Advanced Algorithms

Floriano Zini

Free University of Bozen-Bolzano
Faculty of Computer Science

Academic Year 2013-2014

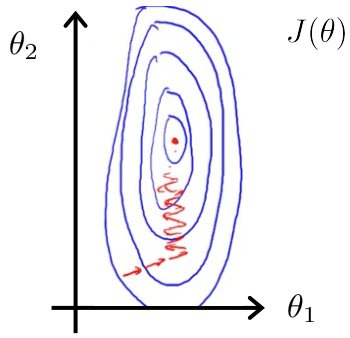
Lecture 12 – Linear regression (cont.)

These slides are taken from **Andrew Ng**, **Machine Learning** on **Coursera** - https://class.coursera.org/ml-003/lecture/preview

Practice advice 1: Feature Scaling

Idea: Make sure features are on a similar scale so that gradient descent can converge more quickly

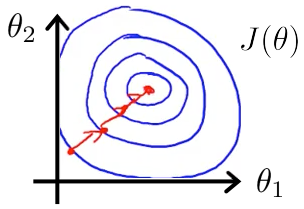
E.g. x_1 = size (0-2000 feet^2)
x_2 = number of bedrooms (1-5)



$$x_1 = \frac{size(feet^2)}{2000}$$

$$x_2 = \text{\# bedrooms / 5}$$

$$0 \le x_1 \le 1 \qquad 0 \le x_2 \le 1$$



Rule of thumb: get every feature into approximately a $-1 \le x_i \le 1$ range

Practice advice 2: Mean Normalization

Replace x_i with $x_i - \mu_i$ to make features have approximately zero mean (Do not apply to $x_0 = 1$)

E.g. $x_1 = \frac{size-1000}{2000}$ Average size = 1000

$x_2 = \frac{\#bedrooms-2}{5}$ Average # bedrooms = 2

$$-0.5 \le x_1 \le 0.5, -0.5 \le x_2 \le 0.5$$

General rule: $x_i \leftarrow \frac{x_i - \mu_i}{s_i}$

μ_i average value of x_i in training set

s_i range (max - min) or standard deviation of x_i

Suppose you are using a learning algorithm to estimate the price of houses in a city. You want one of your features x_i to capture the age of the house. In your training set, all of your houses have an age between 30 and 50 years, with an average age of 38 years. Which of the following would you use as features, assuming you use feature scaling and mean normalization?

- ○ $x_i = \text{age of house}$
- ○ $x_i = \dfrac{\text{age of house}}{50}$
- ○ $x_i = \dfrac{\text{age of house} - 38}{50}$
- ○ $x_i = \dfrac{\text{age of house} - 38}{20}$

Suppose you are using a learning algorithm to estimate the price of houses in a city. You want one of your features x_i to capture the age of the house. In your training set, all of your houses have an age between 30 and 50 years, with an average age of 38 years. Which of the following would you use as features, assuming you use feature scaling and mean normalization?

- ○ $x_i = \text{age of house}$
- ○ $x_i = \dfrac{\text{age of house}}{50}$
- ○ $x_i = \dfrac{\text{age of house} - 38}{50}$
- ✔ $x_i = \dfrac{\text{age of house} - 38}{20}$

Practice advice 3: Learning Rate

Gradient descent

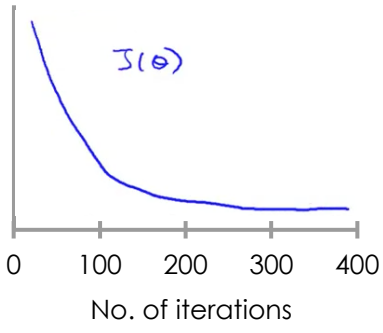
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

- "Debugging": How to make sure gradient descent is working correctly
- How to choose learning rate α

Practice advice 3: Learning Rate

Making sure gradient descent is working correctly

$\min_\theta J(\theta)$



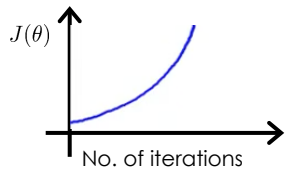
Example automatic convergence test:

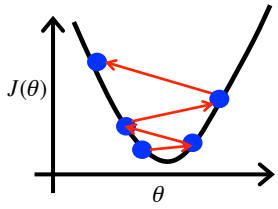
Declare convergence if $J(\theta)$ decreases by less than 10^{-3} in one iteration.

J(θ) should decrease after every iteration

Practice advice 3: Learning Rate

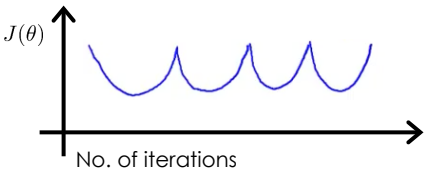
Making sure gradient descent is working correctly.





Gradient descent not working

Use smaller α

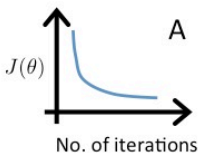


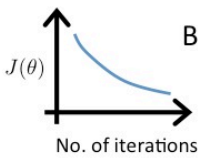
Use smaller α

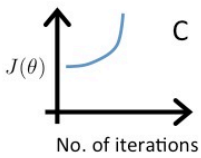
For sufficiently small α, $J(\theta)$ should decrease on every iteration

But if α is too small, gradient descent can be slow to converge

Suppose a friend ran gradient descent three times, with $\alpha = 0.01$, $\alpha = 0.1$, and $\alpha = 1$, and got the following three plots (labeled A, B, and C):



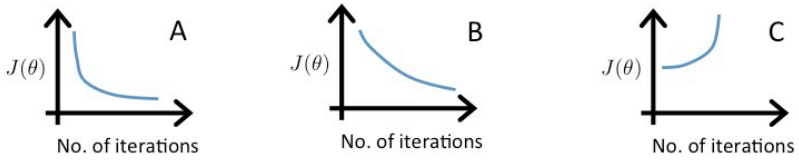




Which plots corresponds to which values of α?

- A is $\alpha = 0.01$, B is $\alpha = 0.1$, C is $\alpha = 1$.
- A is $\alpha = 0.1$, B is $\alpha = 0.01$, C is $\alpha = 1$.
- A is $\alpha = 1$, B is $\alpha = 0.01$, C is $\alpha = 0.1$.
- A is $\alpha = 1$, B is $\alpha = 0.1$, C is $\alpha = 0.01$.

Suppose a friend ran gradient descent three times, with $\alpha = 0.01$, $\alpha = 0.1$, and $\alpha = 1$, and got the following three plots (labeled A, B, and C):



Which plots corresponds to which values of α?

- ○ A is $\alpha = 0.01$, B is $\alpha = 0.1$, C is $\alpha = 1$.
- ✔ A is $\alpha = 0.1$, B is $\alpha = 0.01$, C is $\alpha = 1$.
- ○ A is $\alpha = 1$, B is $\alpha = 0.01$, C is $\alpha = 0.1$.
- ○ A is $\alpha = 1$, B is $\alpha = 0.1$, C is $\alpha = 0.01$.

Practice advice 3: Learning Rate

Summary:

- If α is too small: slow convergence
- If α is too large: $J(\theta)$ may not decrease on every iteration; may not converge.

To choose α, try

…, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, ….