Lecture 2: Linear Regression and Gradient Descent

Supervised learning

- Mô hình của một thuật toán học giám sát:

  Tranning Set ⇒ Learning Algorithm ⇒ Hypothesis

- Hypothesis có tác dụng thực hiện một tác vụ nào đó nhằm thảo mãn yêu cầu của dữ liệu đầu vào (Dự đoán, phân loại, ...)

Linear Regression

- Biểu diễn một Hypothesis:

$$h(x) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... = \sum_{i=0}^{n} \theta_i X_i$$

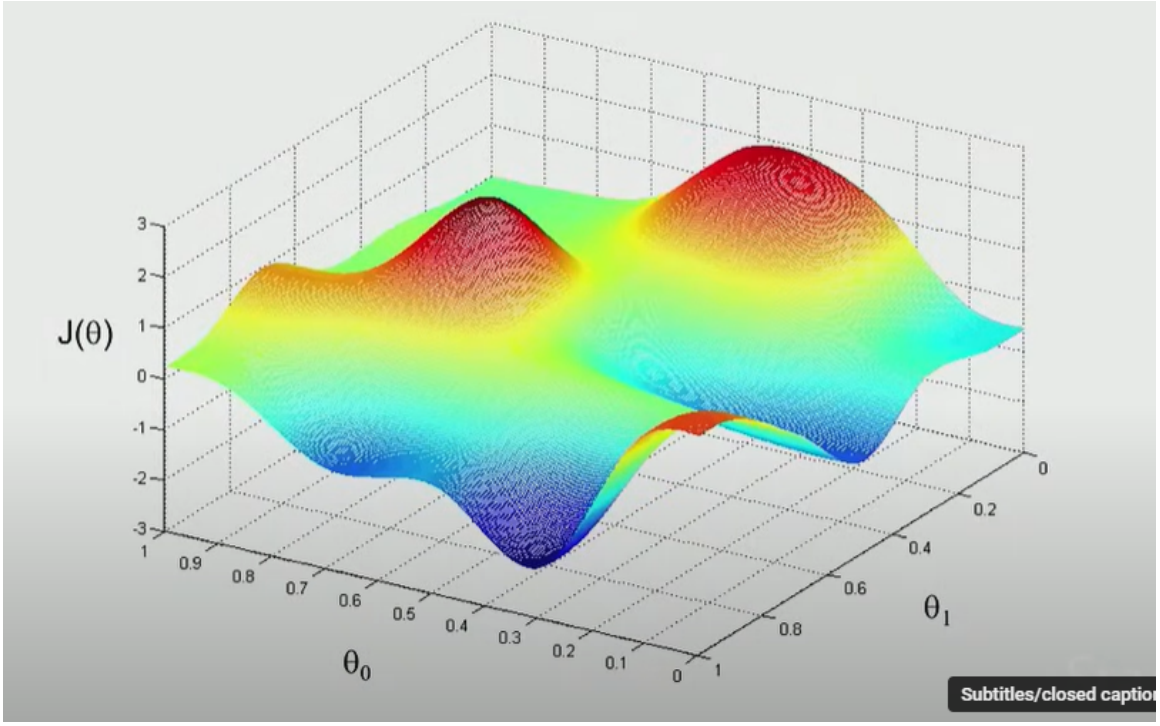
  Trong đó θ_0 và θ_1, θ_2 là tham số của mô hình, X_1, X_2 là các thuộc tính của dữ liệu và X_0 = 1. n là số thuộc tính của dữ liệu

- Nhiệm vụ của Learning Algorithm chính là tìm ra bộ tham số θ_i tối ưu nhất cho bài toán nhằm thực hiện yêu cầu dữ liệu đầu vào
- Linear Regression muốn minimize giá trị sau

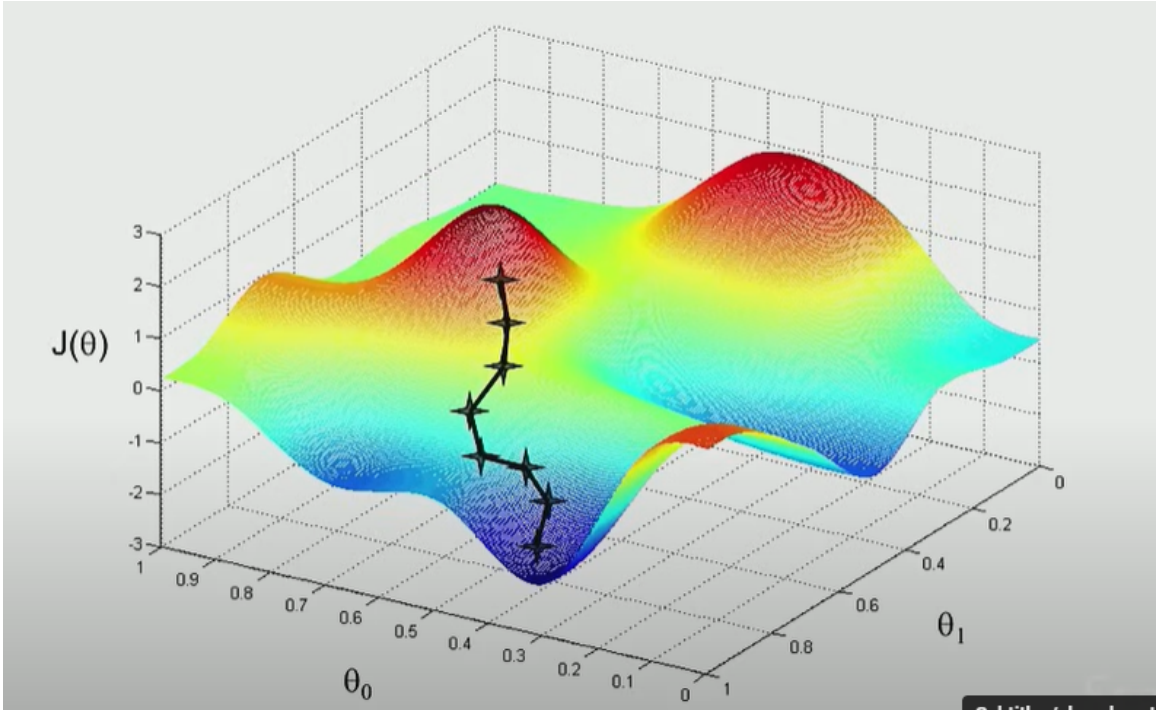
$$\frac{1}{2} \sum_{i=1}^{M} (h(x^i) - y)^2$$

Gradient Descent

- Bắt đầu bằng cách giá trị bất kì của tập θ
- Chuyển đổi dần dần các giá trị trong tập θ để giảm giá trị trên



- Yêu cầu của thuật toán là tìm các giá trị θ mà minimize được giá trị $J(\theta)$ như trong hình trên. Thuật toán sẽ từng bước tìm ra hướng đi để giảm tối đa giá trị này giống như hình sau:



- Để cập nhật giá trị θ, công thức sau được sử dụng:

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

Ở đây, α là tốc độ học (learning rate) và $\frac{\partial}{\partial \theta_j} J(\theta)$ là đạo hàm riêng của hàm chi phí $J(\theta)$ theo θ_j.

- Quá trình cập nhật trên được áp dụng cho đến khi mô hình hội tụ

Batch Gradient Descent

- Yêu cầu tính giá trị θ trên tất cả các điểm dữ liệu trước khi thực hiện cập nhật

Stochastic Gradient Descent (SGD)

- Thay vì tính θ giảm trên tất cả các điểm dữ liệu rồi mới cập nhật, thuật toán SGD thực hiện cập nhất trên từng điểm dữ liệu
- Thuật toán SGD có một vấn đề đó là nó sẽ không thực sự hội tụ do cập nhật trên từng điểm dữ liệu, thay vào đó ở các epoch cuối nó sẽ dịch chuyển xung quanh điểm local minimal

Normal equation

Phương trình thông thường trong gradient descent là một phương pháp toán học được sử dụng để tìm giá trị tối ưu của các tham số trong mô hình hồi quy tuyến tính mà không cần sử dụng vòng lặp hoặc tối ưu hóa gradient. Nó tính toán giá trị tối ưu bằng cách sử dụng công thức toán học trực tiếp.

Phương trình Normal dùng để tìm giá trị tối ưu cho tham số θ bằng cách giải phương trình sau:

$$\theta = (X^T X)^{-1} X^T Y$$

Trong phương trình trên, X là ma trận đầu vào của dữ liệu, Y là vector kết quả và θ là vector tham số cần tìm. $(X^T X)^{-1}$ đại diện cho nghịch đảo của ma trận $X^T X$.