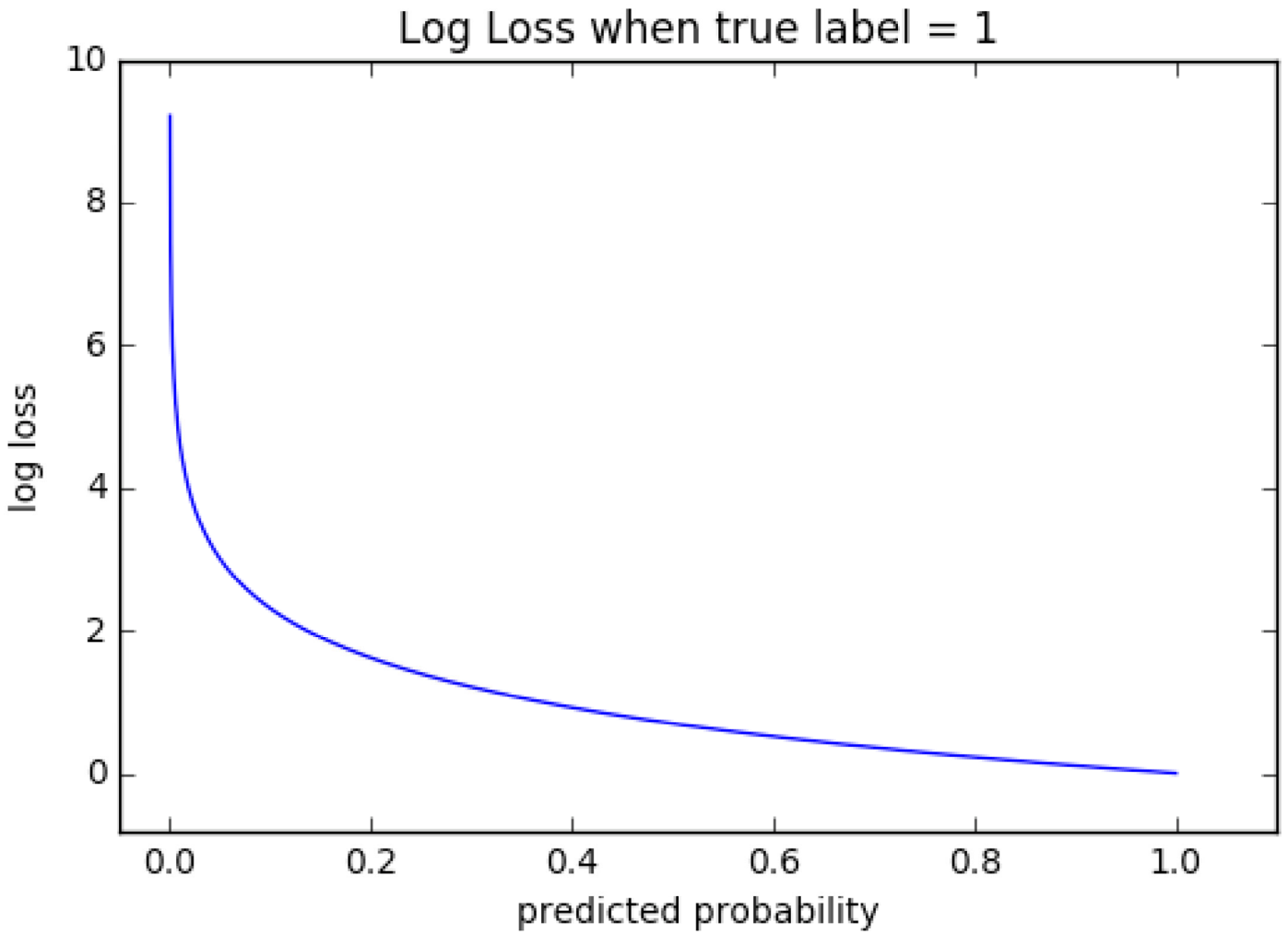
Entropy and Cross Entropy

- In information theory, we like to describe the "*surprise*" of an event. An event is more surprising the less likely it is, meaning it contains more information.
  - Low Probability Event (surprising): More information.
  - Higher Probability Event (unsurprising): Less information.
- Information $h(x)$ can be calculated for an event x, given the probability of the event $P(x)$ as follows:
  - $h(x) = -\log(P(x))$

Entropy and Cross Entropy

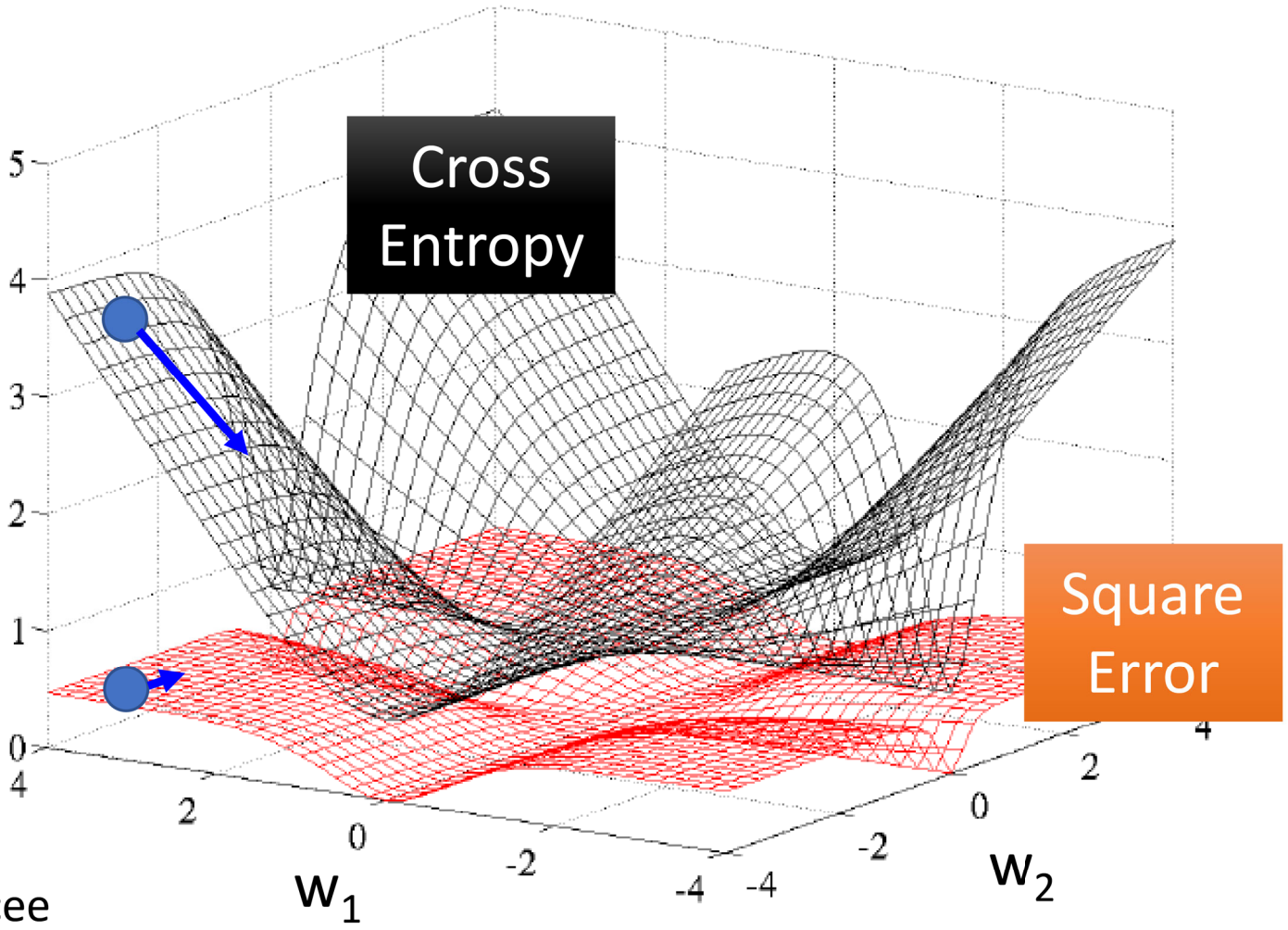
- **Entropy** is the number of bits required to transmit a randomly selected event from a probability distribution.
  - A skewed distribution has a low entropy, whereas a distribution where events have equal probability has a larger entropy.
- **Cross-entropy**: cross-entropy loss increases as the predicted probability diverges from the actual label.
  - Calculates the number of bits required to represent or transmit from one distribution (prediction) compared to another distribution (ground-truth).

Cross-Entropy



Cross Entropy v.s. Square Error

Total Loss



http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf

How to Find Gradient?

$$\min_{\hat{\mathbf{w}}} \left(\sum_i \log\left(1 + \exp(\mathbf{x}_i^T \hat{\mathbf{w}})\right) - y_i \mathbf{x}_i^T \hat{\mathbf{w}} \right)$$

How to Find Gradient?

$$\min_{\hat{\mathbf{w}}} \left(\sum_i \log\left(1 + \exp(\mathbf{x}_i^T \hat{\mathbf{w}})\right) - y_i \mathbf{x}_i^T \hat{\mathbf{w}} \right)$$

Gradient Descent

Iteratively find a $\widehat{w}$ that the gradient of the log probability function close to 0

1. $\widehat{w}^0 = 0$
2. $\widehat{w}^{(k+1)} = \widehat{w}^{(k)} - \alpha \nabla_{\widehat{w}^{(k)}} L(\widehat{w}^{(k)})$

Until reach termination condition

Relation with Neural Network (Spoiler Alert…)

Just Another Example

- https://www.youtube.com/watch?v=zHYF-o7tQ4o

What We can Consider

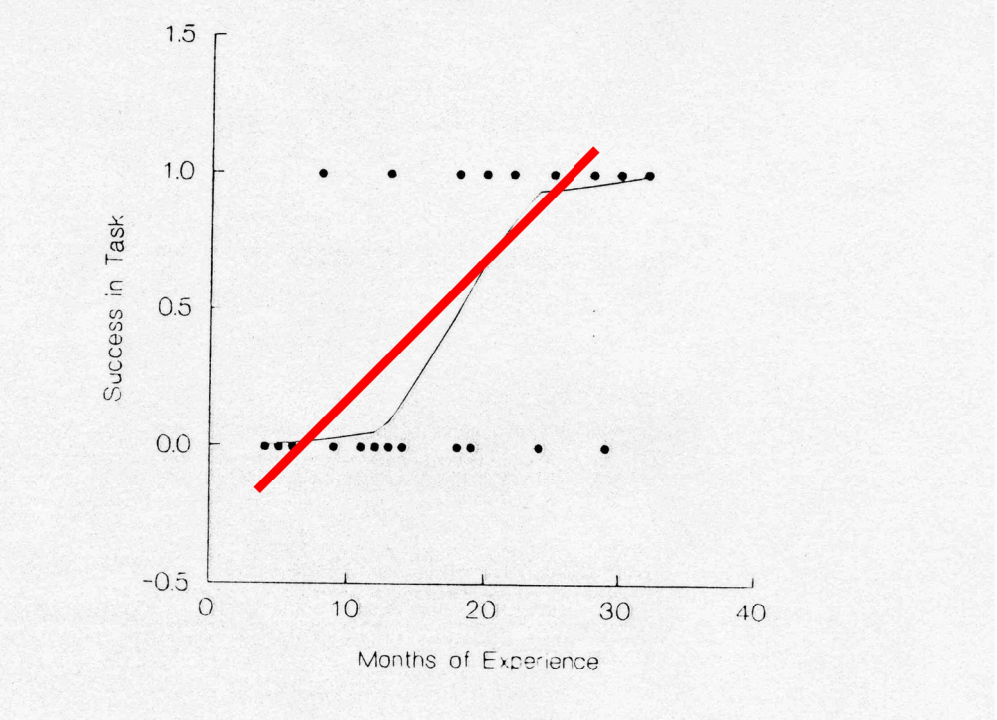
- Shang Dynasty: Bone Crack (maybe related to temperature and humidity…) and Fortune of the Dynasty
- Temperature, humidity, and success: [80F, 10%] and 0; [60F, 70%] and 1;

Example for Deduction: Logistic Regression Training

Example for Deduction

Another Example

- A systems analyst studied the effect of computer programming experience on ability to complete a task within a specified time
- Twenty-five persons selected for the study, with varying amounts of computer experience (in months)
- Results are coded in binary fashion: *Y* = 1 if task completed successfully; *Y* = 0, otherwise



Another Example

- Results from a standard package give:
  - $b = -3.0597$ and $w = 0.1615$
- Estimated logistic regression function:

$$\hat{y} = \frac{1}{1+e^{3.0597-0.1615X}}$$

- For example, the fitted value for $X = 14$ is:

$$\hat{y} = \frac{1}{1+e^{3.0597-0.1615(14)}} = 0.31$$

(Estimated probability that a person with 14 months experience will successfully complete the task)

Similarities

- Linear Regression and Logistic Regression both are *supervised learning* algorithms.
- Linear Regression and Logistic Regression, both the models use *linear* models within the parameterization process

Differences

- Linear Regression is used to handle *regression* problems whereas Logistic regression is used to handle the *classification* problems.
- Linear regression provides a *continuous* output, but Logistic regression provides *discrete* output.
- The purpose of Linear Regression is to find the best-fitted line while Logistic regression is one step ahead and fitting the line values to the sigmoid curve.
- The method for calculating loss function in linear regression is the *mean squared error* whereas for logistic regression it is *maximum likelihood estimation*.

Conclusion

- Background
- Key Ideas
- Gradient Descent for Logistic Regression