Logistic Regression

Ratchainant Thammasudjarit, Ph.D.

Learning Objectives

- Understand the concepts of classification problem
- Understand the theory and applications of Logistic Regression
- Understand the classification model evaluation
- Understand how to build classifier and conduct machine learning experiment using sklearn

Classification Problems

- Classification is the task of categorization in which ideas or objects are
  - Recognizable
  - Differentiable
  - Understandable

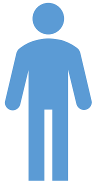
Classification Problems

- Churn Prediction
  - Will this customer decide to stay or leave our business

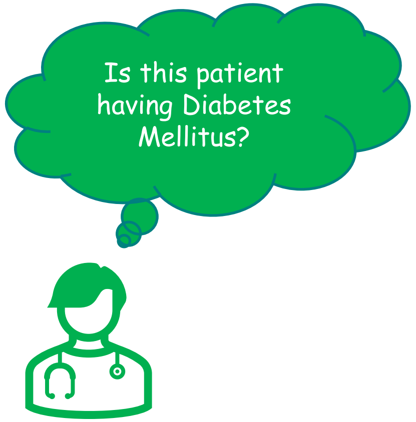
AccTypes	nComplaints
Premium	5

Classification Problems

- Disease Detection
  - Is this patient healthy or having diabetes

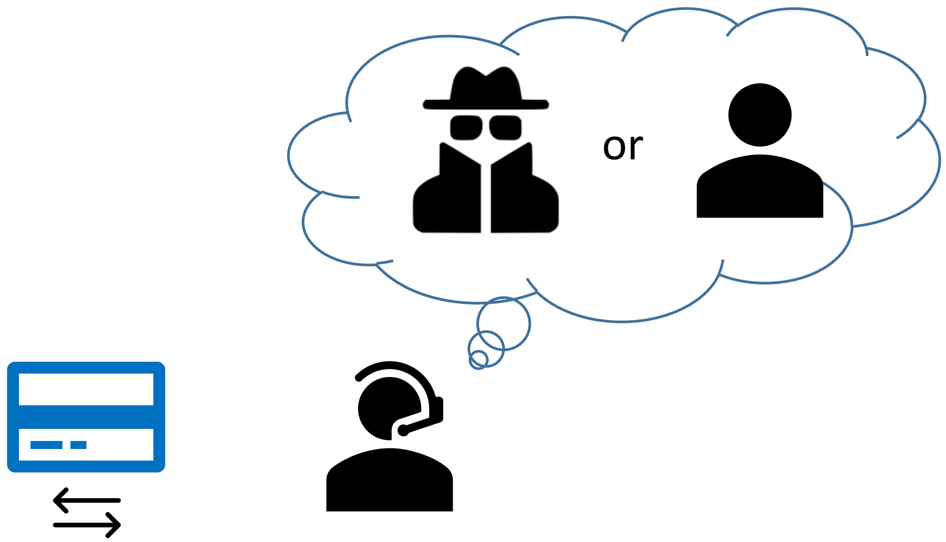


BMI	Sex	LDL
35	Male	133



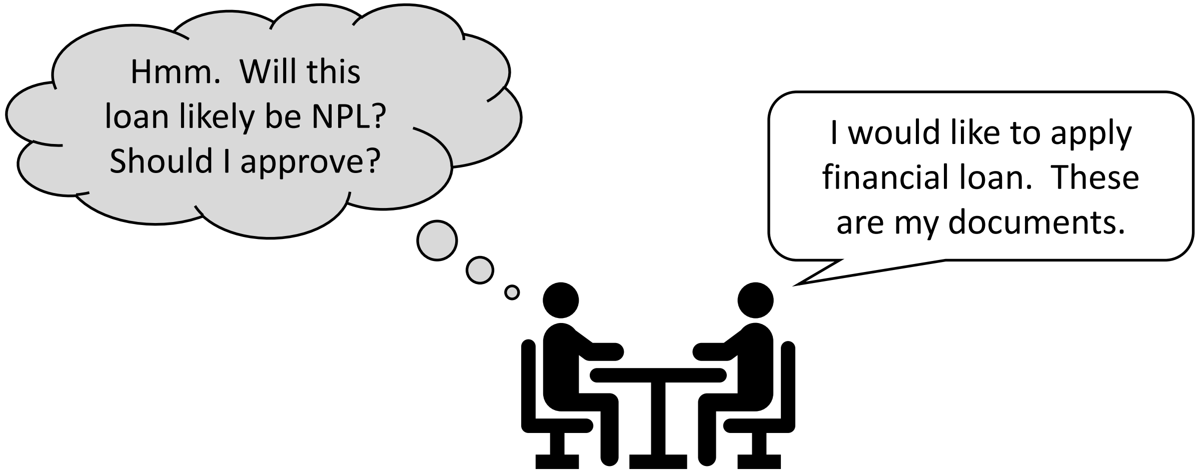
Classification Problems

- Fraud Detection
  - Is this a normal or fraudulent transaction



Classification Problems

- Bank Loan
  - Is this application ended up with NPL



Classification Problems

Notations

Let $\mathcal{D} = \langle \mathbf{X}, \mathbf{y} \rangle$ be a dataset.

$\mathbf{X} = \langle \mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_m \rangle$ be n feature vectors with dimension $n \times m$.

$\mathbf{x}_j$ be a column vector with dimension $n \times 1$ that represents the feature j where $j = 1, 2, \dots, m$.

$\mathbf{x}_i$ is the row vector with dimension $1 \times m$ that represents the feature vector i where $i = 1, 2, \dots, n$.

$\mathbf{y}$ be a column vector with dimension $n \times 1$ thar represents the target class

y be any possible value in $\mathbf{y}$

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{y}$	
	3	yes	no	
	5	no	yes	
$\mathbf{X}$	…		…	$\mathbf{y}$
	0	yes	no	

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{y}$	
$\mathbf{x}$	3	yes	no	y
	5	no	yes	
	…			
	0	yes	no	

Example: Disease Detection
Smoking Frequency Weekly $(\mathbf{x}_1)$: $x \in I^0$ or $x \in I^+$
Chest pain $(\mathbf{x}_2)$: $x \in \{\text{yes}, \text{no}\}$
Lung Cancer $(\mathbf{y})$: $y \in \{\text{yes}, \text{no}\}$

Learning

- Goal: (For classification task) To learn a mapping function from $\mathbf{x}$ to y
- Two types of machine learning
  - Parametric Machine Learning
  - Non-parametric Machine Learning

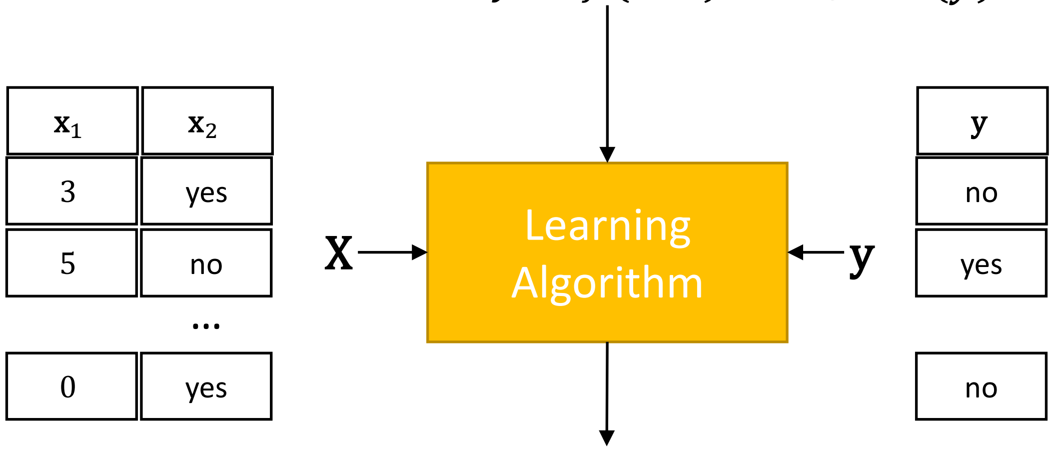
Learning

Parametric Machine Learning

Given $\langle \mathbf{X}, \mathbf{y} \rangle$, learn the optimum parameters $\boldsymbol{\theta}^*$ of a model $y = f(\mathbf{x}; \theta)$

- Example: Logistic Regression

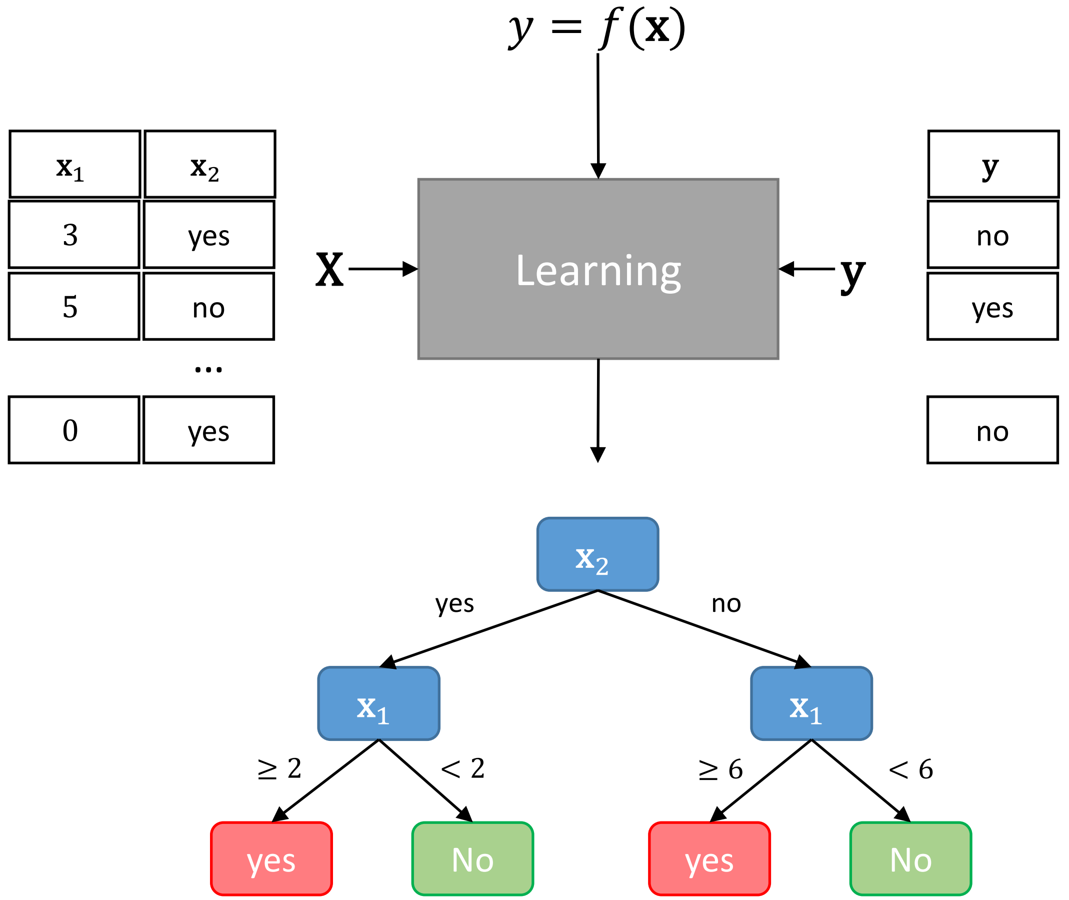
$$y = f(\mathbf{x}; \boldsymbol{\theta}) \quad \text{e.g.} \quad P(y) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \cdots \theta_m x_m)}}$$



$$\boldsymbol{\theta}^* = \langle \theta_0, \theta_1, \dots, \theta_m \rangle$$

Learning

- Example: Decision Tree

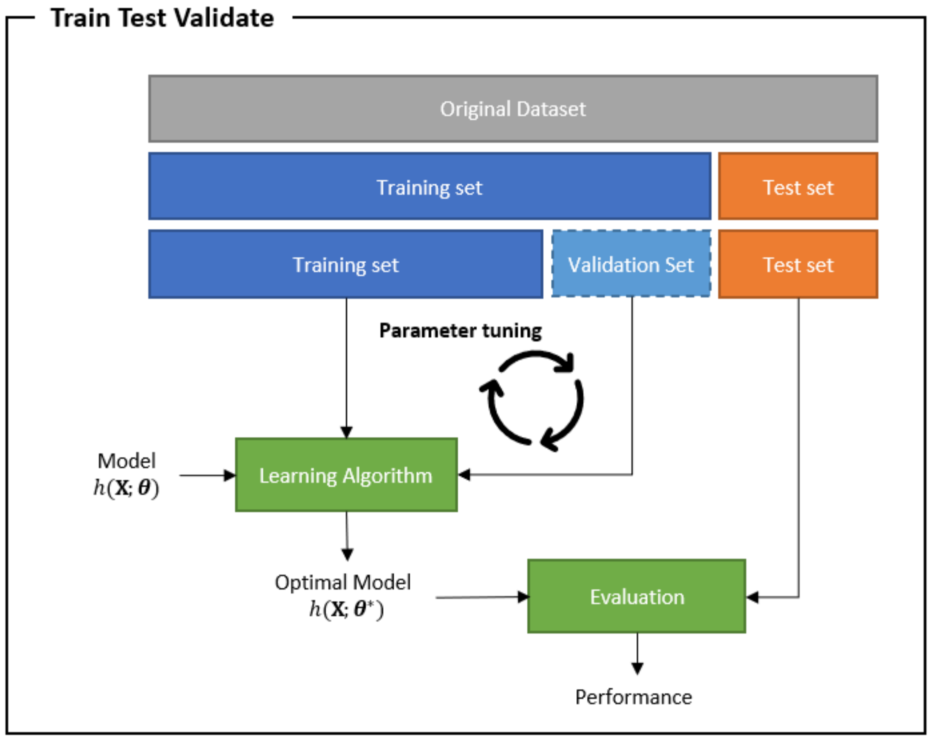


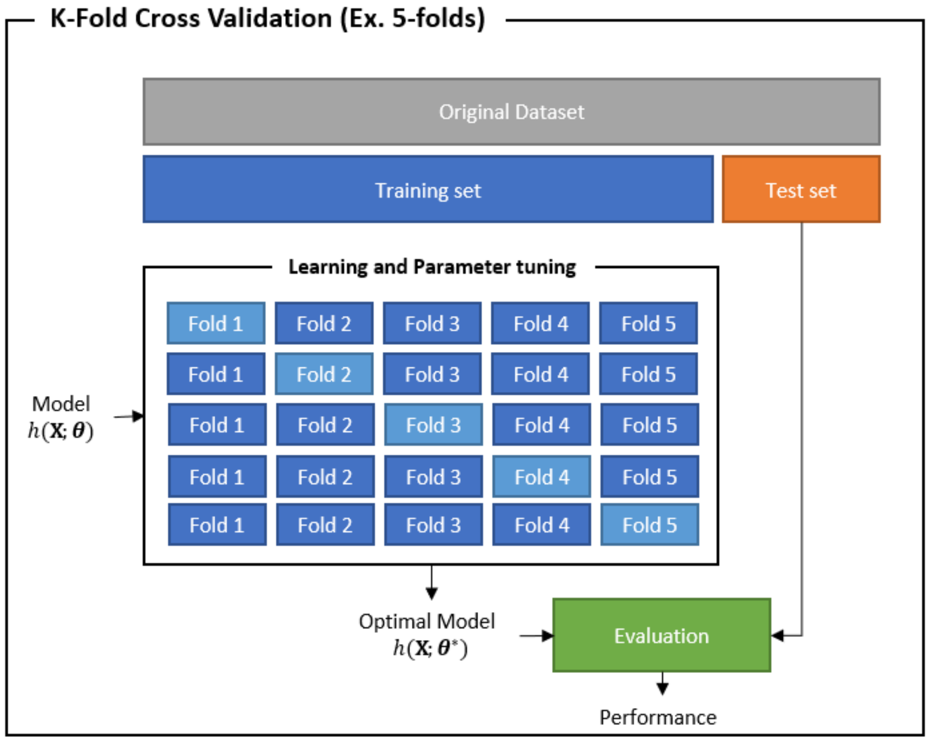
Non-parametric Machine Learning

Given $\langle \mathbf{X}, \mathbf{y} \rangle$, learn the mapping function $y = f(\mathbf{x})$

Learning

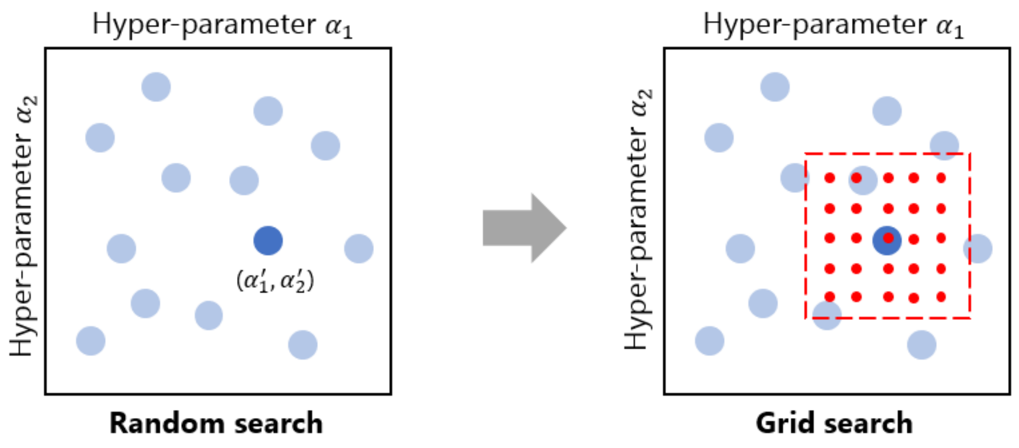
- Methodologies





Tuning

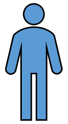
- Strategies



Prediction

- Given a sample $\mathbf{x}$ and a decision threshold τ
    - Estimate the probability of target class $P(y|\mathbf{x};\boldsymbol{\theta}^*)$ for parametric ML
    - Estimate the probability of target class $P(y|\mathbf{x})$ for non-parametric ML
    - Infer the most likely target class $\hat{y}$

Example: Disease detection from Smoking frequency weekly $(\mathbf{x}_1)$ and Chest pain or not $(\mathbf{x}_2)$



x_1	x_2
2	yes

$$P(y|\mathbf{x};\boldsymbol{\theta}^*) = \frac{1}{1+e^{-(0.1+0.45x_1+0.2x_2)}}$$

$$\therefore P(y|\mathbf{x};\boldsymbol{\theta}^*) = \frac{1}{1+e^{-(0.1+0.45(2)+0.2(1))}}$$

$$= 0.769$$

Given decision threshold (τ): 0.5

$$P(y|\mathbf{x};\boldsymbol{\theta}^*) \geq \tau$$

$$\therefore \hat{y} = \text{yes}$$

Inference: This patient is highly likely to have lung cancer.

Types of Classification Problem

- There are 3 types of classification problem
  - Binary Classification: There are 2 possible values for y which are **mutually exclusive**
  - Multiclass Classification: There are more than 2 possible values for y which are **mutually exclusive**
  - Multilabel Classification: There are at least 2 possible values for y which are **NOT mutually exclusive**

Types of Classification Problem

- Binary Classification

$\mathbf{x}_1$	…	$\mathbf{x}_j$	…	$\mathbf{x}_m$	$\mathbf{y}$
$x_{1,1}$	…	$x_{1,j}$	…	$x_{1,m}$	y_1
…	…	…	…	…	…
$x_{i,1}$	…	$x_{i,j}$	…	$x_{i,m}$	y_i
…	…	…	…	…	…
$x_{n,1}$	…	$x_{n,j}$	…	$x_{n,m}$	y_n

Example: Disease detection
A customer decision is either sick or healthy, e.g., $\mathbf{y} = \langle y | y \in \{yes, no\} \rangle$

Types of Classification Problem

- Multiclass Classification

$\mathbf{x}_1$	…	$\mathbf{x}_j$	…	$\mathbf{x}_m$	$\mathbf{y}$
$x_{1,1}$	…	$x_{1,j}$	…	$x_{1,m}$	y_1
…	…	…	…	…	…
$x_{i,1}$	…	$x_{i,j}$	…	$x_{i,m}$	y_i
…	…	…	…	…	…
$x_{n,1}$	…	$x_{n,j}$	…	$x_{n,m}$	y_n

Example: Flower classification
A flower is only one of k possible species, e.g., $\mathbf{y} = \langle y | y \in \{\text{Serosa}, \text{Versicolor}, \text{Virginica}\} \rangle$

Types of Classification Problem

- Multilabel Classification

$\mathbf{x}_1$	…	$\mathbf{x}_j$	…	$\mathbf{x}_m$	$\mathbf{y}_1$	…	$\mathbf{y}_k$
$x_{1,1}$	…	$x_{1,j}$	…	$x_{1,m}$	$y_{1,1}$	…	$y_{1,k}$
…	…	…	…	…	…	…	…
$x_{i,1}$	…	$x_{i,j}$	…	$x_{i,m}$	$y_{i,1}$	…	$y_{i,k}$
…	…	…	…	…	…	…	…
$x_{n,1}$	…	$x_{n,j}$	…	$x_{n,m}$	$y_{n,1}$	…	$y_{n,k}$

Example: Lesion detection
One chest X-ray is possible to detect up to k radiology findings, e.g., Fibrosis, Edema, Cardiomegaly, …, etc.

$$\mathbf{y}_1 = \langle y|y \in \{0,1\}\rangle$$

$$\mathbf{y}_2 = \langle y|y \in \{0,1\}\rangle$$

…

$$\mathbf{y}_k = \langle y|y \in \{0,1\}\rangle$$

Summary

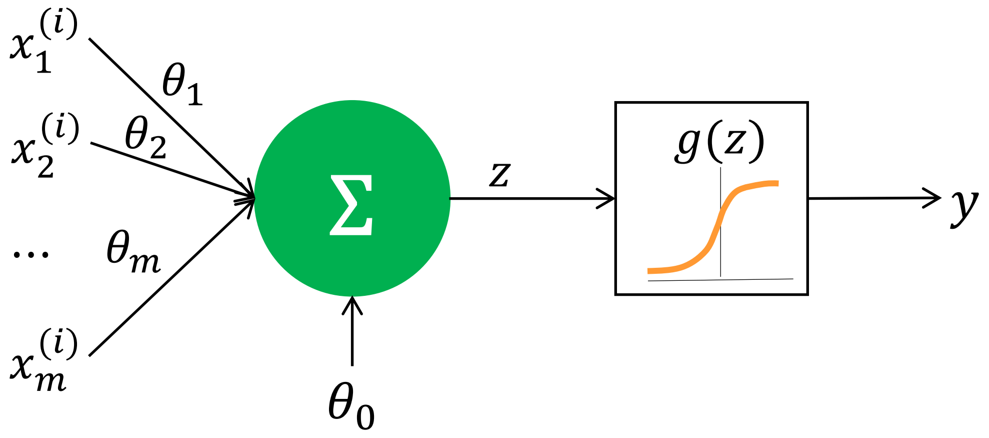
- The key concepts to take away
  - Classification is the task to predict categorical values
  - Parametric Machine Learning Algorithm learns the optimum model parameters $\boldsymbol{\theta}^*$ for $y = f(\mathbf{x}; \boldsymbol{\theta})$
  - Non-parametric Machine Learning Algorithm learns the mapping function $y = f(\mathbf{x})$
  - Classification attempt to predict $P(y|\mathbf{x}; \boldsymbol{\theta}^*)$
  - Use a decision threshold τ to infer $\hat{y}$ from $P(y|\mathbf{x}; \boldsymbol{\theta}^*)$
  - The type of classification problem depends on the target class $\mathbf{y}$

Logistic Regression

- Logistic Regression was originally designed to solve binary classification problems
    - Will this customer leave our business (Churn or Stay)
    - Is this patient healthy (Sick or Healthy)
    - Is this loan application profitable (NPL, non-NPL)
    - Etc.

Logistic Regression

- Model
  - A feature is either $\mathbf{x}_j = \langle x | x \in \mathbb{R} \rangle$ or categorical variable
  - The target class $\mathbf{y} = \langle y | y \in \{0,1\} \rangle$
  - $z = \boldsymbol{\theta} \cdot \mathbf{x}$ is the dot product between model parameters $\boldsymbol{\theta}$ and feature vector $\mathbf{x}$
  - $g(z)$ is the logistic function (a.k.a sigmoid)



Logistic Regression

Linear Combination

$\theta_1 x_1 + \cdots + \theta_m x_m$

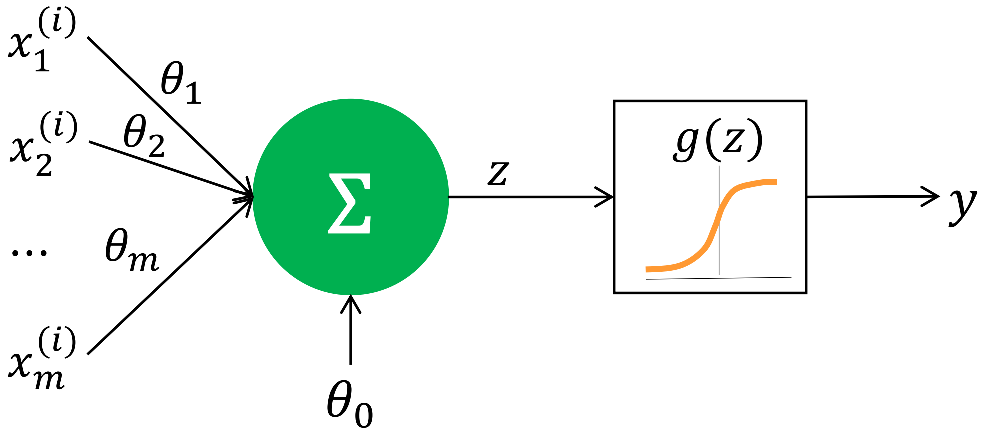
Interception

θ_0

Parameters

$\theta = \langle \theta_0, \theta_1, \ldots, \theta_m \rangle$

$$z = \theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m$$



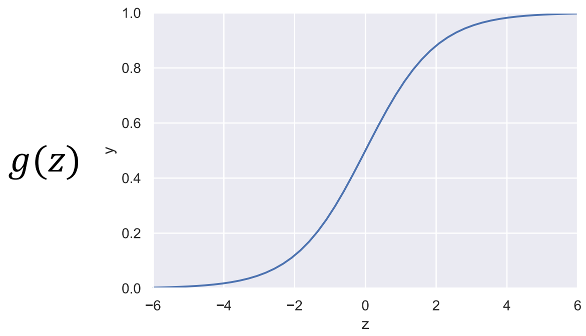
Training Data

$\mathbf{x}_{smk}$	$\mathbf{x}_{cp}$	y_{lc}
3	1	0
5	0	1
	…	
2	1	0

$$P(y|\mathbf{x};\boldsymbol{\theta}) = \frac{1}{1+e^{-z}}$$

$$= \frac{1}{1+e^{-(\theta_0+\theta_{smk}\mathbf{X}_{smk}+\theta_{cp}\mathbf{X}_{cp})}}$$

Logistic Function (Sigmoid)



Logistic Regression

- A sample use case

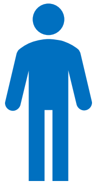
$\boldsymbol{\theta}^*$

$\theta_0 = 0.2$

$\theta_{cmp} = 0.1$

$\theta_{prm} = 0.5$

$$P(y|\mathbf{x};\boldsymbol{\theta}^*) = \frac{1}{1 + e^{-(0.2 + 0.1\mathbf{X}_{smk} + 0.5\mathbf{X}_{cp})}}$$



$smk = 3$

$cp = yes$

$$P(y|\mathbf{x};\boldsymbol{\theta}^*) = \frac{1}{1 + e^{-(0.2 + 0.1(3) + 0.5(1))}}$$

$$= 0.73$$

This patient is more likely to have lung cancer. Treatment must be given.

Logit Transformation

- The original form of logistic regression

$$P(y|\mathbf{x};\boldsymbol{\theta}) = \frac{1}{1 + e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}} \qquad (1)$$

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

- The odds of an event x is defined as follows

Logit Transformation

- Rewrite the original form of logistic regression

$$1 - P(y|\mathbf{x};\boldsymbol{\theta}) = 1 - \frac{1}{1 + e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}}$$

$$= \frac{e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}}{1 + e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}} \qquad (2)$$

$$\frac{P(y|\mathbf{x};\boldsymbol{\theta})}{1 - P(y|\mathbf{x};\boldsymbol{\theta})} = \frac{\dfrac{1}{1 + e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}}}{\dfrac{e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}}{1 + e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}}}$$

- Divide (1) by (2)

Logit Transformation

$$\mathrm{Odds}(y|\mathbf{x};\boldsymbol{\theta}) = \frac{1}{e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}}$$

$$= e^{(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)} \qquad (3)$$

The inverse of the standard logistic regression.

The logit model also called log-odds since it is equal to the logarithm of the odds p/(1-p) where p is the probability.

Therefore, logit is a function that maps probability values from (0, 1) to real numbers (-inf, inf).

$$\ln \mathrm{Odds}(y|\mathbf{x};\boldsymbol{\theta}) = \ln e^{(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}$$

$$\mathrm{logit} P(y|\mathbf{x};\boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m \qquad (4)$$

- Take the natural logarithmic function to (3)

Logit Transformation

- The model parameters of logistic regression are $\theta_0, \theta_1, \dots, \theta_m$
- The θ_0 can be interpreted in 2 ways
  - θ_0 indicates the $\ln \mathrm{Odds}(y|\mathbf{x}; \boldsymbol{\theta}^*)$ when all features are equal to zero the result lies in the decision boundary since theta = 0
  - θ_0 indicates the baseline natural log Odds when all features are unknown or ignored
- Interpretation of any θ_j for $j = 1, 2, \dots, m$ depends on the data type

  The weight of the feature, example: y=2+1.5x_age. This can be interpreted as 1.5 is the weight where y increases with 1.5 since it is directly proportional to y
  - If $\mathbf{x}_j$ is a numerical feature, the θ_j indicates its contribution to the change of $\ln \mathrm{Odds}(y|\mathbf{x}; \boldsymbol{\theta}^*)$

  The model creates a decision boundary. For instance, if the patient is male (1) then all males have 1. The other values will be on the other side of the decision boundary.
  - If $\mathbf{x}_j$ is a categorical feature, the θ_j indicates the difference of $\ln \mathrm{Odds}(y|\mathbf{x}; \boldsymbol{\theta}^*)$ compared to the **baseline of $\mathbf{x}_j$**

Note: The baseline of $\mathbf{x}_j$ is any value x encoded as zero for dummy variable of $\mathbf{x}_j$

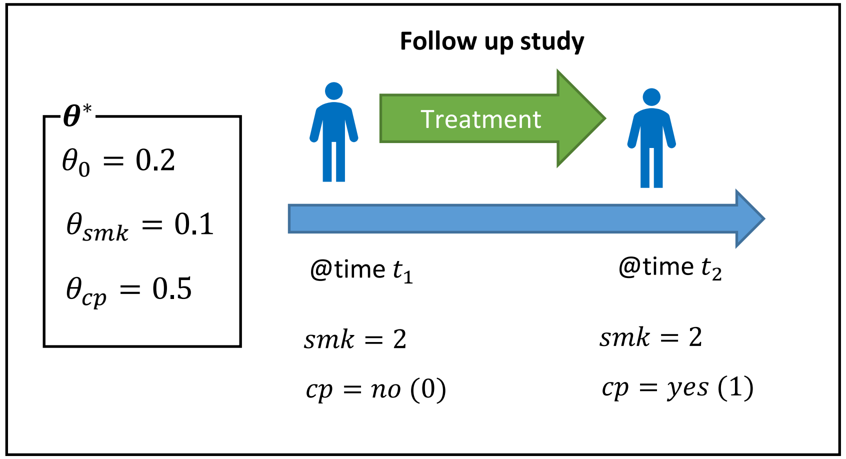
Risk Ratio

- Given a feature $\mathbf{x}_j$ while **other features were known and fixed**
  - Risk Ratio (RR) describes the difference contribution to the outcome for $\mathbf{x}_j$

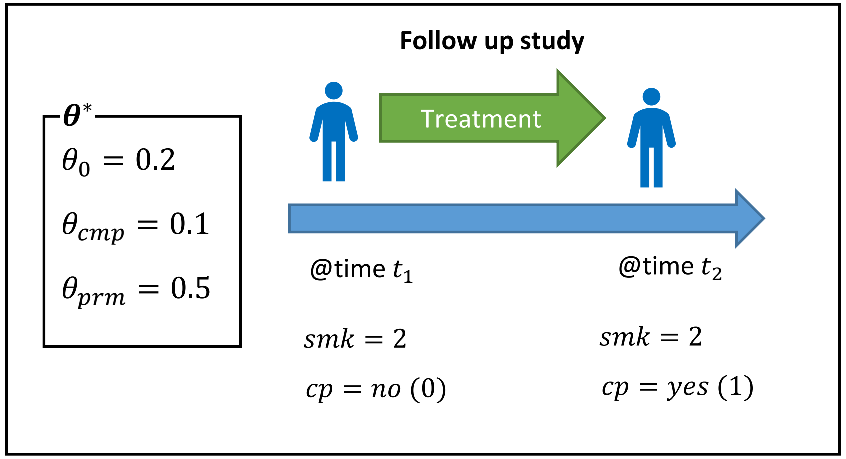
Example:

A patient without chest pain smokes 2 times a week

Recently this patient found himself with chest pain



Risk Ratio



- The risk of lung cancer at time t_1 is

$$P(y = \text{yes}|\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{1 + e^{0.2-0.1(2)-0.5(0)}} = 0.598$$

- The risk of lung cancer at time t_2 is

$$P(y = \text{yes}|\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{1 + e^{0.2-0.1(2)-0.5(1)}} = 0.711$$

- The estimated risk ratio $\left(\widehat{RR}\right)$ is

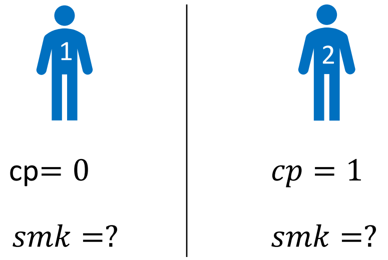
$$\widehat{RR} = \frac{P(y = \text{yes}|\text{smk=2}, cp = 1)}{P(y = \text{yes}|\text{smk=2}, cp = 0)} = \frac{0.711}{0.598} = 1.189$$

Therefore, the patient with chest pain have 1.189 times risk to lung cancer compared to patient without chest pain assuming all patients smokes 2 times a week

Odds Ratio

- The odds ratio compares between 2 group based on a feature $\mathbf{x}_j$ where **other features were fixed but unspecified**

  Example: What is the effect of chest pain to lung cancer considering patients who have the same smoking frequency regardless its values?



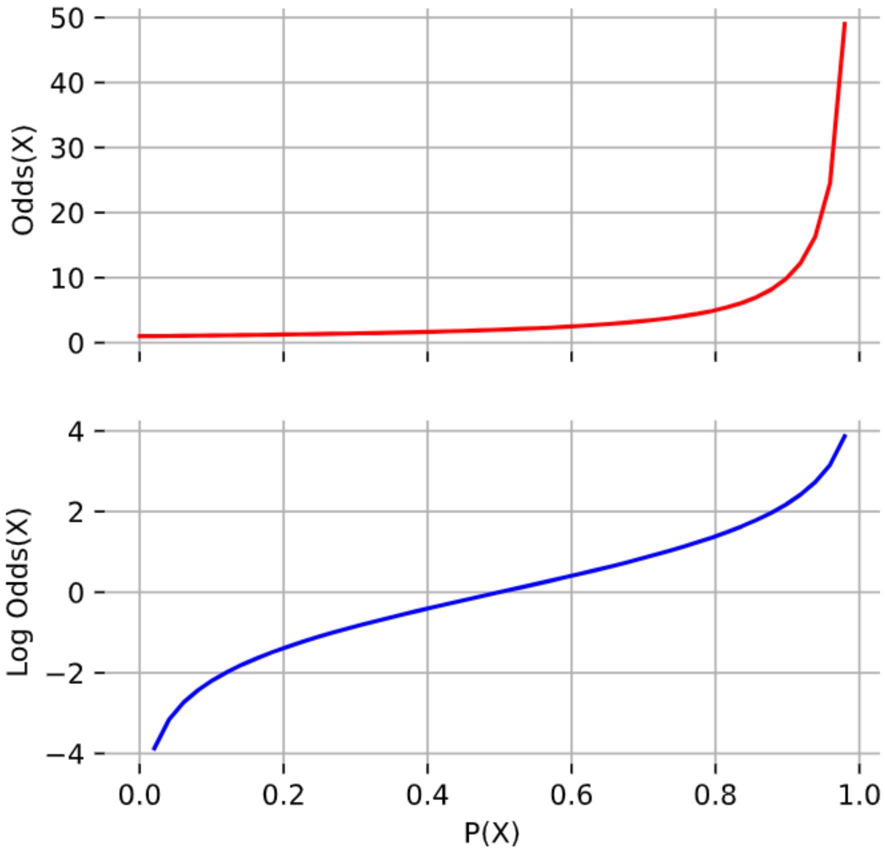
smk **were the same but unspecified**

Odds Ratio

- Relationship between Odds and Probability
  - Let x be an event

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

$$P(x) = \frac{\mathrm{Odds}(x)}{1 + \mathrm{Odds}(x)}$$



Odds Ratio

- Recall the logit transformation

$$\mathrm{logit}P(y|\mathbf{x};\boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

- Each θ_j indicates the different of natural log odds compared to the baseline

Example: $\mathrm{logit}P(y|\mathbf{x};\boldsymbol{\theta}) = 0.2 + 0.1\mathbf{x}_{smk} + 0.5\mathbf{x}_{cp}$

  - The odd ratio of premium users to the ordinary users is

$$\widehat{OR} = e^{\theta_{cp}}$$

$$= e^{0.5}$$

$$= 1.648$$

Therefore, the odds to lung cancer of patient with chest pain is 1.648 times higher than patient without chest pain assuming those patient have the same smoking frequency.

Maximum Likelihood Estimation

- Consider a coin flipping problem, flipping a coin one time is called an **experiment**

- Let θ be a parameter of getting Head (H) from a single experiment
- Let X be a random variable that represents the outcome of coin flipping
- $P(x;\theta)$ represents the probability of outcome $x \in \{H, T\}$ based on the coin parameter

Maximum Likelihood Estimation

- If a coin is fair ($\theta = 0.5$)
  - The probability of getting Head $P(x = H; \theta) = 0.5$
  - Simplify version for the probability notation is $P(H; \theta) = 0.5$
- If a coin is unfair, e.g., $\theta = 0.7$
  - The probability of getting Tail $P(x = T; \theta) = 1 - 0.7 = 0.3$
  - Simplify version for the probability notation is $P(T; \theta) = 0.3$

Maximum Likelihood Estimation

- Let $\mathcal{L}_{\mathbf{x}}(\theta)$ be the likelihood of parameter (θ) given observations $(\mathbf{x})$

  Example: Flipping a coin 5 times yields H, H, T, H, T

  Q: How much likely does that coin bias to H

  Let $x = 1$ be the Head, $x = 0$ be the Tail

  The likelihood function of flipping a coin 1 time follows Bernoulli Distribution as follows

$$\begin{aligned}\mathcal{L}_{\mathbf{x}}(\theta) &= P(x; \theta) \\ &= \theta^{x}(1-\theta)^{1-x}\end{aligned}$$

Maximum Likelihood Estimation

- Assume that each flipping is independent each other
  - The likelihood function of flipping a coin n times is defined as

$$
\begin{aligned}
\mathcal{L}_{\mathbf{x}}(\theta) &= P(\mathbf{x};\theta) \\
&= P(x_1;\theta)P(x_2;\theta)\dots P(x_n;\theta) \\
&= \prod_{i=1}^{n} P(x_i;\theta) \\
&= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)}
\end{aligned}
$$

Maximum Likelihood Estimation

- The likelihood function of observations H, H, T, H, T is

$$\mathcal{L}_{\mathbf{X}}(\theta) = \theta^{1}(1-\theta)^{(1-1)} \cdot \theta^{1}(1-\theta)^{(1-1)} \cdot \theta^{0}(1-\theta)^{(1-0)} \cdot \theta^{1}(1-\theta)^{(1-1)} \cdot \theta^{0}(1-\theta)^{(1-0)}$$

$$= \theta \cdot \theta \cdot (1-\theta) \cdot \theta \cdot (1-\theta)$$

$$= \theta^{3}(1-\theta)^{2}$$

- Apply logarithmic function on both side does not change the θ that maximizes $\mathcal{L}_{\mathbf{X}}(\theta)$
- Let $L_{\mathbf{X}}(\theta)$ be the log-likelihood of θ

$$\ln \mathcal{L}_{\mathbf{X}}(\theta) = \ln \theta^{3}(1-\theta)^{2}$$

$$L_{\mathbf{X}}(\theta) = 3 \ln \theta + 2 \ln(1-\theta)$$

Maximum Likelihood Estimation

- Finding the θ that maximize $L_{\mathbf{X}}(\theta)$ requires the 1st order derivative with respect to θ and set it to zero

$$\frac{d}{d\theta} L_{\mathbf{X}}(\theta) = 0$$

$$\frac{d}{d\theta} 3 \ln \theta + \frac{d}{d\theta} 2 \ln(1-\theta) = 0$$

$$\frac{3}{\theta} - \frac{2}{1-\theta} = 0$$

$$\therefore \theta = \frac{3}{5}$$

Maximum Likelihood Estimation

- Applying the Maximum Likelihood Estimation (MLE) to Logistic Regression
  - For any sample $\mathbf{x}_i$ that has training label equals to 1

$$P(y = 1|\mathbf{x}; \boldsymbol{\theta}) = f(\theta^T \cdot \mathbf{x})$$

  - For any sample $\mathbf{x}_i$ that has training label equals to 0

$$P(y = 0|\mathbf{x}; \boldsymbol{\theta}) = 1 - f(\theta^T \cdot \mathbf{x})$$

where

$$f(\theta^T \cdot \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \cdot \mathbf{x}}}$$

Maximum Likelihood Estimation

- For any training label y

$$P(y|\mathbf{x};\boldsymbol{\theta}) = f(\boldsymbol{\theta}^T \cdot \mathbf{x})^y \cdot (1 - f(\boldsymbol{\theta}^T \cdot \mathbf{x}))^{(1-y)}$$

- For the dataset of n samples, assume that each sample is independent
  - The likelihood function is defined as

$\mathbf{x}_1$	...	$\mathbf{x}_j$	...	$\mathbf{x}_m$	$\mathbf{y}$
$x_{1,1}$	...	$x_{1,j}$	...	$x_{1,m}$	$y^{(1)}$
...	...	...	...	...	...
$x_{i,1}$	...	$x_{i,j}$	...	$x_{i,m}$	$y^{(i)}$
...	...	...	...	...	...
$x_{n,1}$	...	$x_{n,j}$	...	$x_{n,m}$	$y^{(n)}$

$$\mathcal{L}_{\mathbf{X}}(\boldsymbol{\theta}) = P(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})$$

$$= \prod_{i=1}^{n} f(\boldsymbol{\theta}^T \cdot \mathbf{x})^{y^{(i)}} \cdot (1 - f(\boldsymbol{\theta}^T \cdot \mathbf{x}))^{(1-y^{(i)})}$$

Maximum Likelihood Estimation

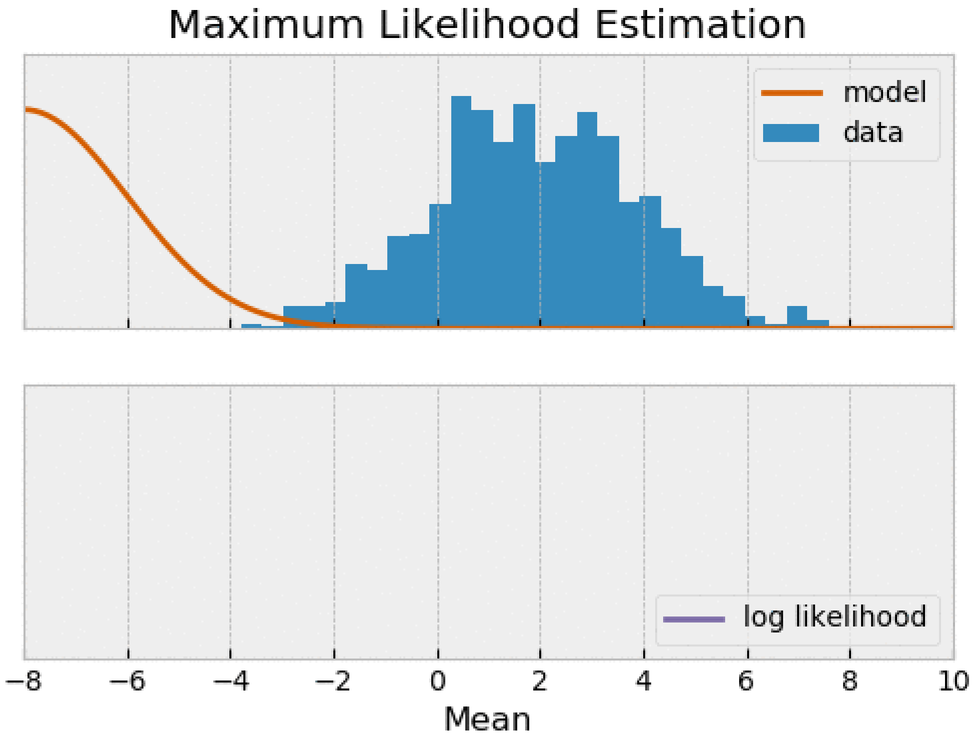
- From the likelihood function, the log-likelihood function is

$$
\begin{aligned}
L_{\mathbf{x}}(\boldsymbol{\theta}) &= \ln \prod_{i=1}^{n} f(\boldsymbol{\theta}^T \cdot \mathbf{x})^{y^{(i)}} \cdot (1 - f(\boldsymbol{\theta}^T \cdot \mathbf{x}))^{(1-y^{(i)})} \\
&= \sum_{i=1}^{n} \ln \left[f(\boldsymbol{\theta}^T \cdot \mathbf{x})^{y^{(i)}} \cdot (1 - f(\boldsymbol{\theta}^T \cdot \mathbf{x}))^{(1-y^{(i)})} \right] \\
&= \sum_{i=1}^{n} \left[\ln f(\boldsymbol{\theta}^T \cdot \mathbf{x})^{y^{(i)}} + \ln(1 - f(\boldsymbol{\theta}^T \cdot \mathbf{x}))^{(1-y^{(i)})} \right] \\
&= \sum_{i=1}^{n} \left[y^{(i)} \ln f(\boldsymbol{\theta}^T \cdot \mathbf{x}) + (1 - y^{(i)}) \ln(1 - f(\boldsymbol{\theta}^T \cdot \mathbf{x})) \right]
\end{aligned}
$$

- Learning any θ_j requires partial derivative of $L_{\mathbf{x}}(\theta)$ with respect to θ_j and set to zero

Maximum Likelihood Estimation

- Fitting the model to data using MLE shows improvement of the log-likelihood

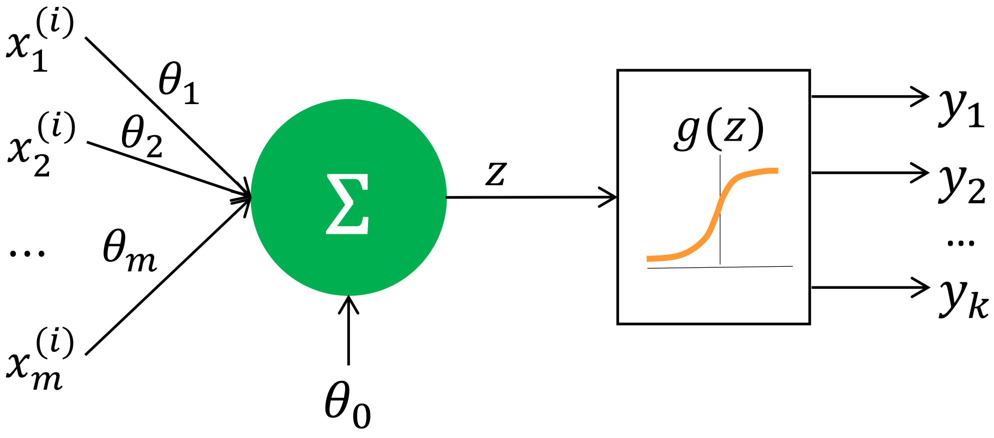


Multinomial Logistic Regression

- The original logistic regression is applicable when y is dichotomous
- Multinomial Logistic Regression (a.k.a. Softmax Regression) is designed for categorical target class

Multinomial Logistic Regression

- Model



$$\mathbf{z} = \boldsymbol{\theta} \cdot \mathbf{x}$$
$$\mathbf{t} = e^{\mathbf{z}}$$
$$\mathbf{a} = g(\mathbf{z}) = \frac{e^{\mathbf{z}}}{\sum_{i=1}^{k} t_i}$$

Example: Suppose there are 4 classes ($k = 4$)

Assume that we get $\mathbf{z}$ as follows

$$\mathbf{z} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix} \quad \therefore \mathbf{t} = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix} = \begin{bmatrix} 148.4 \\ 7.4 \\ 0.4 \\ 20.1 \end{bmatrix} \quad \therefore \sum_{i=1}^{k} t_i = 176.3$$

Note: Summation of **a** is always equal to 1

$$\therefore \mathbf{a} = \frac{\mathbf{t}}{176.3} = \begin{bmatrix} 148.4/176.3 \\ 7.4/176.3 \\ 0.4/176.3 \\ 20.1/176.3 \end{bmatrix} = \begin{bmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{bmatrix}$$

$$P(y_1|\mathbf{x};\boldsymbol{\theta}^*) = 0.842$$
$$P(y_2|\mathbf{x};\boldsymbol{\theta}^*) = 0.042$$
$$P(y_3|\mathbf{x};\boldsymbol{\theta}^*) = 0.002$$
$$P(y_4|\mathbf{x};\boldsymbol{\theta}^*) = 0.114$$

$\hat{y}$ is inferred from the maximum probability

Summary

- The key concepts to take away
  - Original logistic regression is applicable for dichotomous outcome
  - A model parameter θ_0 can be interpreted as
    - $\ln \text{Odds}(y|\mathbf{x};\boldsymbol{\theta}^*)$ when all features are equal to zero
    - Baseline natural log Odds when all features are unknown or ignored
  - A model parameter θ_j where $j = 1, 2, \dots, m$ can be interpreted as
    - Numerical feature: Contribution of $\mathbf{x}_j$ to the change of $\ln \text{Odds}(y|\mathbf{x};\boldsymbol{\theta}^*)$
    - Categorical feature: Difference of $\ln \text{Odds}(y|\mathbf{x};\boldsymbol{\theta}^*)$ compared to the baseline of $\mathbf{x}_j$

Summary

- The risk ratio describes the difference contribution to the outcome for $\mathbf{x}_j$ when **other features were known and fixed**
- The odds ratio compares between 2 group based on a feature $\mathbf{x}_j$ where **other features were fixed but unspecified**