Logistic Regression

Outline

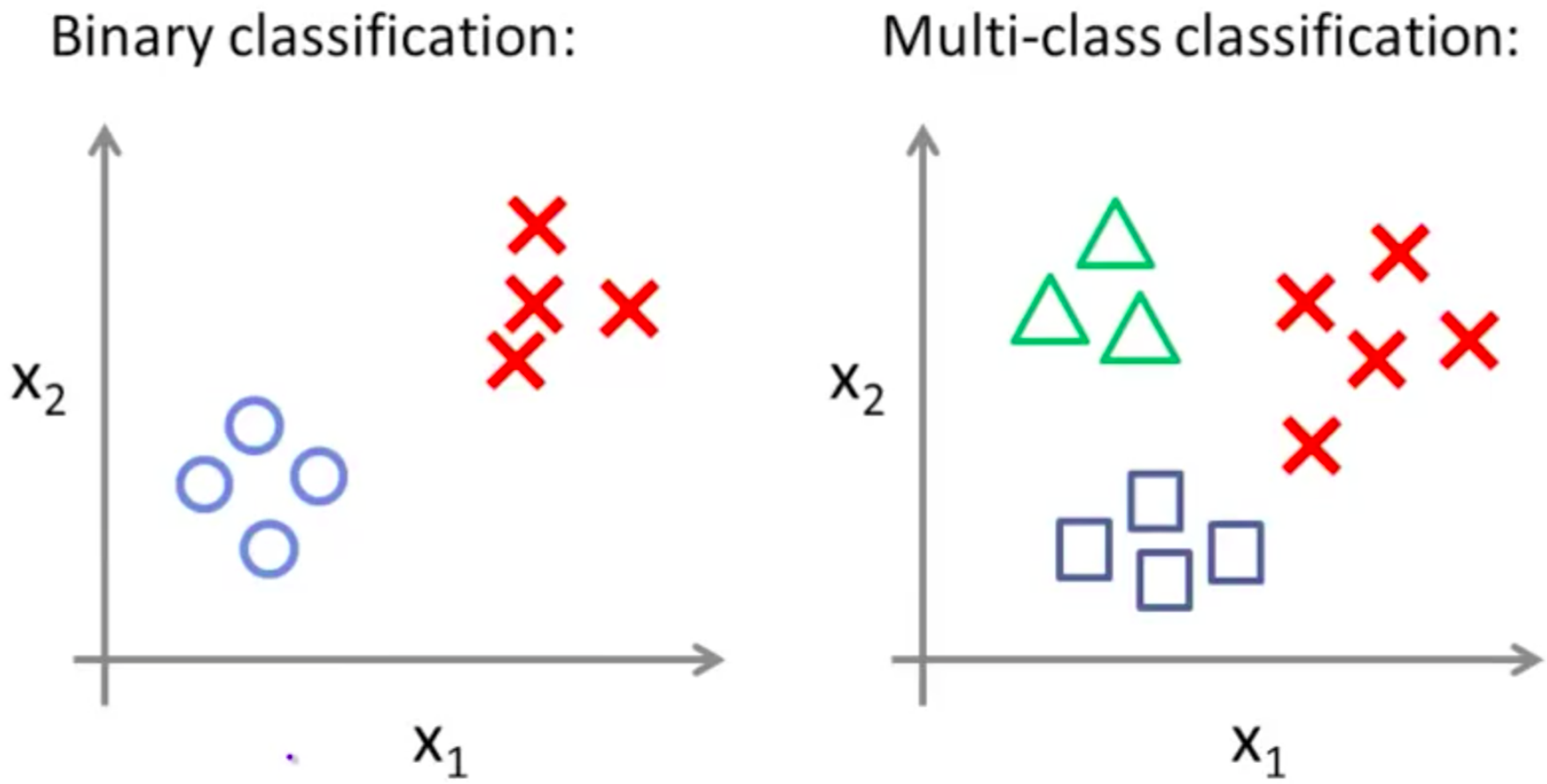
- Assumptions
- Logistic Regression Function
- Cost Function
- Optimization
- Applications

What is a Logistic Regression ?

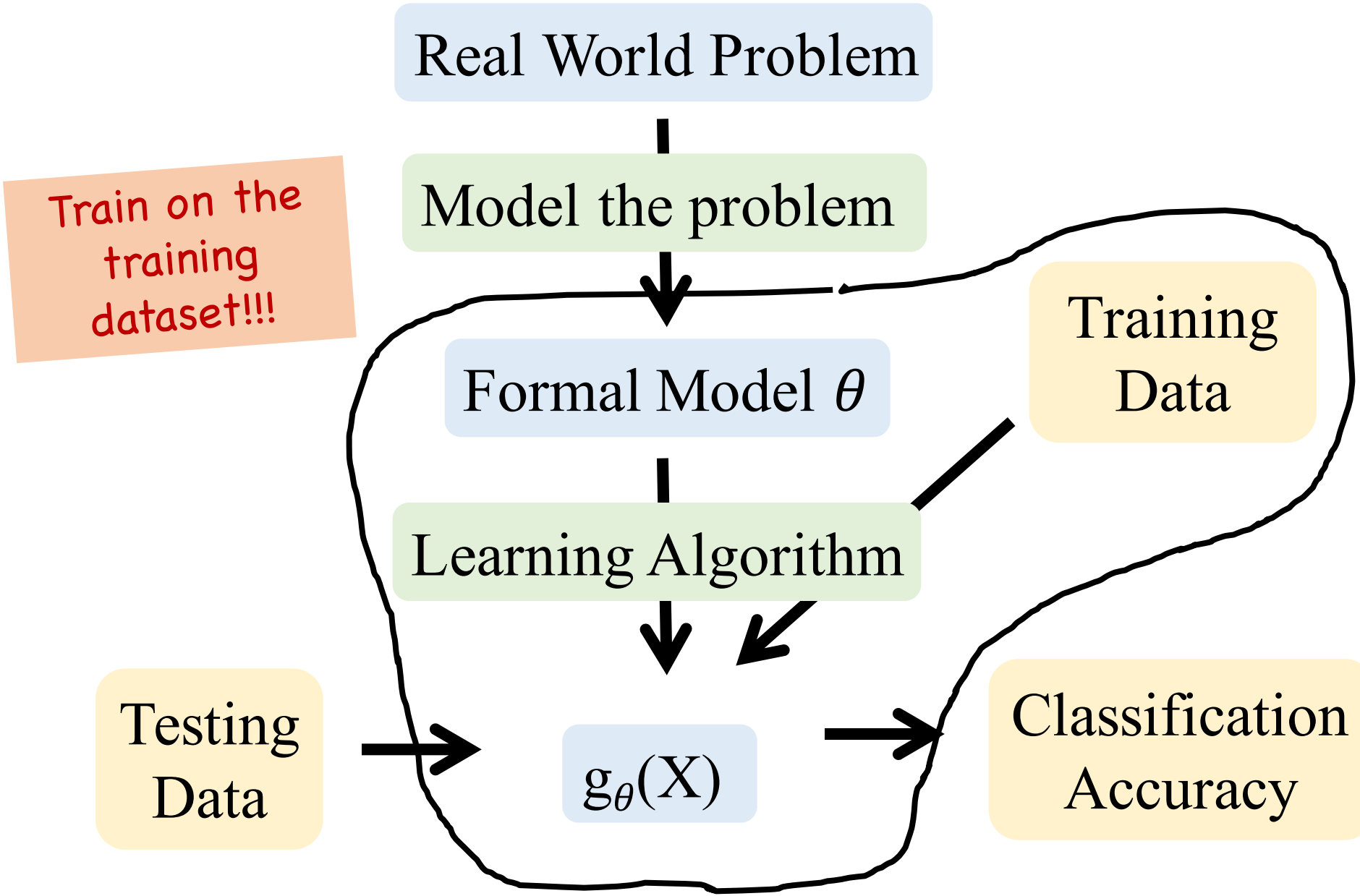
- A model for doing probabilistic binary classification
- Predicts **label probabilities** rather than a hard value of the label
- The model's prediction is a probability, defined using the **sigmoid function**

$$Y=1/(1+\exp^{-(W^T x)})$$

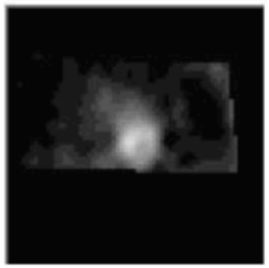
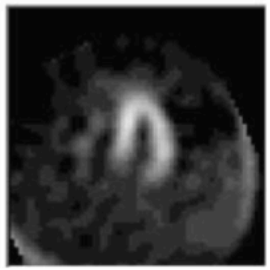
Binary vs Multiclass ?



Training

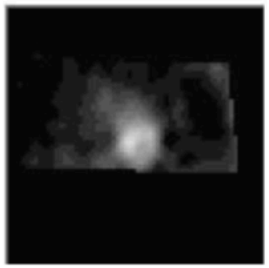
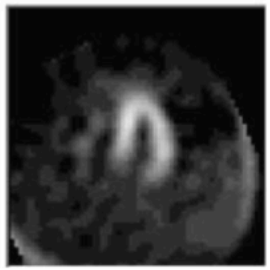

Real World Problem
Model the problem
Train on the training dataset!!!
Formal Model θ
Training Data
Learning Algorithm
Testing Data
$g_\theta(X)$
Classification Accuracy

Healthy Heart Classifier

	ROI 1	ROI 2	ROI m	Output
			… 	
Heart 1	0	1	1	0
Heart 2	1	1	1	0
		⋮		⋮
Heart n	0	0	0	1

$$g_\theta(\mathrm{X})$$

Healthy Heart Classifier

	ROI 1	ROI 2	ROI m	Output
			… 	
New Heart	1	0	1	1

$$g_\theta(X)$$

Binary Logistic Regression Assumptions:

- The dependent variable should be **binary (dichotomous)** in nature (e.g., presence vs. absent).
- There should be **no outliers** in the data.
- There should be no high correlations **(multicollinearity)** among the predictors. This can be assessed by a correlation matrix among the predictors.

Logistic Regression

- At the center of the logistic regression analysis is the task estimating the **log odds of an event**.
- Mathematically, logistic regression estimates a **multiple linear regression function** defined as:

$$\mathbf{Log\left(\frac{P(y=1)}{1-P(y=1)}\right)=W^{T}X}$$

Logistic Regression

- The sigmoid first computes a **real-valued “score”**, and **“squashes”** it between**(0,1)** to turn this score into a probability score
- **Model parameter is the unknownw.**

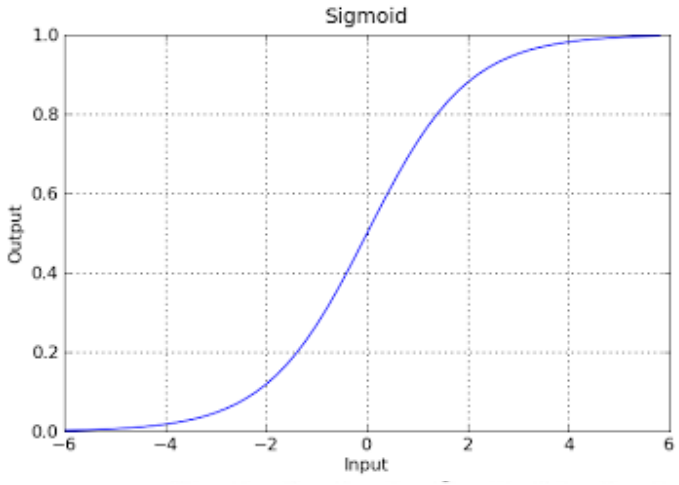
  Need to learn it from training data

Sigmoid Function

Logistic function (Sigmoid)

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

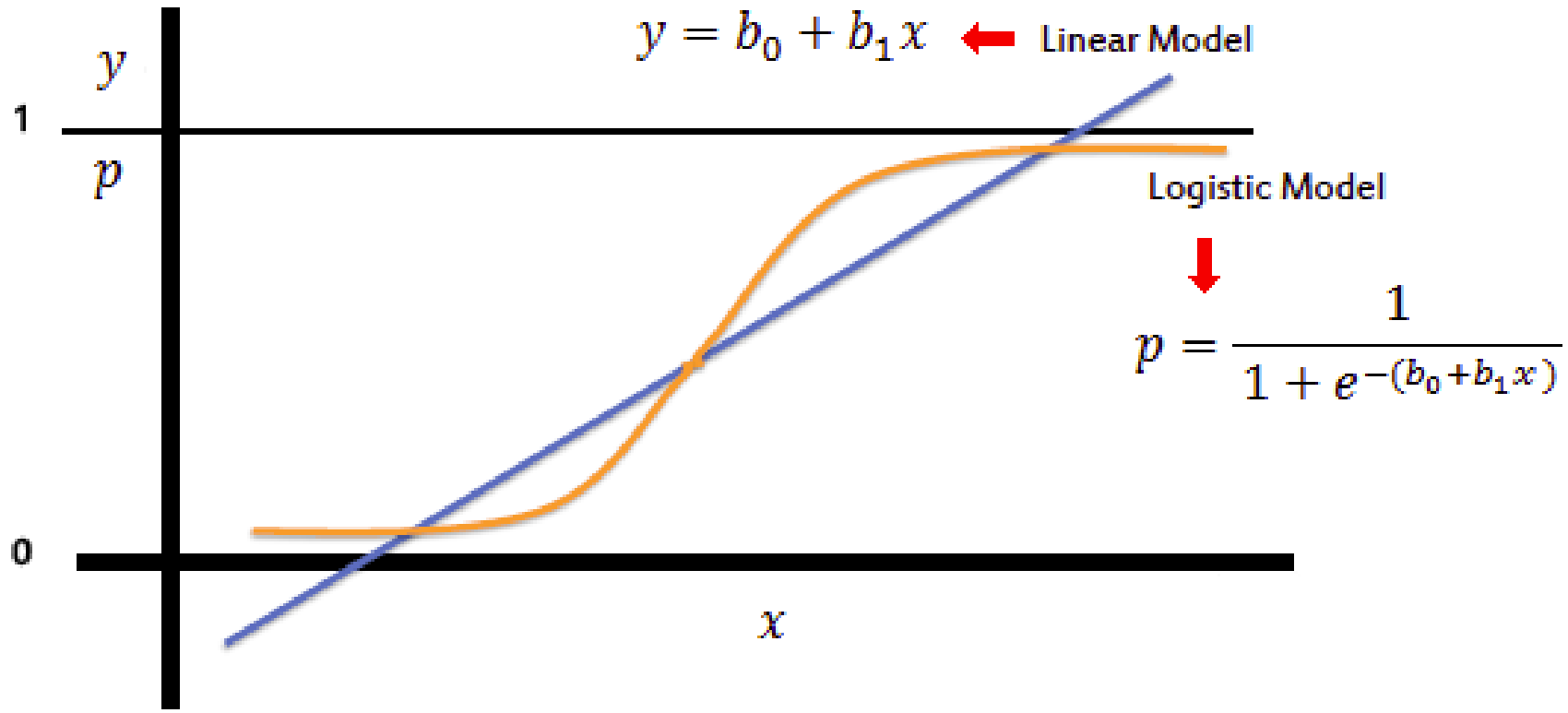
$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}(e^{-z} \\ &= \frac{1}{(1+e^{-z})} \cdot \left(1 - \\ &= g(z)(1-g(z)). \end{aligned}$$



The derivative of logistic function has a nice feature:

$$g'(z) = g(z)(1 - g(z))$$

Linear vs Logistic Regression

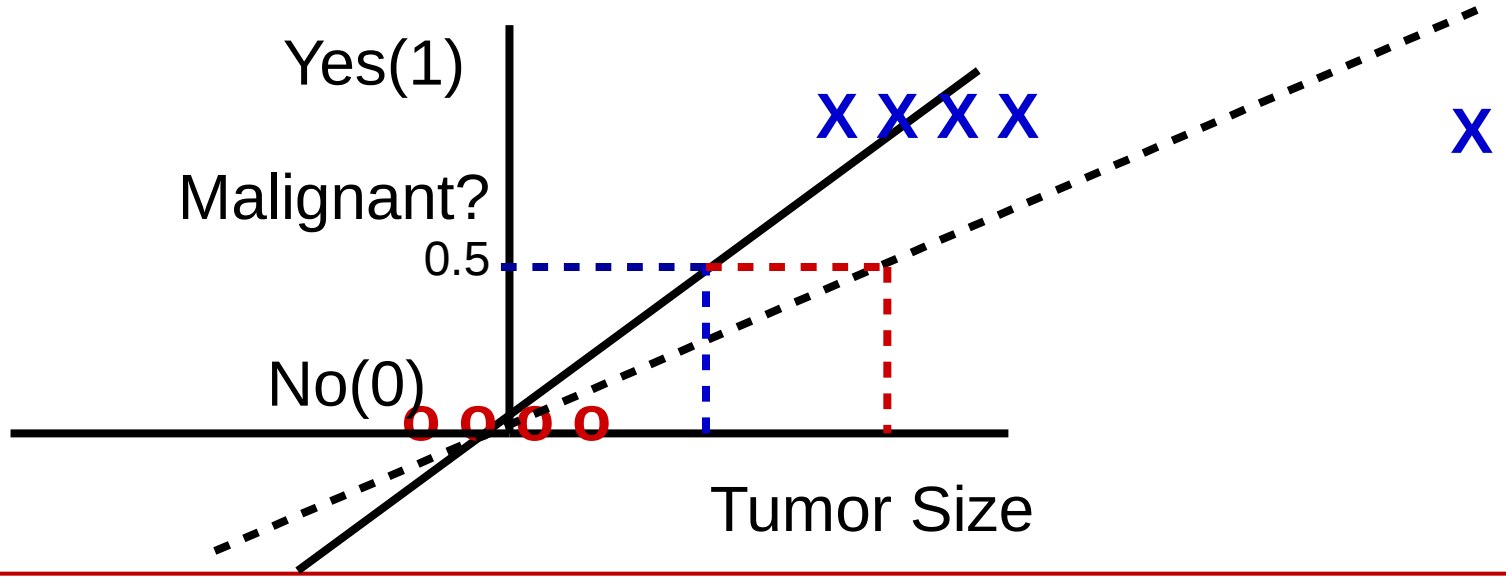


Linear vs Logistic Regression

- Issue 1 of Linear Regression

  As you can see on the graph, your prediction

  would leave out “malignant tumors” as the gradient becomes less steep with an additional data point on the extreme right

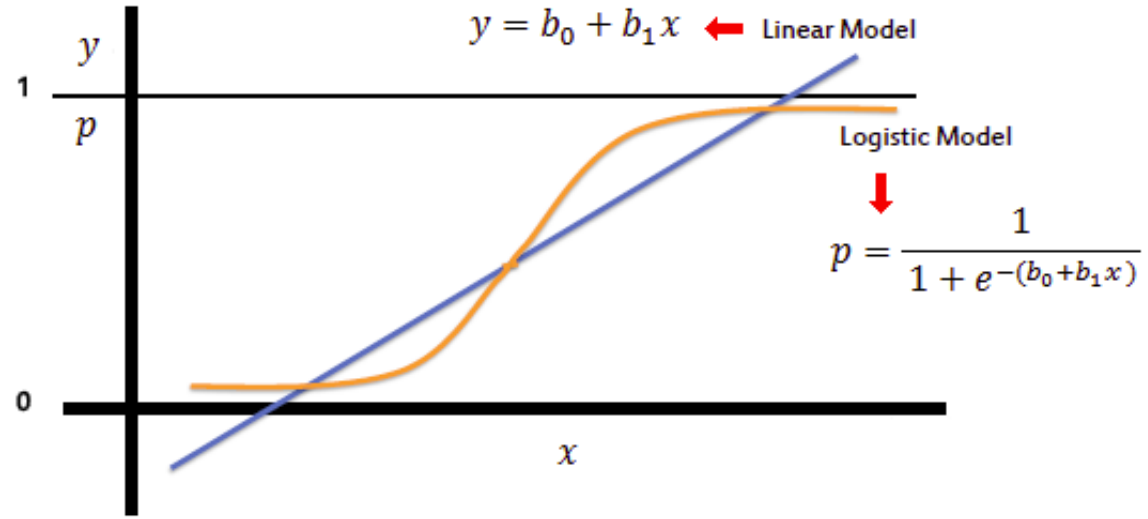


Linear vs Logistic Regression

- Issue 2 of Linear Regression

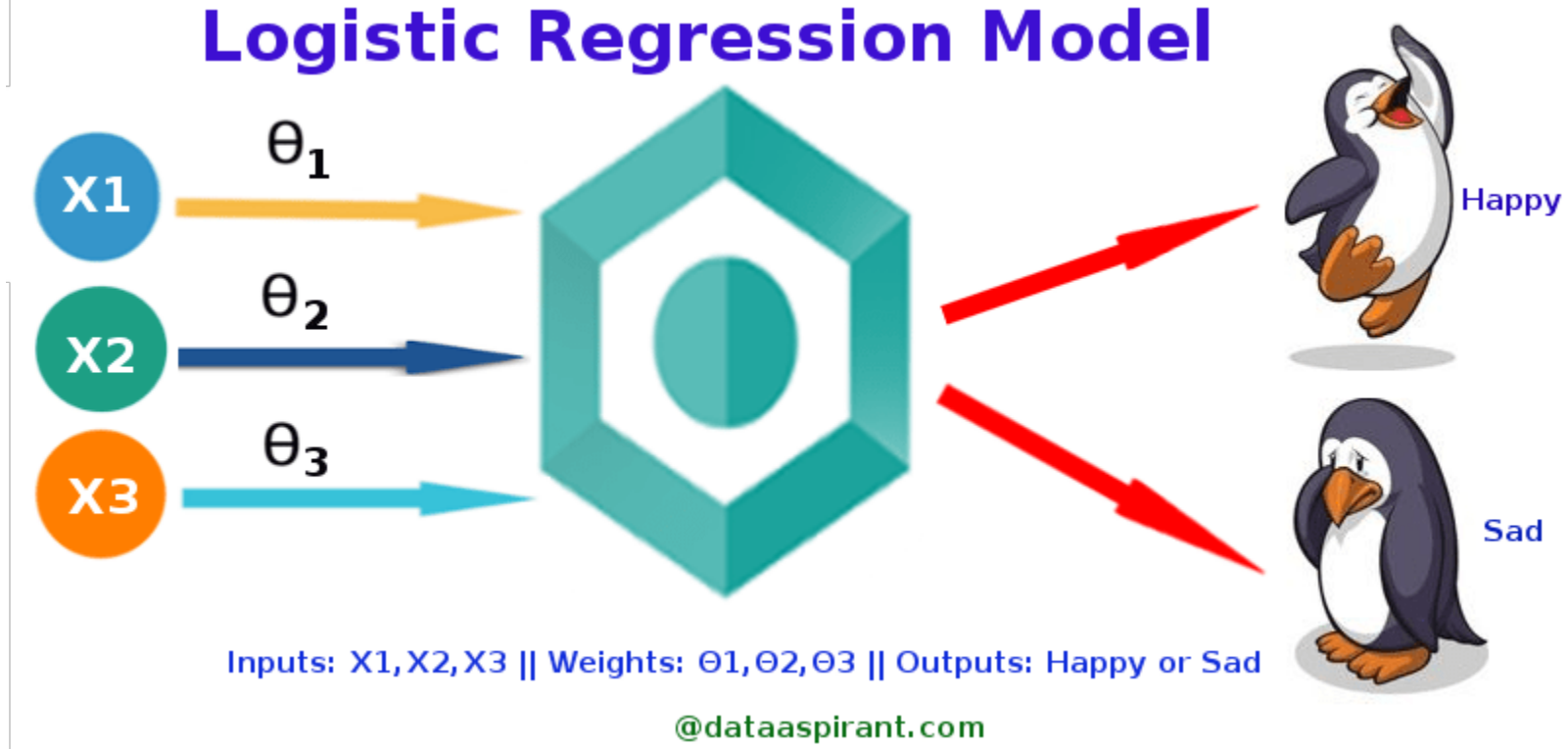
  Hypothesis can be larger than 1 or smaller than zero

  Hence, we have to use logistic regression



Types of Logisitc Regression?

- Binary Logistic Regression
  - The categorical response has only two 2 possible outcomes. Example: Spam or Not
- Multinomial Logistic Regression
  - Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)
- Ordinal Logistic Regression
  - Three or more categories with ordering. Example: Movie rating from 1 to 5


Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1, X2, X3 || Weights: Θ1, Θ2, Θ3 || Outputs: Happy or Sad
@dataaspirant.com

Applications of Logistic Regression

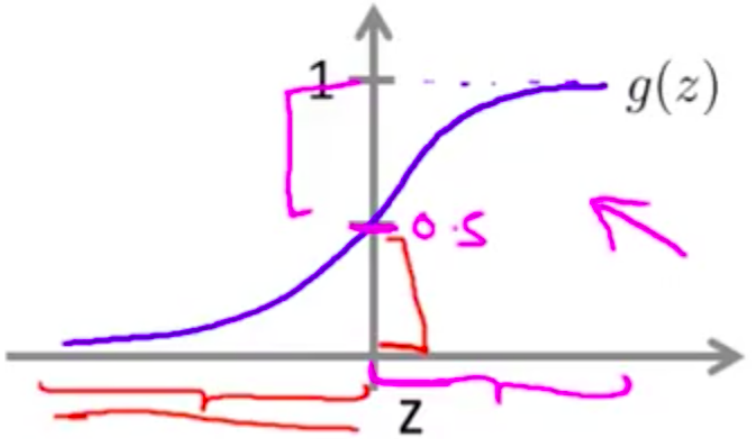
- Candidate is winning the election or not?
- Image Segmentation and Categorization
- Geographic Image Processing
- Handwriting recognition
- Healthcare : Analyzing a group of over million people for myocardial infarction within a period of 10 years is an application area of logistic regression.
- Prediction whether a person is depressed or not
- So for any binary Categorical classification you can

Decision Boundary for Logistic Regression?

Logistic regression

$h_\theta(x) = g(\theta^T x)$ = $P(y=1 \mid x; \theta)$

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



Suppose predict "$y = 1$" if $h_\theta(x) \geq 0.5$

$\theta^T x \geq 0$

$g(z) \geq 0.5$ when $z \geq 0$

$h_\theta(x) = g(\theta^T x) \geq 0.5$ whenever $\theta^T x \geq 0$

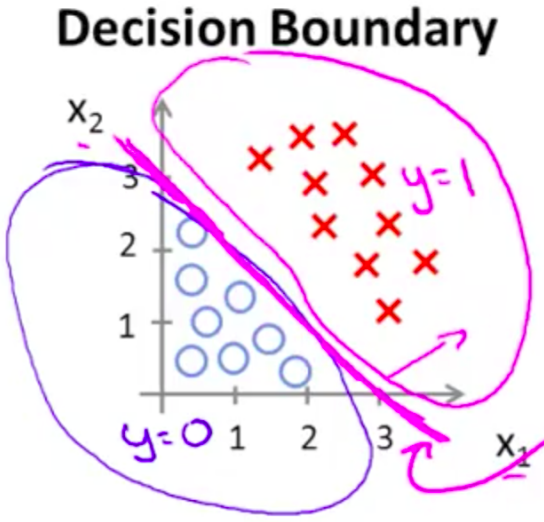
predict "$y = 0$" if $h_\theta(x) < 0.5$

$h_\theta(x) = g(\theta^T x)$

$\theta^T x < 0$

$g(z) < 0.5$

Decision Boundary for Logistic Regression?



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

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$-3 \quad 1 \quad 1$

Decision boundary

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

$\theta^T x$

$x_1 + x_2 \geq 3$

x_1, x_2

$h_\theta(x) = 0.5$

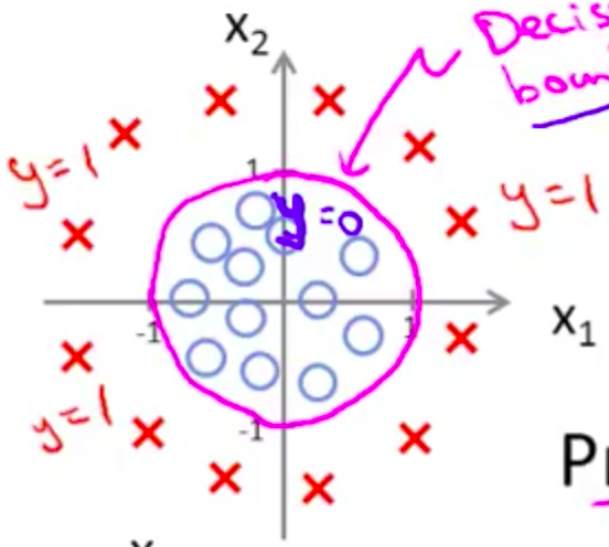
$x_1 + x_2 = 3$

$x_1 + x_2 < 3$

$y = 0$

Decision Boundary for Logistic Regression?

Non-linear decision boundaries



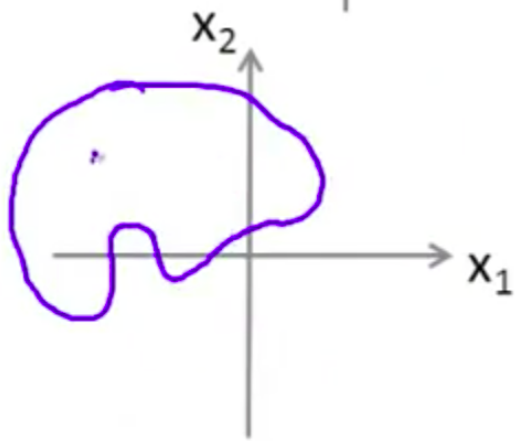
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

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

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

$x_1^2 + x_2^2 = 1$

$x_1^2 + x_2^2 \geq 1$



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \dots)$$

Logistic Regression Training

Initialize: $\theta_j = 0$ for all $0 \le j \le m$

Repeat many times:

gradient[j] = 0 for all $0 \le j \le m$

For each training example (x, y):

For each parameter j:

gradient[j] $+= x_j \left(y - \frac{1}{1 + e^{-\theta^{\mathrm{T}} \mathbf{x}}} \right)$

θ_j += η * gradient[j] for all $0 \le j \le m$

Classification with Logistic Regression

- Training: determine parameters θ_j (for all $0 \le j \le m$)
  - After parameters θ_j have been learned, test classifier
- To test classifier, for each new (test) instance ***X***:
  - Compute: $p = P(Y = 1 \mid \boldsymbol{X}) = \dfrac{1}{1 + e^{-z}}$, where $z = \theta^T \mathbf{x}$
  - Classify instance as: $\hat{y} = \begin{cases} 1 & p > 0.5 \\ 0 & \text{otherwise} \end{cases}$
  - Note about evaluation set-up: parameters θ_j are **not** updated during "testing" phase

Logistic Regression

1 Make logistic regression assumption

$$P(Y = 1|X = \mathbf{x}) = \sigma(\theta^T \mathbf{x})$$
$$P(Y = 0|X = \mathbf{x}) = 1 - \sigma(\theta^T \mathbf{x})$$

2 Calculate the log probability for all data

$$LL(\theta) = \sum_{i=0}^{n} y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log[1 - \sigma(\theta^T \mathbf{x}^{(i)})]$$

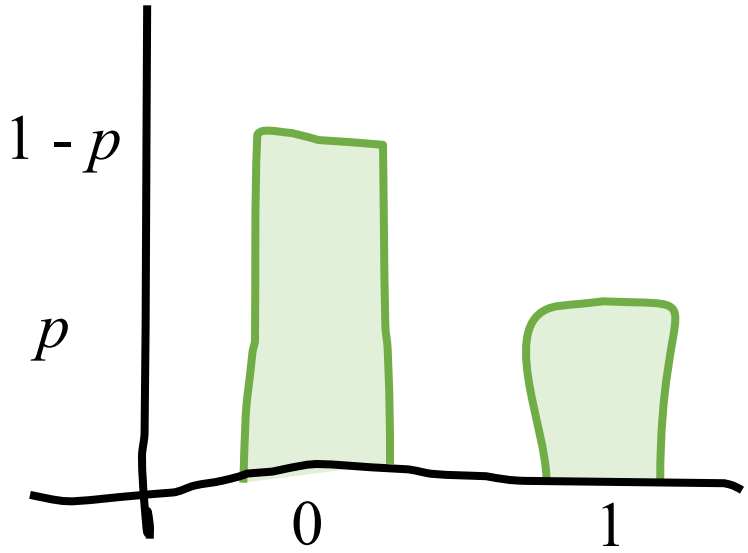
3 Get derivative of log probability with respect to thetas

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=0}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)}) \right] x_j^{(i)}$$

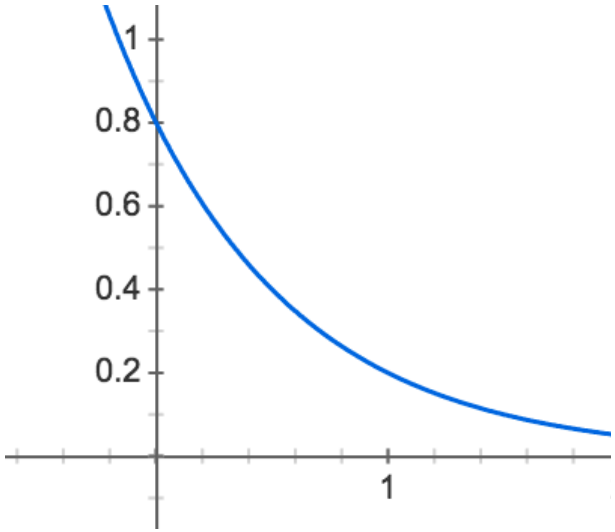
Recall: PMF of Bernoulli

- Y ~ Ber(*p*)
- Probability mass function: $P(Y = y)$

PMF of Bernoulli



PMF of Bernoulli (*p* = 0.2)



$$P(Y = y) = p^y(1-p)^{1-y}$$

$$P(Y = y) = 0.2^y(0.8)^{1-y}$$

Log Probability of Data

$$P(Y = 1|X = \mathbf{x}) = \sigma(\theta^T \mathbf{x})$$
$$P(Y = 0|X = \mathbf{x}) = 1 - \sigma(\theta^T \mathbf{x})$$

Implies

$$P(Y = y|X = \mathbf{x}) = \sigma(\theta^T \mathbf{x})^y \cdot \left[1 - \sigma(\theta^T \mathbf{x})\right]^{(1-y)}$$

For IID data

$$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)}|X = \mathbf{x}^{(i)})$$
$$= \prod_{i=1}^{n} \sigma(\theta^T \mathbf{x}^{(i)})^{y^{(i)}} \cdot \left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right]^{(1-y^{(i)})}$$

Take the log

$$LL(\theta) = \sum_{i=0}^{n} y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log[1 - \sigma(\theta^T \mathbf{x}^{(i)})]$$

Sigmoid has a Beautiful Slope

$$\frac{\partial}{\partial \theta_j}\sigma(\theta^T x)?$$

$$\frac{\partial}{\partial z}\sigma(z) = \sigma(z)[1 - z]$$

True fact about sigmoid functions

$$\frac{\partial}{\partial \theta_j}\sigma(\theta^T x) = \frac{\partial}{\partial z}\sigma(z) \cdot \frac{\partial z}{\partial \theta_j}$$

Chain rule!

$$\frac{\partial}{\partial \theta_j}\sigma(\theta^T x) = \sigma(\theta^T x)[1 - \sigma(\theta^T x)]x_j$$

Plug and chug

Sigmoid, you should be a ski hill

Gradient Update

$$LL(\theta) = \sum_{i=0}^{n} y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log[1 - \sigma(\theta^T \mathbf{x}^{(i)})]$$

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} y \log \sigma(\theta^T \mathbf{x}) + \frac{\partial}{\partial \theta_j} (1 - y) \log[1 - \sigma(\theta^T \mathbf{x}]$$

Imagine only one data point

$$= \left[\frac{y}{\sigma(\theta^T x)} - \frac{1 - y}{1 - \sigma(\theta^T x)} \right] \frac{\partial}{\partial \theta_j} \sigma(\theta^T x)$$

$$= \left[\frac{y}{\sigma(\theta^T x)} - \frac{1 - y}{1 - \sigma(\theta^T x)} \right] \frac{\partial}{\partial \theta_j} \sigma(\theta^T x)$$

$$= \left[\frac{y - \sigma(\theta^T x)}{\sigma(\theta^T x)[1 - \sigma(\theta^T x)]} \right] \sigma(\theta^T x)[1 - \sigma(\theta^T x)] x_j$$

$$= \left[y - \sigma(\theta^T x) \right] x_j$$

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=0}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)}) \right] x_j^{(i)}$$

For many data points

Thank you