Logistic Regression

Eun Yi Kim

Artificial Inteligence
& Computer Vision
Laboratory

Today

- Classification
- Logistic Regression
  - ✓Hypothesis representation
  - ✓Cost function
  - ✓Gradient decent
  - ✓Advanced optimization algorithm
- Multi-class classification

Classification

Classification

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

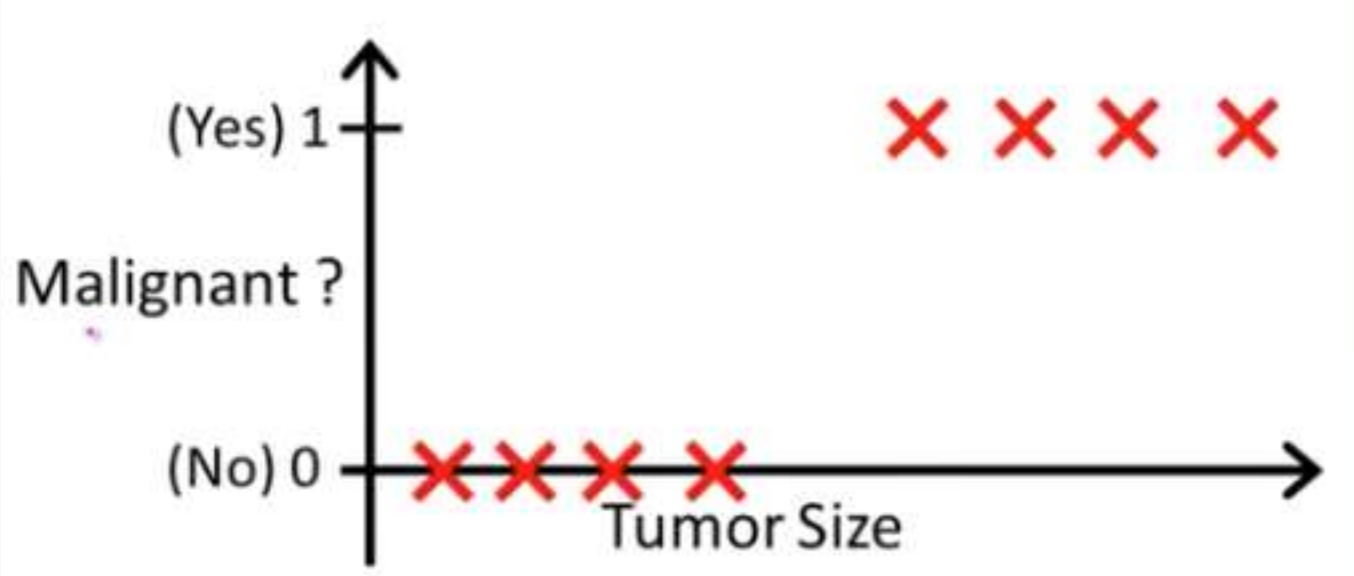
$$y \lfloor \{0,1\}$$

0: “Negative class”

1: “Positive Class”

Classification with Linear Regression

Artificial Inteligence & Computer Vision Laboratory



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

$$\text{If } h_\theta(x) \geq 0.5, \text{ predict "} y = 1 \text{"}$$

$$\text{If } h_\theta(x) < 0.5, \text{ predict "} y = 0 \text{"}$$

Classification with Linear Regression

Artificial Inteligence & Computer Vision Laboratory

- Classification: y = 0 or 1

$$h_\theta(x) \text{ can be } > 1 \text{ or } < 0$$

- Logistic Regression: $0 \le h_\theta(x) \le 1$

Logistic Regression

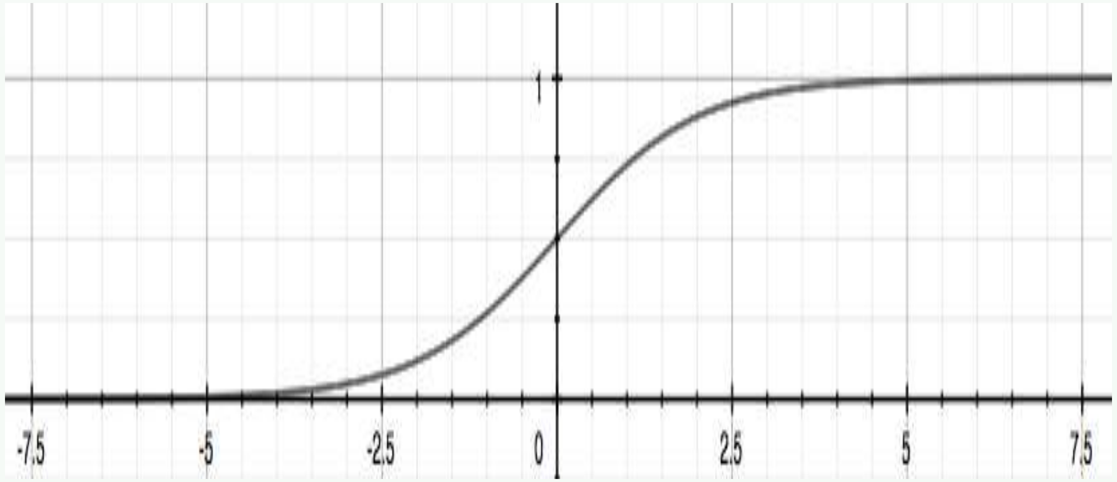
Hypothesis Representation

Logistic Regression Model

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

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

- Sigmoid function
- Logistic function



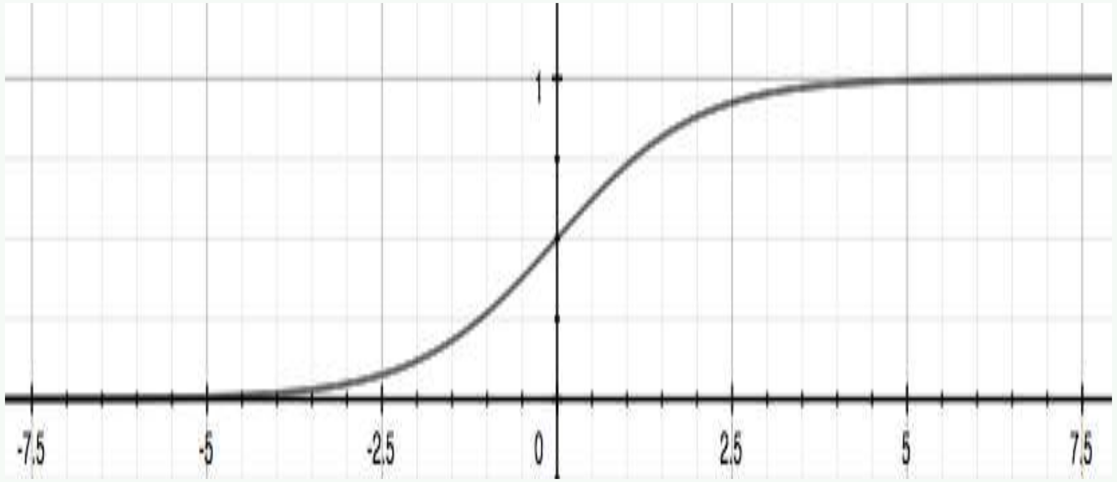
Logistic Regression Model

Artificial Inteligence & Computer Vision Laboratory

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

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

- Sigmoid function
- Logistic function



Interpretation of Hypothesis Output

Artificial Inteligence & Computer Vision Laboratory

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

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

- Example: if $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$$h_\theta(x) = 0.7$$

Tell patient that 70% chance of tumor being malignant

"Probability that y=1, given x, parameterized by θ

$$h_\theta(x) = P(y = 1|x; \theta) = 1 - P(y = 0|x; \theta)$$

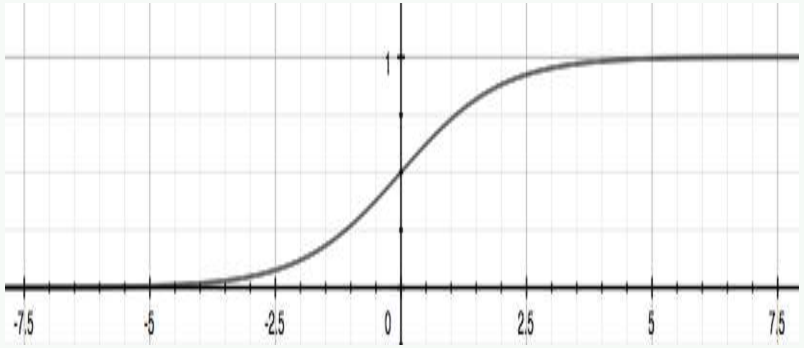
$$P(y = 0|x; \theta) + P(y = 1|x; \theta) = 1$$

Decision Boundary

- Logistic regression

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

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

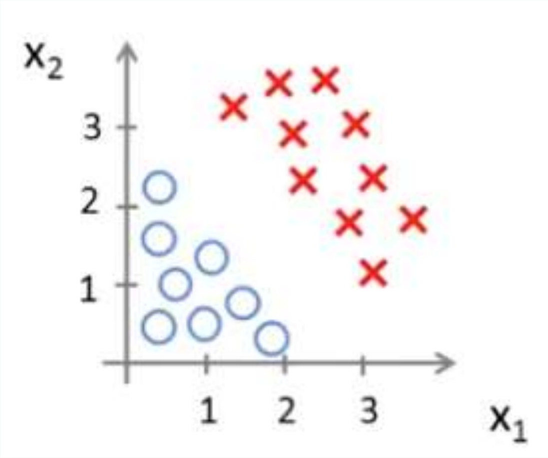


Suppose predict "y=1" if $h_\theta(x) \quad 0.5$

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

Decision Boundary

- Linear decision boundary



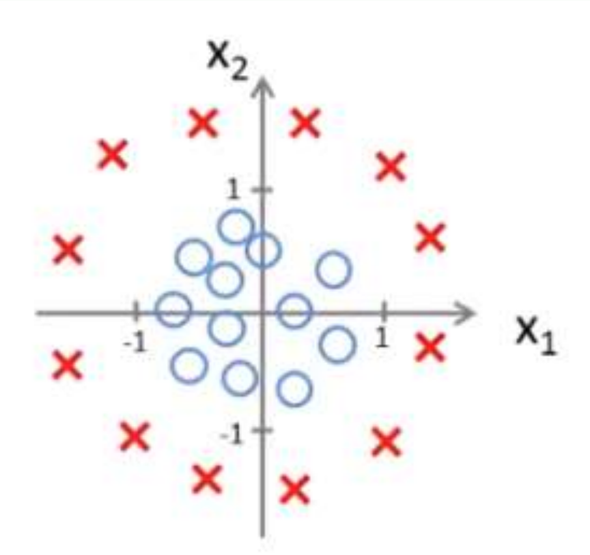
$$h_\theta(x) = g\left(\theta^T x\right) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

predict "y=1" if $-3 + x_1 + x_2 \quad 0$

Decision Boundary

Artificial Inteligence
& Computer Vision
Laboratory

- Non-linear decision boundary



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

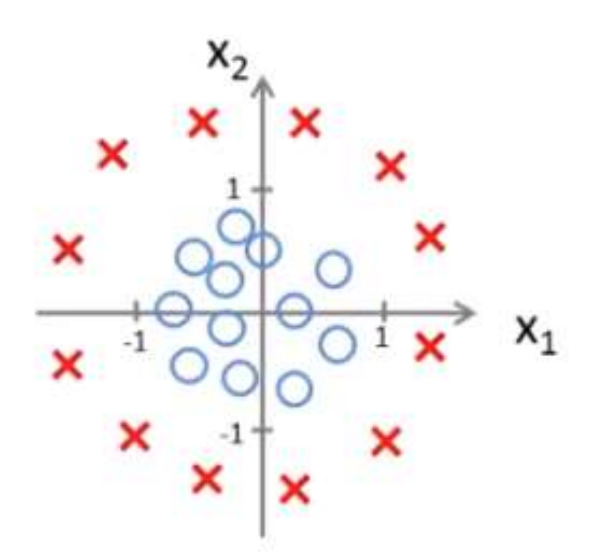
predict “y=1” if $-1 + x_1^2 + x_2^2 \quad 0$

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

Decision Boundary

Artificial Inteligence
& Computer Vision
Laboratory

- Non-linear decision boundary



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

predict “y=1” if $-1 + x_1^2 + x_2^2 \quad 0$

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

Logistic Regression

Cost function

Cost Function

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

m examples

$$x^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} \in \mathbb{R}^{n+1}$$

$$x_0 = 1,\ y \in \{0,1\}$$

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

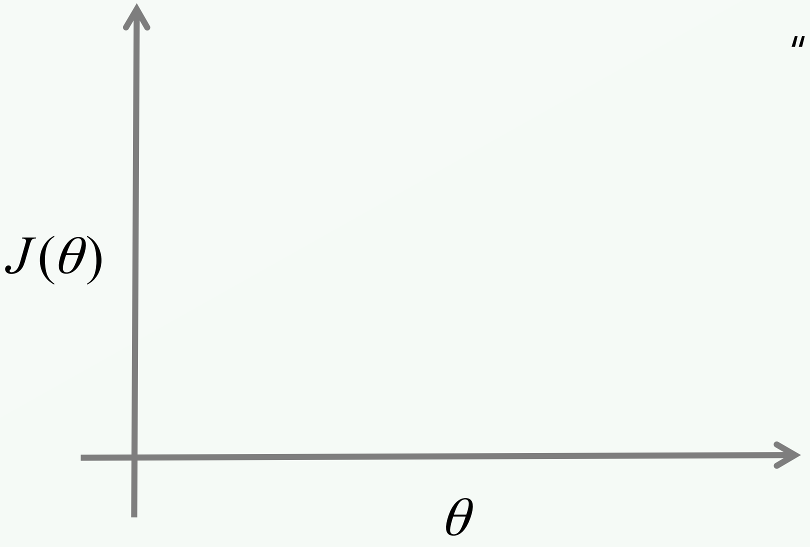
How to choose parameter θ ?

Cost Function

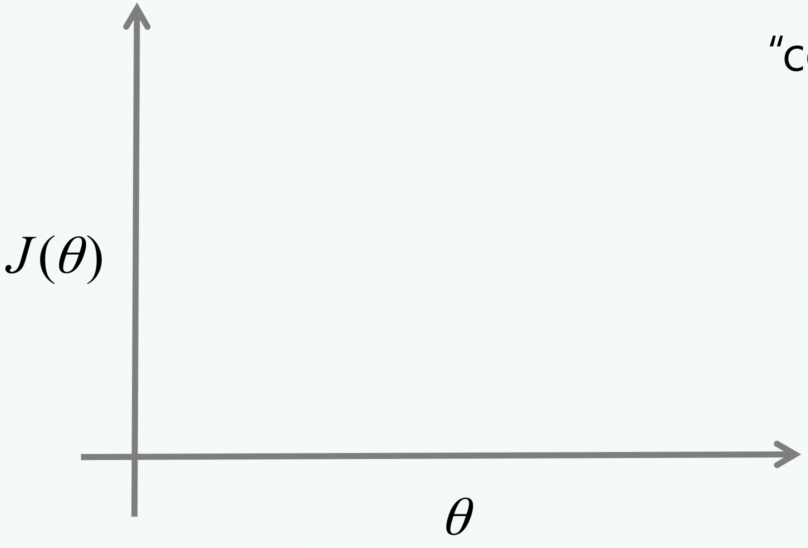
- Linear regression: $J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

- Logistic regression: $J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost\left(h_\theta(x^{(i)}), y^{(i)}\right)$

$$Cost\left(h_\theta(x^{(i)}), y^{(i)}\right) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$$



"non-convex"

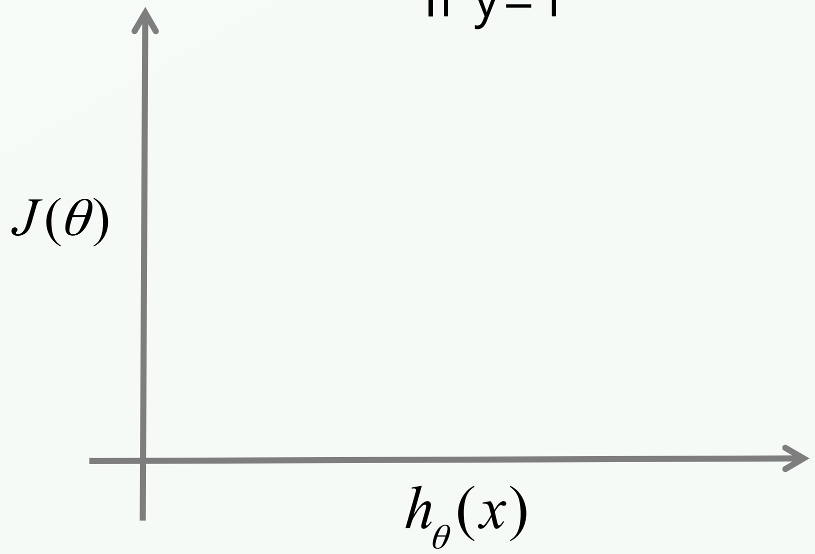


"convex"

Logistic Regression Cost Function

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

If y=1



$\mathrm{Cost}(h_\theta(x), y) = 0$ if $h_\theta(x) = y$

But if $y = 1$ and $h_\theta(x) \to 0$

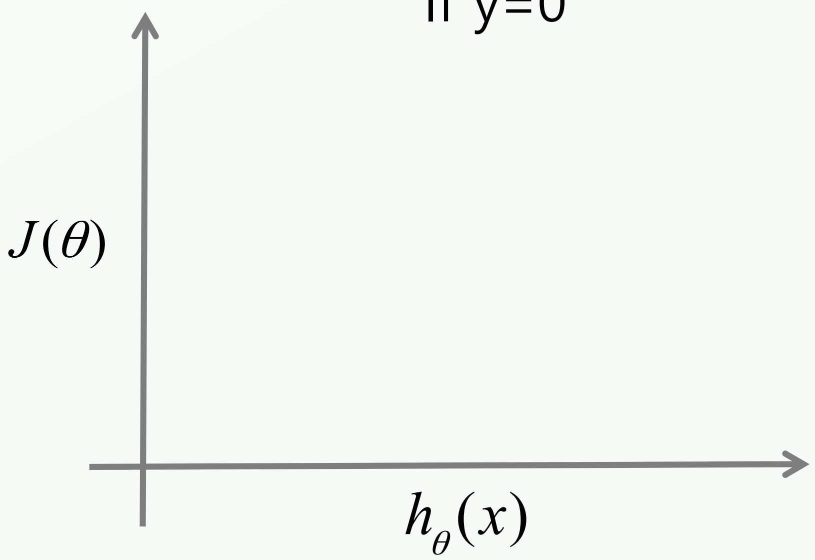
$\mathrm{Cost}(h_\theta(x), y) \to \infty$

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

Logistic Regression Cost Function

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

If y=0



$\mathrm{Cost}(h_\theta(x), y) = 0$ if $h_\theta(x) = y$

but if $y = 0$ and $h_\theta(x) \rightarrow 1$

$\mathrm{Cost}(h_\theta(x), y) \rightarrow \infty$

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

Logistic Regression

Simplified cost function and gradient descent

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} Cost\left(h_\theta(x^{(i)}), y^{(i)}\right)$$

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

Note: y = 0 or 1 always

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost\left(h_\theta(x^{(i)}), y^{(i)}\right)$$

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

To fit parameter θ :

$$\min_{\theta} J(\theta)$$

To make a prediction given new x :

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

Gradient Descent

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

Want $\min_{\theta} J(\theta)$

Repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial\theta_j} J(\theta)$$

}

Simultaneously update all parameters

Gradient Descent

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

Want $\min_{\theta} J(\theta)$

$$\text{Repeat } \{$$

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

$$\}$$

Simultaneously update all parameters

$$\theta := \theta - \frac{\alpha}{m}X^T(g(X\theta) - \vec{y})$$

Algorithm looks identical to linear regression !

Optimization Algorithm

Cost function $J(\theta)$. Want $\min_{\theta} J(\theta)$

Given θ , we have code that can compute

- $J(\theta)$
- $\frac{\partial}{\partial \theta_j} J(\theta)$

Gradient descent:

$$
\begin{aligned}
&\text{Repeat } \{ \\
&\quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\
&\}
\end{aligned}
$$

Optimization Algorithm

Artificial Inteligence & Computer Vision Laboratory

Given θ , we have code that can compute

- $J(\theta)$
- $\frac{\partial}{\partial\theta_j} J(\theta)$

Optimization algorithms:

- Gradient descent
- Conjugate gradient
- BFGS
- L-BFGS

Advantages:

- No need to manually pick α
- Often faster than gradient descent

Disadvantages:

- More complex

Logistic Regression

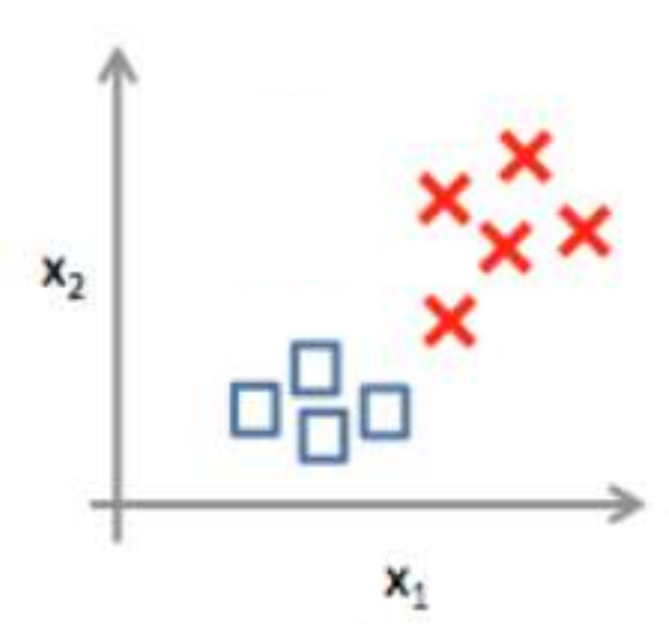
Multi-class classification: One-vs-all

Multiclass Classification

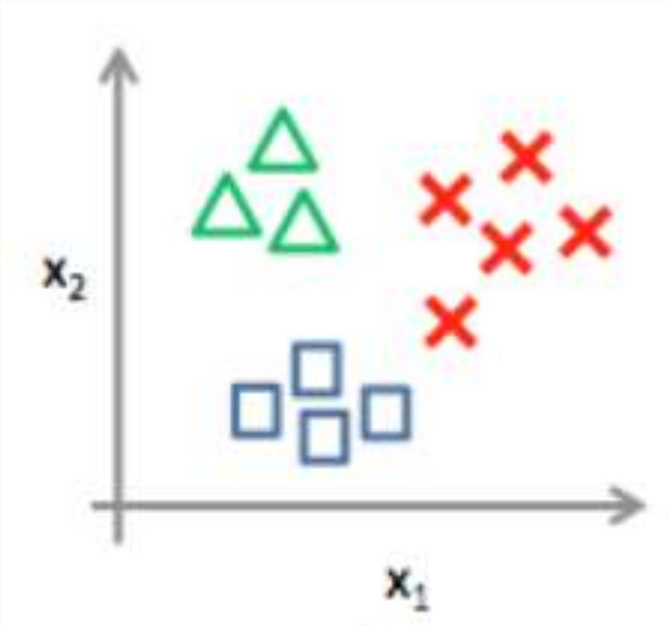
- Email foldering/tagging: Work, Friends, Family, Hobby
- Medical diagrams: Not ill, Cold, Flu, Covid19
- Weather: Sunny, Cloudy, Rain, Snow

Multiclass Classification

Binary classification:



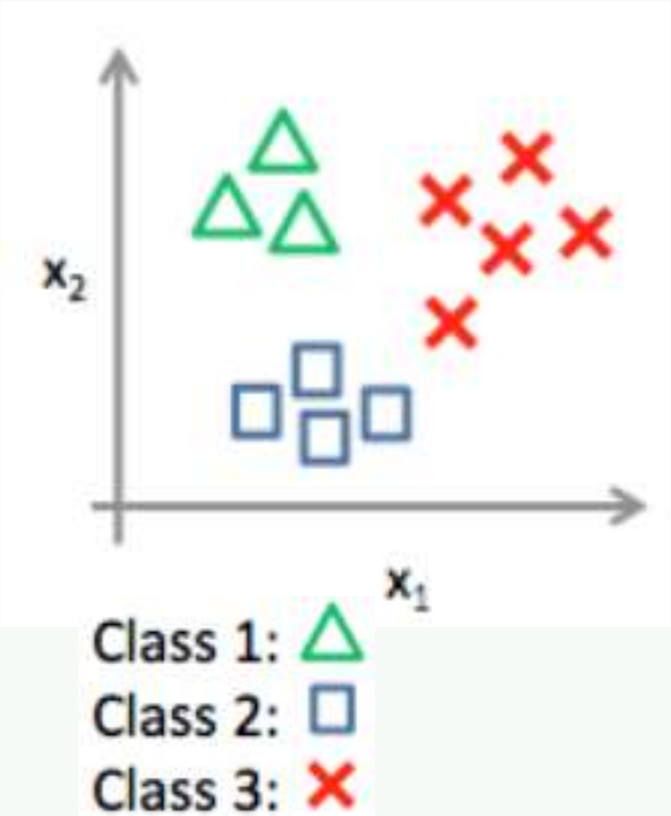
Multi-class classification:

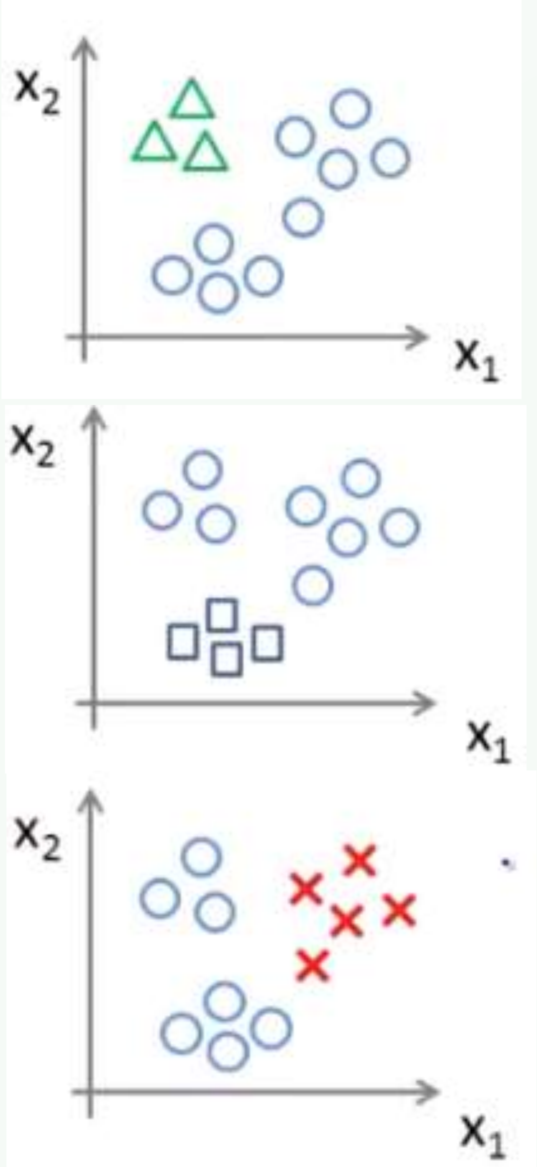


Multiclass Classification

Artificial Inteligence & Computer Vision Laboratory

One-vs-all (one-vs-rest)





$$h_\theta^{(i)}(x) = P(y = i \mid x; \theta) \qquad (i = 1, 2, 3)$$

Multiclass Classification

One-vs-all (one-vs-rest)

- 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 maximize

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

Today

Artificial Inteligence
& Computer Vision
Laboratory

- Classification
- Logistic Regression
  - ✓Hypothesis representation
  - ✓Cost function
  - ✓Gradient decent
  - ✓Advanced optimization algorithm
- Multi-class classification