NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

CZ3005
Artificial Intelligence

Logistic Regression

Asst/P Mahardhika Pratama

Email: mpratama@ntu.edu.sg

Office: N4-02a-08

Instructor

- I got my PhD from UNSW, Australia. I completed my PhD in 2.5 years. After graduation, I did my postdoc for few years at UTS, Australia and then worked as a faculty at Latrobe University, Australia before joining NTU. My research is in the area of autonomous learning and data stream mining. I currently serve as EIC of IJBIDM and a consultant at Lifebytes, Australia.
- My consultation time is at 5pm, Wednesday.

Tutorials

- Tutorial starts from week 10 – 12
- 3 tutorials in the second half: fuzzy logic, logical reasoning, first-order logic

Labs

- One lab in the second half
- Lab is an individual assignment
- Takes place in week 9/10
- Attendance is not compulsory

Final Grade

- 60% Final Exam + 40% Labs (Lab 1 and Lab 2)

Artificial Intelligence

- ❑ Problem Solving
- ❑ Knowledge Representation and Reasoning
- ❑ Acting Logically
- ❑ Uncertain Knowledge and Reasoning
- ❑ Learning
- ❑ Communicating, Perceiving and Acting

Outline

- ❑Classification
- ❑Hypothesis Representation
- ❑Decision Boundary
- ❑Cost Function
- ❑Optimization

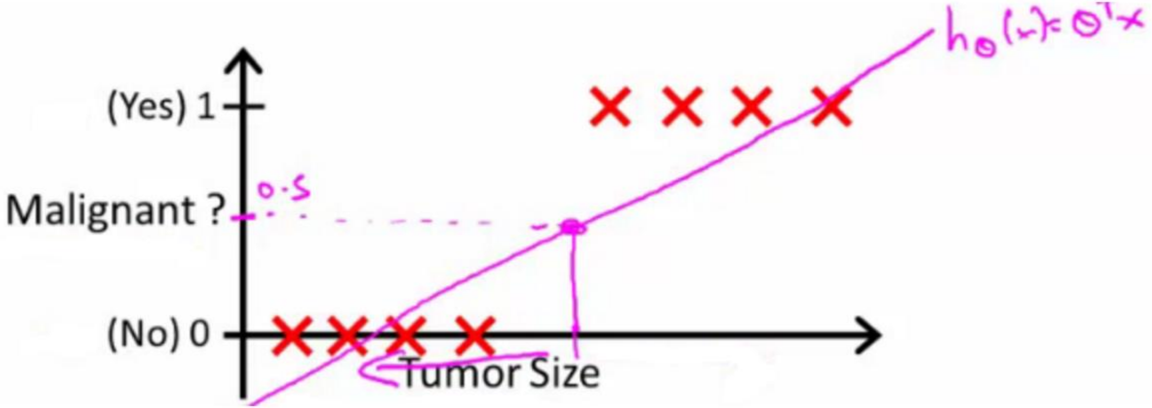
Classification

- Develop the logistic regression algorithm to determine what class a new input should fall into
- Classification problems
  - Email -> spam/not spam?
  - Online transactions -> fraudulent?
  - Tumor -> Malignant/benign
- Y is either 0 or 1
    - 0 = negative class (absence of something)
    - 1 = positive class (presence of something)

Tumour Prediction Problem

- Tumour Size vs Malignancy (0 or 1)
- We can develop linear classifier
  - Use a threshold to determine the class label
  - It seems working

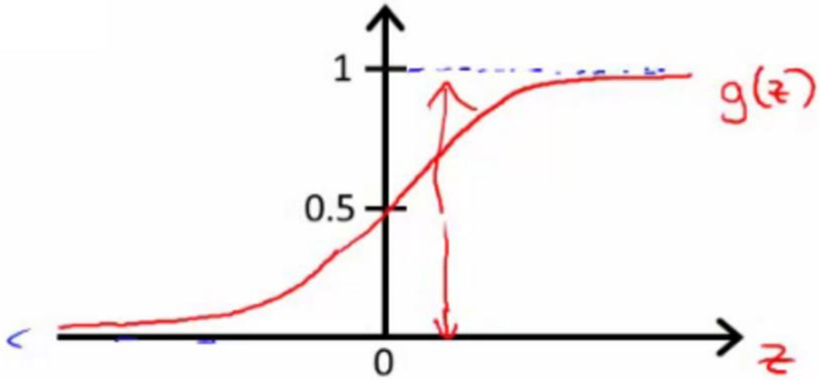
Linear Classifier



- How if we have a single yes for a very small tumour
- Output values beyond 0 or 1
- Logistic regression outputs value between 0 and 1
  - Logistic regression is for classification problem

Hypothesis Representation

- The classifier output is bounded in [0,1]
- The linear classifier : $y = \theta^T x$
- The logistic regression : we use sigmoid function
  - $y = g(\theta^T x)$where $g(z) = \frac{1}{1+e^{-z}}$

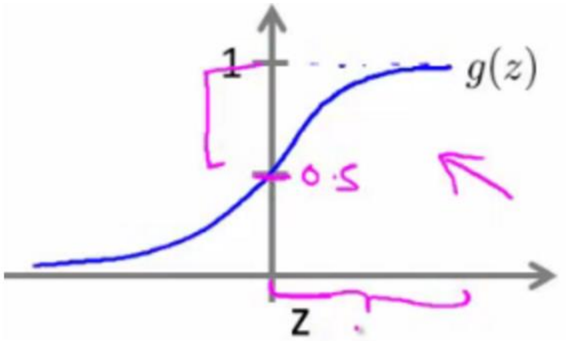


Interpretation

- We treat the hypothesis as the estimated probability of Y=1
- If X is a feature vector with $x_0 = 1$, $x_1 = tumoursize$
- $g(\theta^T x) = 0.7$ means a patient has 70% chance of a tumour being malignant or it can be written in the probabilistic notation as $g(\theta^T x) = P(y = 1|x;\theta)$
- $P(y = 1|z) + P(y = 0|z) = 1, P(y = 0|z) = 1 - P(y = 1|z)$

Decision Boundary

- When probability of y=1 being 1 is greater than 0.5 then we can predict y=1
- $g(z)$ ≥0.5, when $z \geq 0$
- Y=1 when $\theta^T x \geq 0$



Decision Boundary

- $g(\theta^T x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

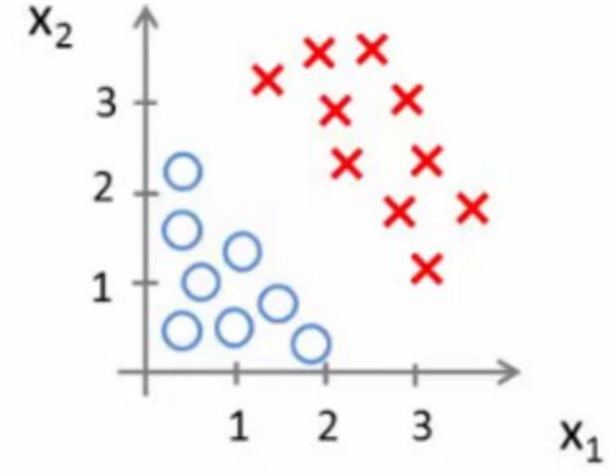
❑ Suppose $\theta_0 = -3, \theta_1 = 1, \theta_2 = 1, \theta = [-3,1,1]$

❑ We predict y=1 if $y = -3 + x_1 + x_2$

❑ We can rewrite it as $x_1 + x_2 \geq 3$

❑ Decision boundary is a property of hypothesis

Decision Boundary

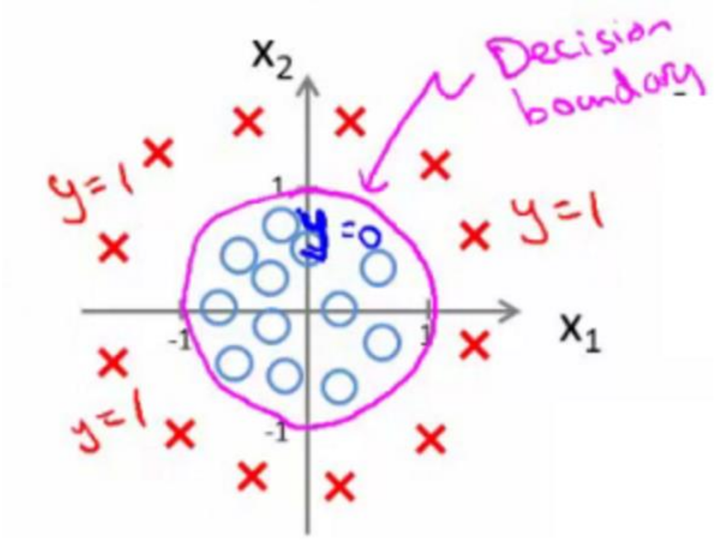


Decision Boundary

- $g(z) = g(\theta_0 + x_1\theta_1 + x_2\theta_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$
- If $\theta = [-1,0,0,1,1]$, $g(z) = -1 + x_1^2 + x_2^2$
- Y=1 if $x_1^2 + x_2^2 \geq 1$.
- This gives us a circle with a radius of 1 around 0
- By using higher order polynomial terms,

we can get even more complex decision boundaries

Decision Boundary



Cost Function

- Define the cost function to tune θ
- Suppose $cost = \frac{1}{2}(g(z) - y)^2$,

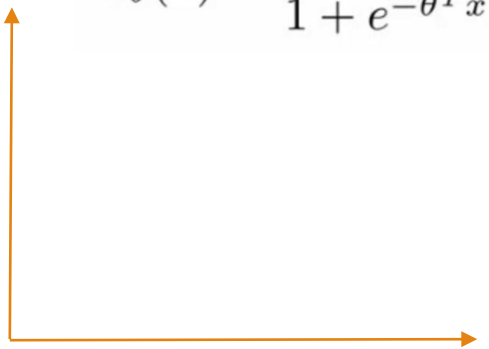
 the cost function is written as $J(\theta) = \sum_{i=1}^{m} Cost(g(z), y)$
- This is a non-convex function, having many local minimums
- Need a convex function to converge to a global minimum

Cost Function

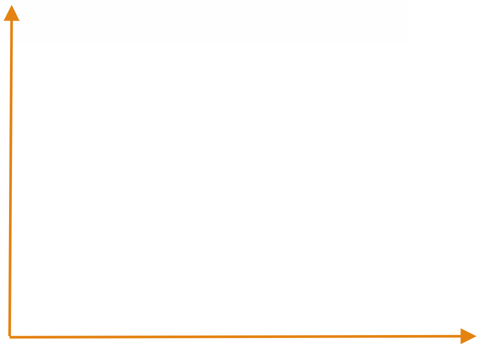
Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$

m examples $\quad x \in \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix} \quad x_0 = 1, y \in \{0, 1\}$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$



Non-convex



convex

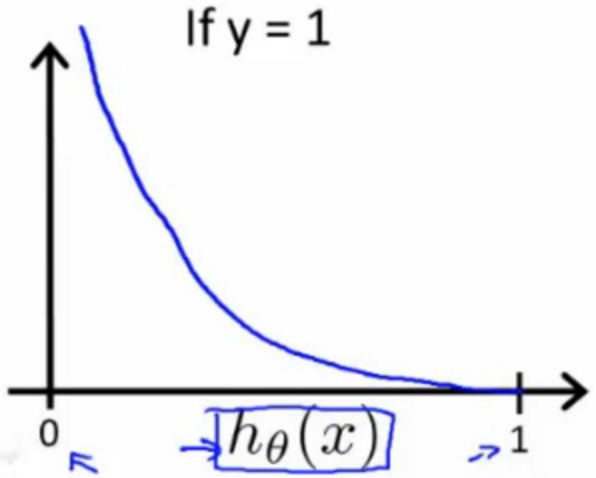
Cost Function

- Logistic regression cost function is as follows

$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost(g(z), y)$, where

$$Cost(g(z), y) = \begin{cases} -\log(g(z)), \ y = 1 \\ -\log\big(1 - g(z)\big), y = 0 \end{cases}$$

- If y=1, g(z)=1, Cost=0
- But if g(z)=0, $g(z) = \infty$
- With this, $J(\theta)$is convex and avoids local minima



Simplified Cost Function

- For binary classification, problem y is either 0 or, 1
- We can compress the cost function into one line

$$Cost(g(z), y) = -y\log\big(g(z)\big) - (1-y)\log\big(1-g(z)\big)$$

So, the cost function is now $J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(g(z), y)$

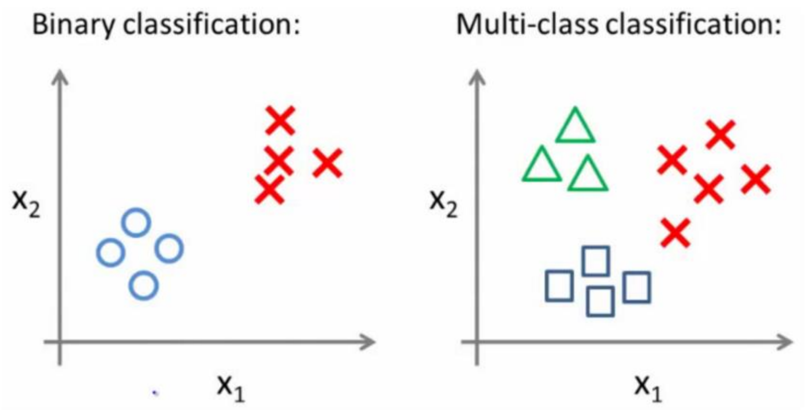
$$= \frac{-1}{m}[\sum_{i=1}^{m} y\log\big(g(z)\big) + (1-y)\log(1-g(z))]$$

- This cost function can be derived using the maximum likelihood estimation, assuming that data follows Bernoullie distribution
- Now, we can find θthat minimizes $J(\theta)$

Gradient Descent

- To minimize $J(\theta)$, we can use the gradient descent method
- Note that $\theta \in \Re^{n+1}$, where *n* is the number of input attributes
- Update rule: $\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta} = \theta_j - \alpha \sum_{i=1}^{m} (g(z) - y) x_{i,j}$
- Where α is a learning rate

Multiclass classification problem



- Multiclass classification problem : Classification with multiple classes
- One versus All classification : split the problem into three binary classification problems
- Triangles vs Squares and Crosses, Squares vs Triangles and Crosses, Crosses vs Triangles and Squares

Multiclass Classification Problem

- Train a logistic regression *g(z)* for each class *I*
- To make a prediction, choose the class *I* that maximizes the probability of *g(z)=1*

$$I = \max_{i=1,2,3} G_i(Z)$$