Logistic Regression Classifier

Mohsen Afsharchi

TOC

- **Linear Regression (Reminder)**

- Logistic Regression
  - Regression based explanation
    - Linear regression and classification
    - Logistic regression model
    - Cost function
    - Parameter optimization
    - Multi-class problem
  - Bayesian based explanation
    - Sigmoid/Logistic function

Linear Regression (1/3)

- The goal is to make quantitative (real valued) predictions on the basis of a (vector of) features or attributes

- Example: predicting house price from 4 attributes

Size (feet^2)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

Features — Target value

- We need to
  - specify the class of functions (e.g., linear)
  - select how to measure prediction loss
  - solve the resulting minimization problem

Linear Regression (2/3)

- Linear regression model

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

For convenience of notation, define $x_0 = 1$.

- How to find the parameters $\theta_0, \theta_1, \ldots, \theta_n$?
  - ➔ Given data, minimize the difference between real values and prediction values (prediction loss)
    : Gradient descent algorithm

- How to measure the prediction loss?
  - ➔ Cost function

$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Linear Regression (3/3)

- Gradient descent algorithm

Parameters: $\theta_0, \theta_1, \ldots, \theta_n$

Cost function:

$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Repeat $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \ldots, \theta_n)$$

$\}$ (simultaneously update for every $j = 0, \ldots, n$)

TOC

- Linear Regression (Reminder)

- Logistic Regression
  - **Regression based explanation**
    - Linear regression and classification
    - Logistic regression model
    - Cost function
    - Parameter optimization
    - Multi-class problem
  - Bayesian based explanation
    - Sigmoid/Logistic function

Linear regression and classification

- Classification

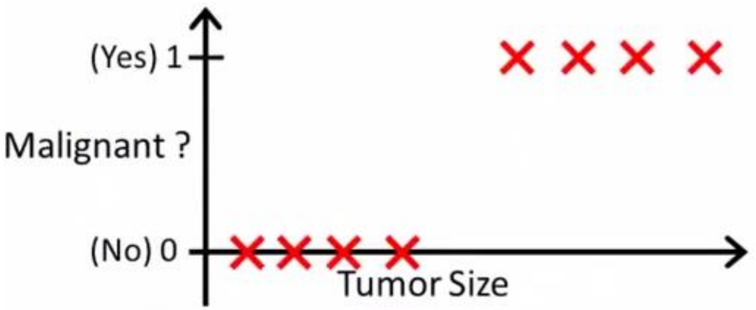
Email: Spam / Not Spam?
Online Transactions: Fraudulent (Yes / No)?
Tumor: Malignant / Benign ?

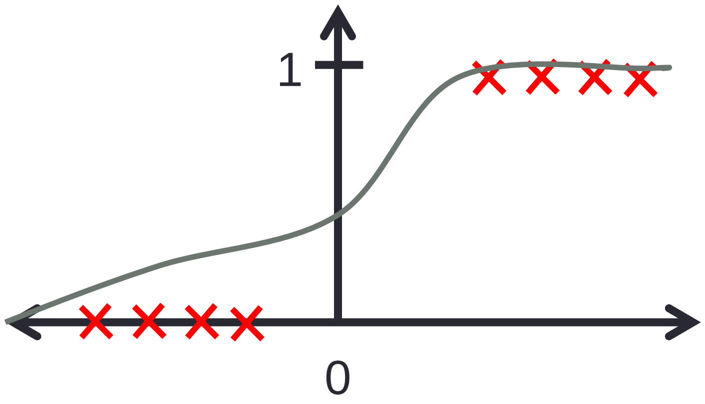
$y \in \{0, 1\}$

0: "Negative Class" (e.g., benign tumor)
1: "Positive Class" (e.g., malignant tumor)

Logistic regression model (1/3)

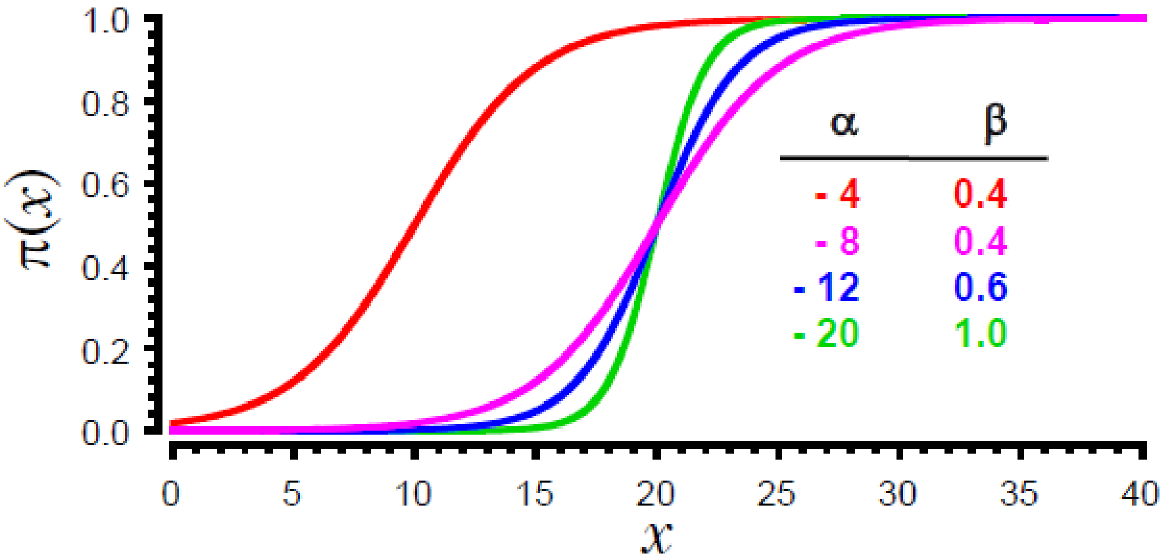
- Linear Regression Model ➔ Logistic Regression Model





Regression Parameters

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$



When $x = -\alpha / \beta$, $\alpha + \beta x = 0$ and hence $\pi(x) = 1/(1+1) = 0.5$

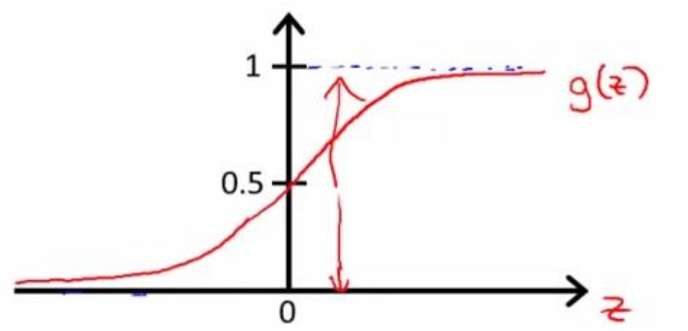
The slope of $\pi(x)$ when $\pi(x)$=.5 is $\beta/4$.

Thus β controls how fast $\pi(x)$ rises from 0 to 1.

Logistic regression model (2/3)

- Want $0 \leq h_\theta(x) \leq 1$

- How?
  - ➔**Logistic** function / Sigmoid function



- Linear Regression: $h_\theta(x) = \theta^T x$

- Logistic Regression: $h_\theta(x) = g(\theta^T x)$; $g(z) = \frac{1}{1+ e^{-z}}$

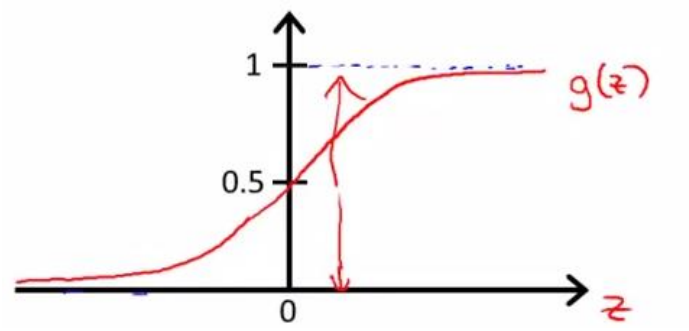
$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

Logistic regression model (3/3)

- Interpretation of Hypothesis output
  - $h_\theta(x)$ = estimated probability that y=1 on input x

    $=P(\, y = 1 \mid x; \theta)$

- Threshold classifier output $h_\theta(x)$ at 0.5:
  - If $h_\theta(x)$ >= 0.5, predict y=1
  - If $h_\theta(x)$ < 0.5, predict y=0



Cost function (1/4)

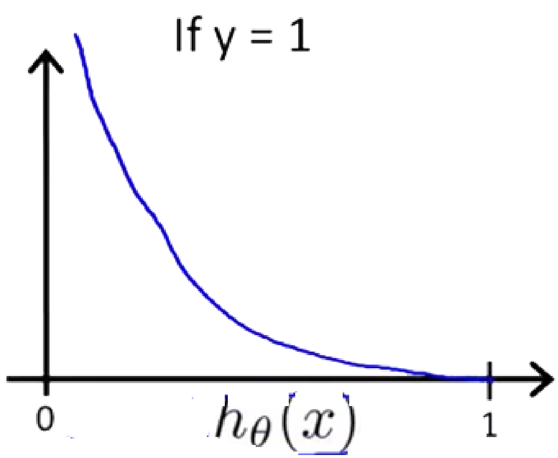
- How to find the best parameters?
  - ➔ Similar to linear regression: gradient descent Algorithm
  - ➔ But, different cost function

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))
\end{aligned}
$$

Cost function (2/4)

Logistic regression cost function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} \boxed{-\log(h_\theta(x)) \quad \text{if } y = 1} \\ -\log(1 - h_\theta(x)) \quad \text{if } y = 0 \end{cases}$$



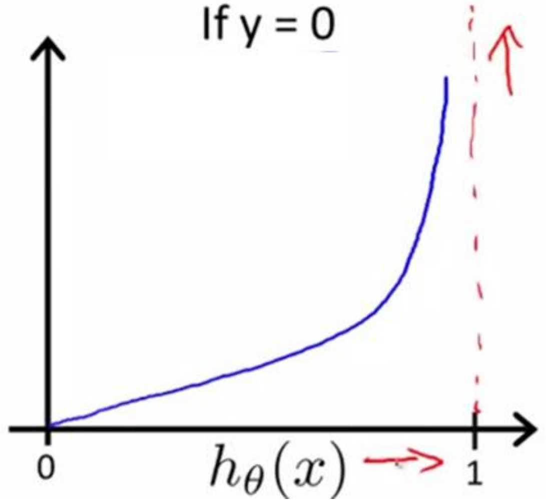
Cost $= 0$ if $y = 1, h_\theta(x) = 1$
But as $h_\theta(x) \to 0$
$Cost \to \infty$

Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1|x;\theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

Cost function (3/4)

Logistic regression cost function

$$\mathrm{Cost}(h_\theta(x^{(i)}, y^{(i)}) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ \boxed{-\log(1 - h_\theta(x)) \quad \text{if } y = 0} \end{cases}$$



Cost function (4/4)

Logistic regression cost function

$$\mathrm{Cost}(h_\theta(x^{(i)}, y^{(i)}) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$\begin{aligned} J(\theta) &= \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) \\ &= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))] \end{aligned}$$

Parameter optimization

$$J(\theta) = \frac{1}{m}\sum_{i=1} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$
$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))]$$

To fit parameters θ :

$$\min_\theta J(\theta)$$

To make a prediction given new x:

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

Gradient Descent

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))]$$

Want $\min_\theta J(\theta)$:

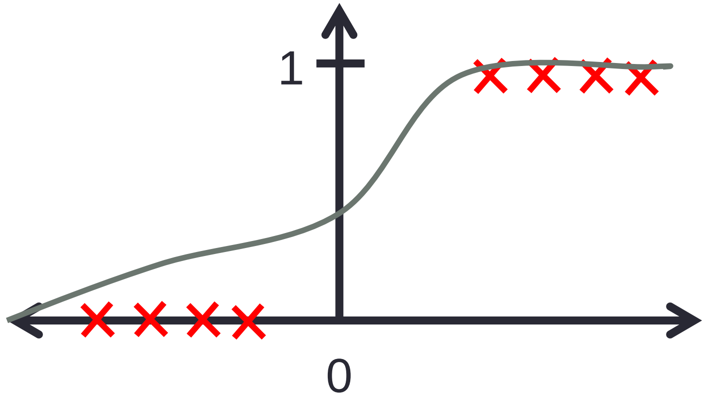
Repeat $\{$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

$\}$

Multi-class problem

- How to adopt logistic regression classification for multi-class problem?



[Binary class]

[multi-class]

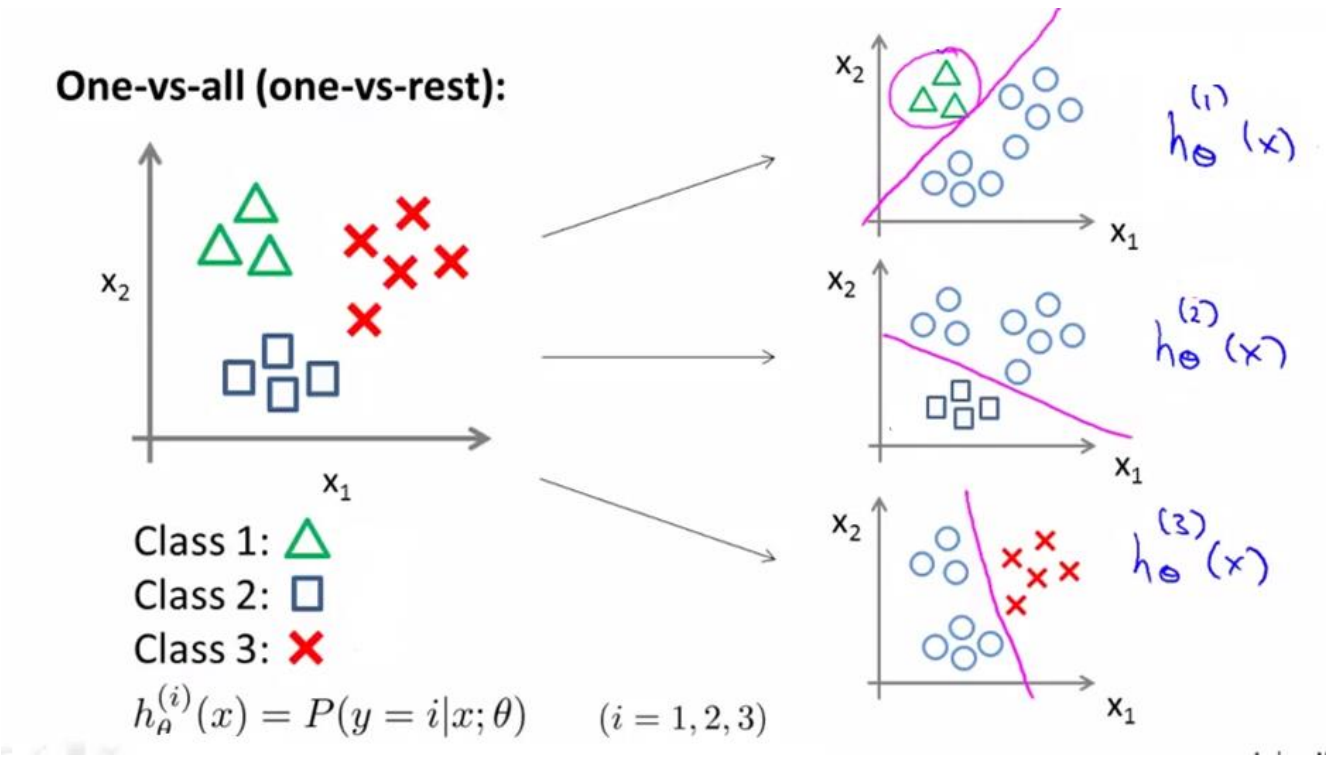
Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x)$ >= 0.5, predict y=1

If $h_\theta(x)$ < 0.5, predict y=0

?

Multi-class problem



Multi-class problem

One-vs-all

Train a logistic regression classifier $h_\theta^{(i)}(x)$ for each class i to predict the probability that $y = i$.

On a new input x, to make a prediction, pick the class i that maximizes

$$\max_i h_\theta^{(i)}(x)$$