Linear Regression & Logistic Regression

Trại Hè Toán và Khoa học MaSSP
Môn Tin học
Hà Nội, Tháng 6/2017

Outline

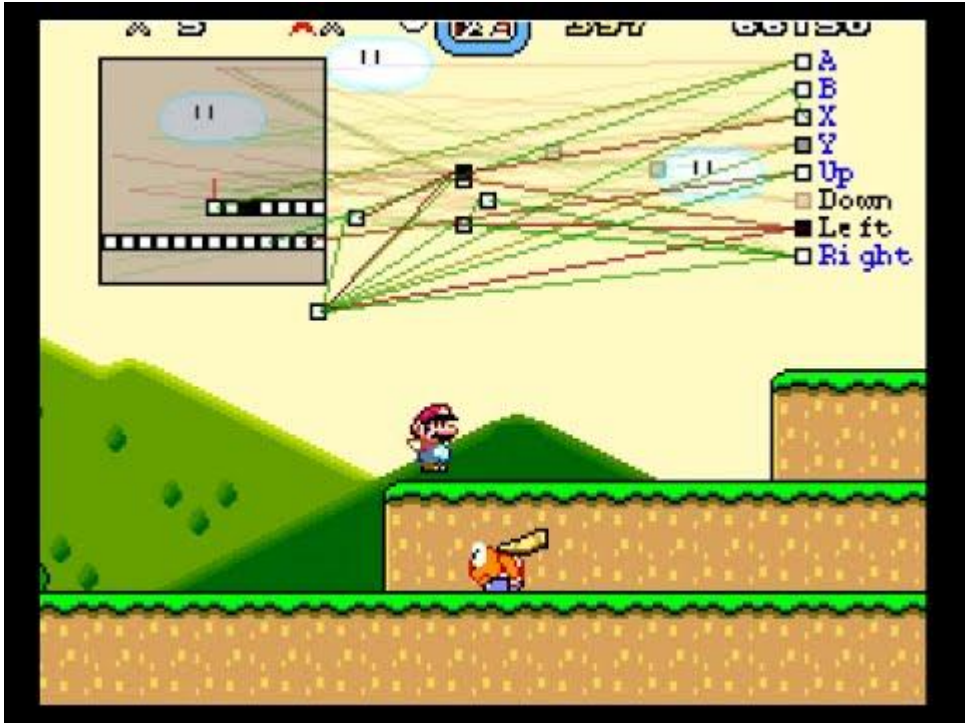
- Giới thiệu bài toán Machine Learning
- Kiến thức cơ bản
  - Linear Regression
  - Logistic Regression
- Những vấn đề nâng cao
- Câu hỏi

Giới thiệu bài toán Machine Learning

Ví dụ:

- Một engine chơi cờ bằng cách phân tích dữ liệu những ván cờ đã chơi, dữ liệu càng lớn chơi càng tốt.
- Một chương trình nhận diện ảnh diễn viên dựa vào thư viện ảnh của các diễn viên, thư viện càng lớn nhận diện càng chính xác.

Giới thiệu bài toán Machine Learning



Giới thiệu bài toán Machine Learning

Khái niệm: Một chương trình máy tính được học từ kinh nghiệm ***E***, tương ứng với một loại công việc ***T*** và thước đo hiệu suất ***P***. Hiệu suất của chương trình này tăng lên nhờ kinh nghiệm.

Machine Learning là dạy cho máy tính có khả năng học mà không cần lập trình một cách tường minh.

“Đừng cho máy tính con cá, hãy cho nó cái cần câu.”

-Hoang Vu-

Giới thiệu bài toán Machine Learning

Phân loại:

- Supervised Learning (Học có giám sát)
  - Regression (Hồi quy)
  - Classification (Phân loại)
- Unsupervised Learning (Học không giám sát)

Giới thiệu bài toán Machine Learning

Ví dụ:

	Regression	Classification	Unsupervised
Cho thông tin hồ sơ ứng viên MaSSP năm ngoái, dự đoán một thí sinh năm nay đỗ hay trượt.		X	
Cho một danh sách các bài báo, gom các bài báo có nội dung giống nhau làm một chủ đề			X
Cho dữ liệu chỉ số cầu thủ của PES, dự đoán giá của một cầu thủ trên TTCN	X		
Cho kết quả xổ số một năm trở lại đây, dự đoán số đề ngày mai		X	

Kiến thức cơ bản

- Linear Regression (Hồi quy tuyến tính)
- Logistic Regression (Hồi quy lôgit)
- Hypothesis Function (Hàm giả thuyết), Cost Function (Hàm giá)
- Gradient Descent (Xuống dốc)

Kiến thức cơ bản - Linear Regression

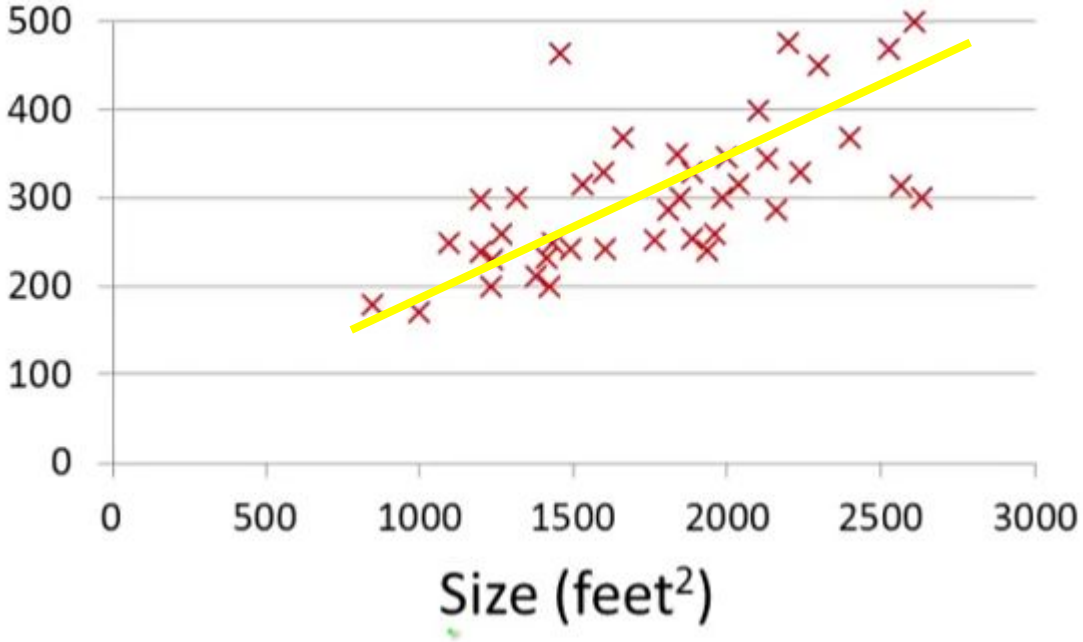
Bài toán:

Cho một tập dữ liệu gồm m ngôi nhà, ngôi nhà thứ i có kích thước $x^{(i)}$ (feet^2) và có giá là $y^{(i)}$ (x \$1000). Hãy dự đoán xem một ngôi nhà nằm ngoài dữ liệu đã cho, có kích thước x, thì có giá là bao nhiêu?

Kiến thức cơ bản - Linear Regression

Kích thước nhà (x)	Giá nhà (y)
2104	460
1416	232
1534	315
852	178
...	...

Kiến thức cơ bản - Linear Regression



→ **Phương pháp: Tìm đường thẳng "khớp" nhất**

Kiến thức cơ bản - Linear Regression

Hypothesis Function: $h_\theta(x) = \theta_0 + \theta_1 x$

Mục tiêu: Tìm θ_0, θ_1 để $h_\theta(x)$ khớp với training set nhất.

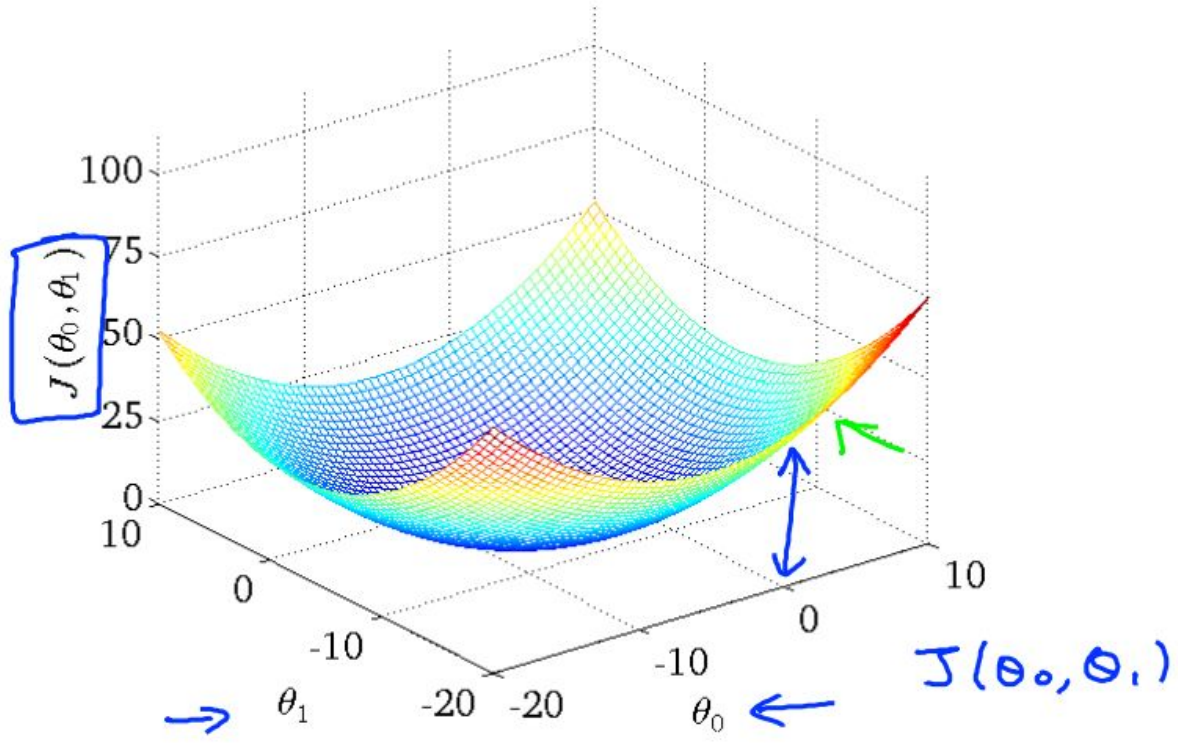
Kiến thức cơ bản - Linear Regression

Thế nào là khớp nhất?

Cost Function: $$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{2} (h_\theta(x^{(i)}) - y^{(i)})^2 \right)$$

Mục tiêu: Cực tiểu hóa $J(\theta_0, \theta_1)$

Kiến thức cơ bản - Linear Regression



Kiến thức cơ bản - Linear Regression

Gradient descent: Cực tiểu hóa hàm $J(\theta_0, \theta_1)$.

- Bắt đầu với giá trị θ_0, θ_1 bất kì.
- Liên tục thay đổi θ_0, θ_1 để $J(\theta_0, \theta_1)$ giảm từng bước nhỏ cho đến khi $J(\theta_0, \theta_1)$ đạt cực tiểu địa phương.

Kiến thức cơ bản - Linear Regression

Thuật toán Gradient Descent:

Repeat

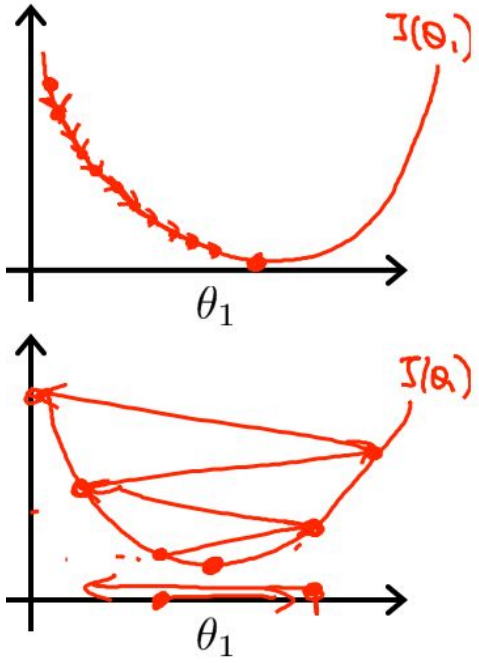
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \text{ với } j = 0 \text{ và } j = 1$$

- α - Learning rates (tốc độ học)
- $\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ Đạo hàm một phía của Cost Function

LƯU Ý: CẬP NHẬT θ_0 VÀ θ_1 **ĐỒNG THỜI**, KHÔNG PHẢI LẦN LƯỢT!

Kiến thức cơ bản - Linear Regression

α thể hiện độ lớn của một bước nhảy. Nếu α quá nhỏ, Gradient Descent sẽ mất rất lâu để đạt cực tiểu. Nếu α quá lớn, Gradient Descent sẽ có thể không bao giờ đạt cực tiểu do nhảy qua điểm cực tiểu quá xa.



Kiến thức cơ bản - Linear Regression

Áp dụng Gradient Descent vào Linear Regression:

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} ((h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}) \mid x_0^{(i)} = 1 \ \forall i$$

Kiến thức cơ bản - Linear Regression

Linear Regression với nhiều biến - nhiều thuộc tính

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
...	...	...	...	...

Kiến thức cơ bản - Linear Regression

- ***m***: Số ví dụ trong training set
- ***n***: Số thuộc tính
- $x_j^{(i)}$: Giá trị của thuộc tính thứ ***j*** trong ví dụ thứ ***i***
- $y^{(i)}$: Output của ví dụ thứ ***i***

Kiến thức cơ bản - Linear Regression

Hypothesis Function: $h_\theta(x_0, x_1, \dots, x_n) = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_n x_n$

Cost Function: $J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{2} (h_\theta(x_0^{(i)}, x_1^{(i)}, \dots, x_n^{(i)}) - y^{(i)})^2 \right)$

Đạo hàm một phía của Cost Function:

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{m} \sum_{i=1}^{m} ((h_\theta(x_0^{(i)}, x_1^{(i)}, \dots, x_n^{(i)}) - y^{(i)}) \cdot x_j^{(i)})$$

$x_0 = 1$

Kiến thức cơ bản - Linear Regression

Gradient Descent:

Repeat

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \ldots, \theta_n) \quad với \; j = 0, 1, \ldots, n$$

LƯU Ý: CẬP NHẬT $\boldsymbol{\theta_0}$, $\boldsymbol{\theta_1}$, ... , **ĐỒNG THỜI**, KHÔNG PHẢI LẦN LƯỢT!

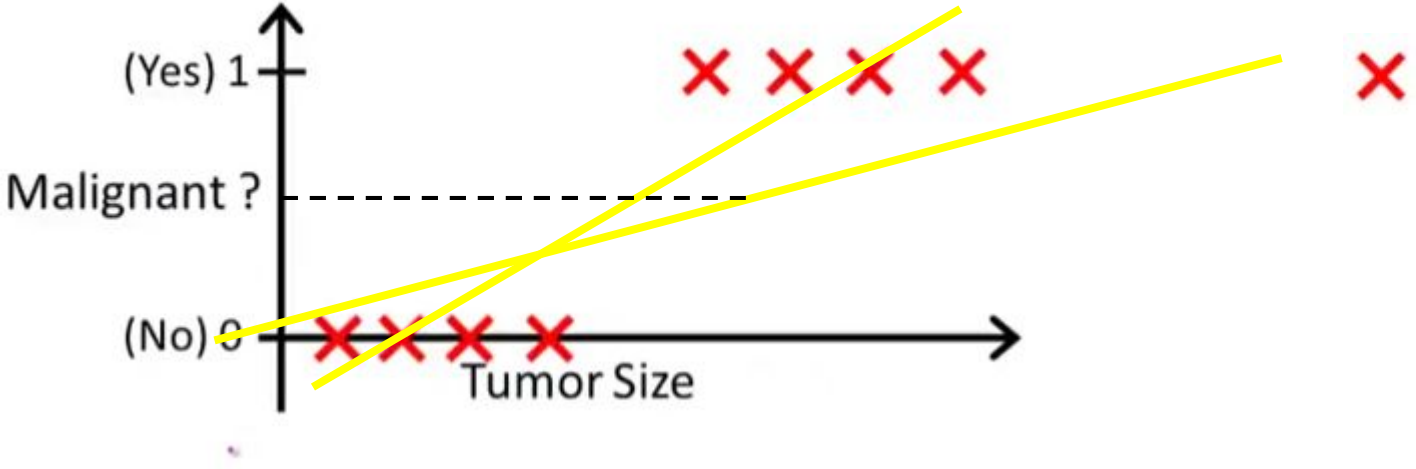
Kiến thức cơ bản - Logistic Regression

Bài toán:

Cho một training set gồm ***m*** khối u ung thư, khối u thứ ***i*** có kích thước $x^{(i)}$ (cm) và có thể là lành tính ($y^{(i)} = 0$) hoặc ác tính ($y^{(i)} = 1$). Hãy dự đoán xem một khối u nằm ngoài dữ liệu đã cho, với kích thước ***x***, là lành tính hay ác tính.

Kiến thức cơ bản - Logistic Regression

Thử áp dụng mô hình Linear Regression:



Kiến thức cơ bản - Logistic Regression

Hypothesis Function của Linear Regression không tốt để giải bài toán Classification

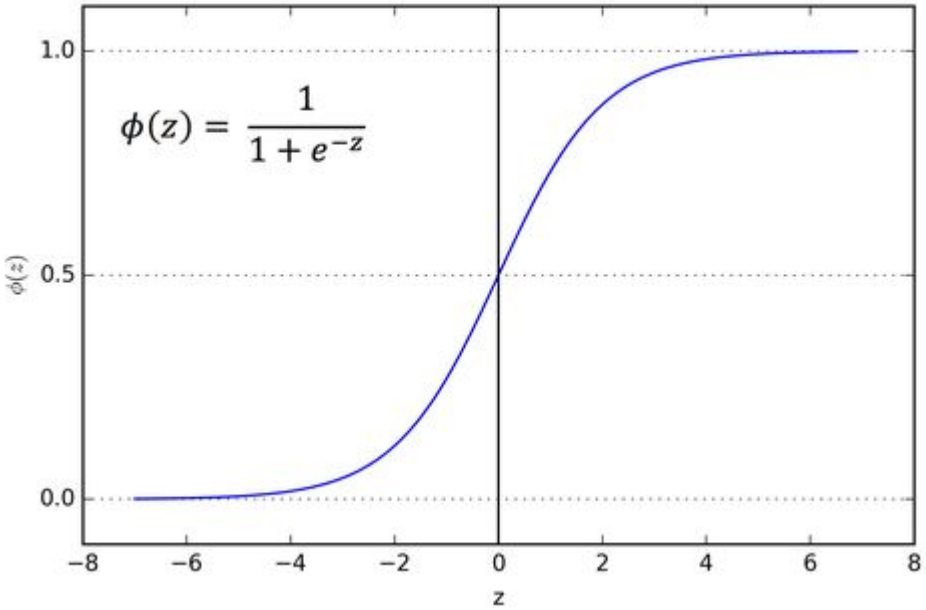
→ Thay đổi Hypothesis Function:

$$h_\theta(x_0, x_1, \ldots, x_n) = \frac{1}{1 + e^{-z}}$$

$$z = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n$$

Logistic Function (Sigmoid Function)

Kiến thức cơ bản - Logistic Regression



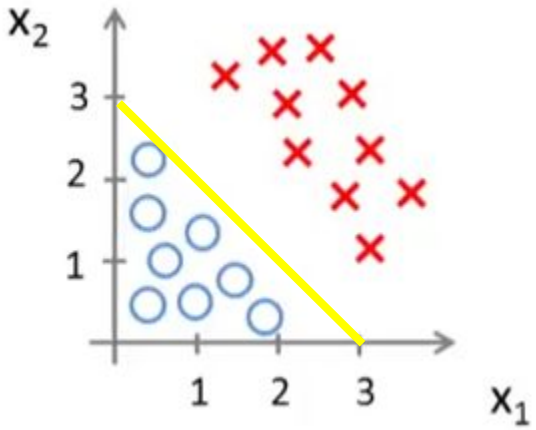
Kiến thức cơ bản - Logistic Regression

$h_\theta(x_0, x_1, \dots, x_n)$ cho biết xác suất để $y = 1$

Ví dụ: $h_\theta(x_0, x_1, \dots, x_n) = 0.7$ → Có 70% khối u với các thuộc tính x là ác tính.

Kiến thức cơ bản - Logistic Regression

Decision Boundary (Đường phân chia quyết định):



$$z = -3 + x_1 + x_2$$

Kiến thức cơ bản - Logistic Regression

Cost Function: Liệu có thể sử dụng Cost Function của Linear Regression?

$$J(\theta_0,\ \theta_1,\ \ldots,\ \theta_n) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}(h_\theta(x_0^{(i)},\ x_1^{(i)},\ \ldots,\ x_n^{(i)}) - y^{(i)})^2\right)$$

Nhược điểm: Hàm nhiều cực tiểu địa phương chứ không duy nhất, làm cho Gradient Descent khó tìm được cực tiểu toàn cục.

Kiến thức cơ bản - Logistic Regression

Sử dụng Cost Function khác:

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

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

→ Làm Gradient Descent giống như Linear Regression

Kiến thức cơ bản - Logistic Regression

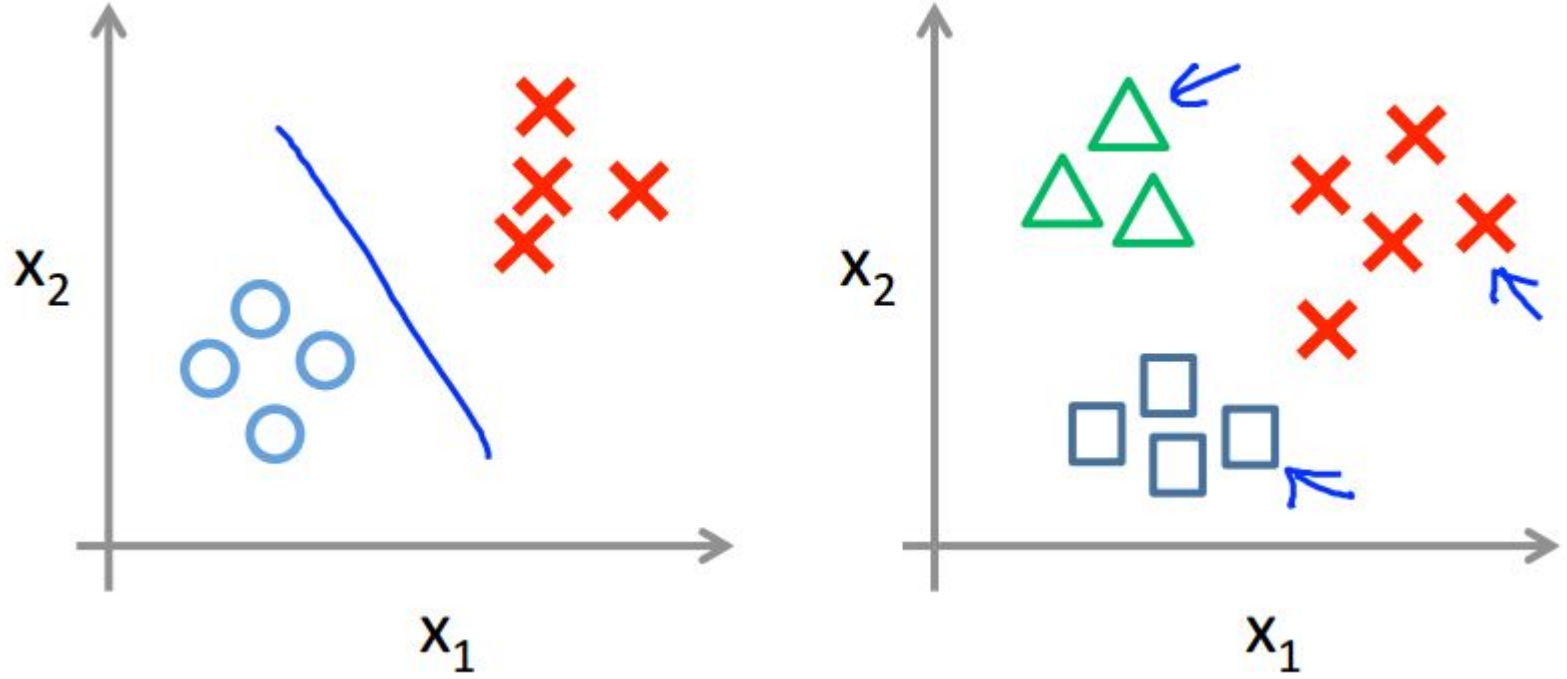
Bài toán: Multiclass Classification

Bài toán xuất hiện khi output ta cần phân loại nhiều hơn hai class.

Ví dụ:

- Dự báo thời tiết nắng, mưa, sương mù, nhiều mây.
- Đoán số đề ngày mai.

Kiến thức cơ bản - Logistic Regression



Kiến thức cơ bản - Logistic Regression

Phương pháp one-vs-all (một mình chấp tất):

- Với mỗi class, ta chuyển output của dữ liệu thành dạng hai class - hoặc là class đang xét, hoặc không phải class đó.
- Chạy Gradient Descent với mỗi class để tính xác suất đối tượng cần phân loại thuộc class đó.
- Chọn class có xác suất xảy ra cao nhất.

Kiến thức cơ bản - Logistic Regression

Phương pháp Softmax Regression (hồi quy Softmax):

Những vấn đề nâng cao

Vectorization (Vector hóa)

Những vấn đề nâng cao

Normal Equation (Phương trình chuẩn)

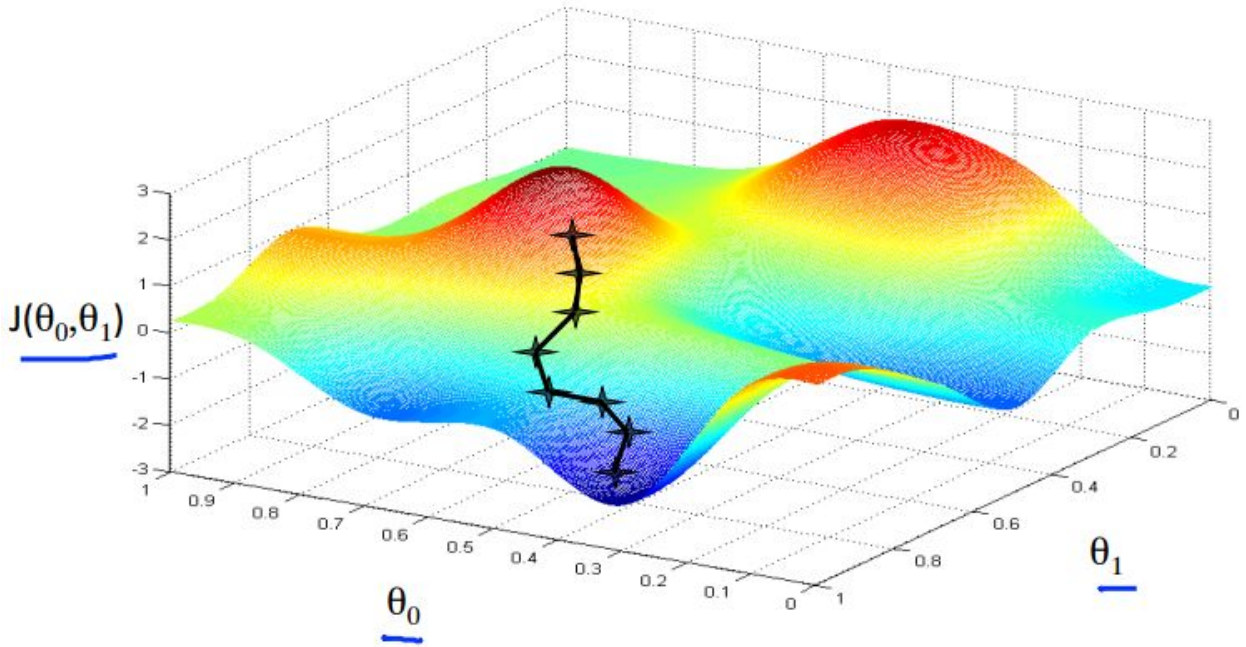
Những vấn đề nâng cao

Feature Scaling (Chuẩn hóa dữ liệu)

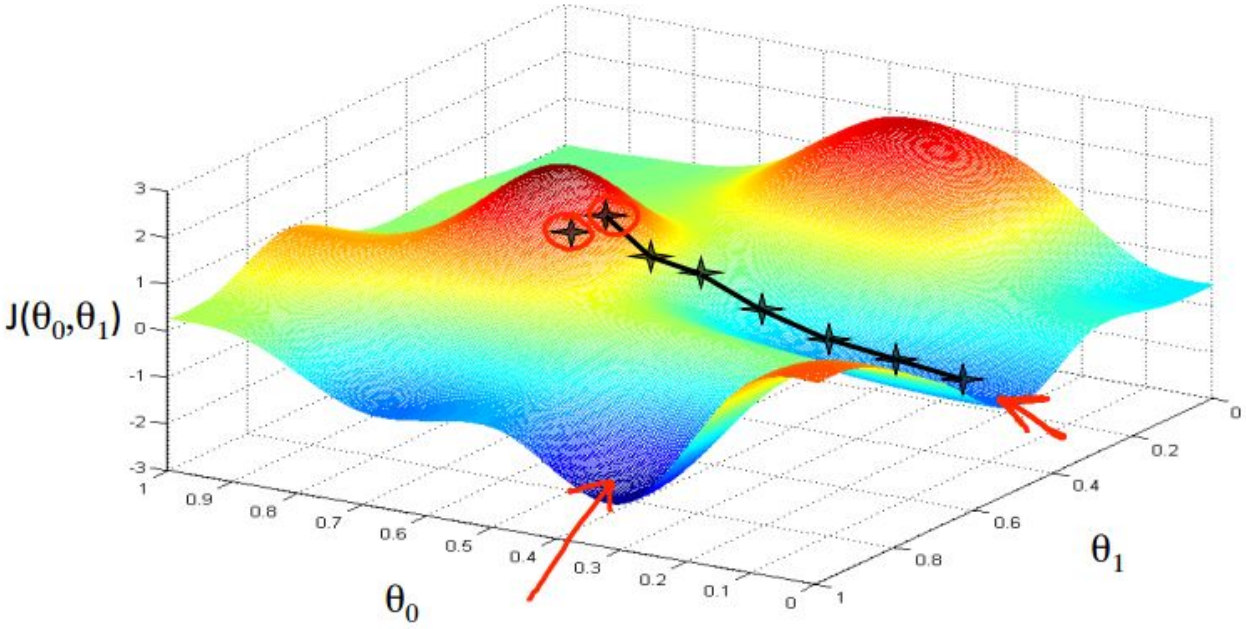
Những vấn đề nâng cao

Overfitting (Quá khớp)

Những vấn đề nâng cao - Gradient Descent tổng quát



Những vấn đề nâng cao - Gradient Descent tổng quát



Câu hỏi

Vì sao Cost Function của Linear Regression và Logistic Regression chỉ có duy nhất một cực tiểu địa phương?

Câu hỏi

Vì sao phải Update đồng thời? Update lần lượt thì có đúng không?

Câu hỏi

Làm thế nào để chọn α? Vì sao Cost Function đảm bảo giảm khi α đủ nhỏ?

Câu hỏi

Vì sao phần mẫu số của Hypothesis Function của Logistic Regression lại lấy lũy thừa ***e***? Lấy lũy thừa số khác có được không? Tại sao Cost Function lại phải lấy log? Bỏ log đi có được không?

Câu hỏi

Trong Normal Equation, tại sao có thể đảm bảo phương trình có nghiệm? Nhỡ $\boldsymbol{A^T A}$ không khả nghịch thì sao? Khi nào thì $\boldsymbol{A^T A}$ không khả nghịch?

Câu hỏi

Tại sao Feature Scaling lại nhanh hơn?

Câu hỏi

Tại sao khi Regularization, $\boldsymbol{\theta}$ nhỏ đi thì lại đỡ bị overfitting hơn? Tại sao đường cong lại đỡ uốn hơn?