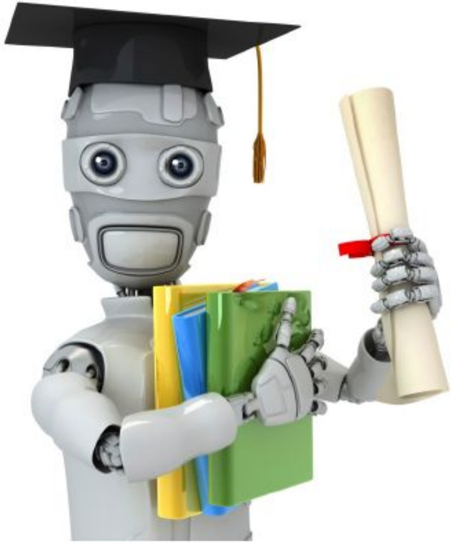
Introduction to ML

George Igwegbe
Artificial Intelligence
Saturday (AI6), Lagos

WK3

1. Welcome
2. What is ML?
3. Supervised Learning
4. Unsupervised Learning
5. Model Representation
6. Cost function
7. Cost function 2
8. Gradient Descent
9. Gradient Descent 2
10. Gradient Descent for Linear Regression
11. What's Next
12. Matrices and Vectors
13. Addition and Scalar Multiplication
14. Matrix Vector Multiplication
15. Matrix Vector Multiplication 2
16. Matrix Vector Multiplication 3
17. Matrix Vector Properties
18. Inverse and Transpose
19. Multiple Features



Machine Learning

BUZZWORD

Multivariate Linear Regression

Singular

Degenerate

Normal Equation

Condition of adding matrices

Condition of multiplying matrices

Dimension of matrix

Transpose of matrix

1-indexed vs 0 -indexed

Inverse of matrix

MULTIPLE FEATURES

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

- Difficult to visualize->3D
- Notations becomes complicated

LINEAR ALGEBRA

Size (feet²)	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

$$X = \begin{bmatrix} 2104 & 5 & 1 & 45 \\ 1416 & 3 & 2 & 40 \\ 1534 & 3 & 2 & 30 \\ 852 & 2 & 1 & 36 \end{bmatrix} \qquad y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 172 \end{bmatrix}$$

MATRIX

VECTOR

MATRIX: Rectangular array of numbers

$$X = \begin{bmatrix} 2104 & 5 & 1 & 45 \\ 1416 & 3 & 2 & 40 \\ 1534 & 3 & 2 & 30 \\ 852 & 2 & 1 & 36 \end{bmatrix}$$

Dimension of matrix: number of rows x number of columns

$\boxed{\text{(? x ?) matrix}}$

Matrix Elements (entries of matrix)

$$A = \begin{bmatrix} 1402 & 191 \\ 1371 & 821 \\ 949 & 1437 \\ 147 & 1448 \end{bmatrix}$$

$A_{ij} =$ "i, j entry" in the i^{th} row, j^{th} column.

$A_{11} = 1402$

$A_{12} = 191$

$A_{32} = 1437$

$A_{41} = 147$

$\cancel{A_{43}}$ = undefined (error)

VECTOR: An n x 1 matrix.

$$y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 172 \end{bmatrix}$$

(? x ?) matrix

Dimension of matrix: number of rows x number of columns

VECTOR: An nx1 matrix.

$$y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

n = 4

← 4-dimensional vector.

$y_i = i^{th}$ element

$y_1 = 460$

$y_2 = 232$

$y_3 = 315$

1-indexed vs 0-indexed:

$$y[1] \quad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \qquad y = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} \leftarrow y[0]$$

1-indexed 0-indexed

Matrix Addition

$$\underbrace{\begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix}}_{\substack{3\times 2 \\ \text{matrix}}} + \underbrace{\begin{bmatrix} 4 & 0.5 \\ 2 & 5 \\ 0 & 1 \end{bmatrix}}_{3\times 2} = \underbrace{\begin{bmatrix} 5 & 0.5 \\ 4 & 10 \\ 3 & 2 \end{bmatrix}}_{3\times 2}$$

$$\cancel{\underbrace{\begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix}}_{3\times 2} + \underbrace{\begin{bmatrix} 4 & 0.5 \\ 2 & 5 \end{bmatrix}}_{2\times 2}} = \text{error}$$

Scalar Multiplication

$$3 \times \begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix} = \boxed{\begin{bmatrix} 3 & 0 \\ 6 & 15 \\ 9 & 3 \end{bmatrix}} = \begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix} \times 3$$

3×2 $\quad$ 3×2

$$\boxed{\begin{bmatrix} 4 & 0 \\ 6 & 3 \end{bmatrix} / 4} = \frac{1}{4} \begin{bmatrix} 4 & 0 \\ 6 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{3}{2} & \frac{3}{4} \end{bmatrix}$$

Condition of multiplication of matrix

$$A \times x = y$$

$$[\quad] \times [\quad] = [\quad]$$

m x n matrix (m rows, n columns)	n x 1 matrix (n-dimensional vector)	m-dimensional vector

To get y_i, multiply A's i^{th} row with elements of vector x, and add them up.

Example of Matrix Multiplication

$$\begin{bmatrix} 1 & 2 & 1 & 5 \\ 0 & 3 & 0 & 4 \\ -1 & -2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} \boxed{14} \\ \boxed{13} \\ \boxed{-7} \end{bmatrix} = \begin{bmatrix} 14 \\ 13 \\ -7 \end{bmatrix}$$

$\boxed{3} \times 4$ $\quad 4 \times 1 \quad 3 \times 1$

$$1 \times 1 + 2 \times 3 + 1 \times 2 + 5 \times 1 = 14$$

$$0 \times 1 + 3 \times 3 + 0 \times 2 + 4 \times 1 = 13$$

$$-1 \times 1 + (-2) \times 3 + 0 \times 2 + 0 \times 1 = -7$$

Application of Matrix Multiplication

House sizes:

→ 2104
→ 1416
→ 1534
→ 852

$$h_\theta(x) = -40 + 0.25x$$

$h_\theta(x)$

Matrix 4x2 × Vector 2x1 = 4x1 matrix

$$\begin{bmatrix}1 & 2104\\1 & 1416\\1 & 1534\\1 & 852\end{bmatrix} \times \begin{bmatrix}-40\\0.25\end{bmatrix} = \begin{array}{c|l} \left[\begin{array}{c} -40\times1 + 0.25\times2104 \\ -40\times1 + 0.25\times1416 \\ \\ \\ \end{array}\right] & \begin{array}{l} \leftarrow h_\theta(2104) \\ \leftarrow h_\theta(1416) \\ \\ \\ \end{array} \end{array}$$

Condition of multiplication of matrix

$$A \times B = C$$

A	B	C
m x n matrix (m rows, n columns)	n x o matrix (n rows, o columns)	m x o matrix

Example of Matrix Multiplication

$$\overset{2\times 2}{\begin{bmatrix} 1 & 3 \\ 2 & 5 \end{bmatrix}} \overset{2\times 2}{\begin{bmatrix} 0 & 1 \\ 3 & 2 \end{bmatrix}} = \overset{2\times 2}{\begin{bmatrix} 9 & 4 \\ 15 & 12 \end{bmatrix}}$$

$$\begin{bmatrix} 1 & 3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} = \begin{bmatrix} 1\times 0 + 3\times 3 \\ 2\times 0 + 5\times 3 \end{bmatrix} = \cancel{\begin{bmatrix} 9 \\ 15 \end{bmatrix}} \begin{bmatrix} 9 \\ 15 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1\times 1 + 3\times 2 \\ 2\times 1 + 5\times 2 \end{bmatrix} = \begin{bmatrix} 7 \\ 12 \end{bmatrix}$$

House sizes:

$$\begin{matrix} 2104 \\ 1416 \\ 1534 \\ 852 \end{matrix}$$

Have 3 competing hypotheses:

1. $h_\theta(x) = -40 + 0.25x$
2. $h_\theta(x) = 200 + 0.1x$
3. $h_\theta(x) = -150 + 0.4x$

Matrix

Matrix

$$\begin{bmatrix} 1 & 2104 \\ 1 & 1416 \\ 1 & 1534 \\ 1 & 852 \end{bmatrix} \times \begin{bmatrix} -40 & 200 & -150 \\ 0.25 & 0.1 & 0.4 \end{bmatrix} = \begin{bmatrix} 486 & 410 & 692 \\ 314 & 342 & 416 \\ 344 & 353 & 464 \\ 173 & 285 & 191 \end{bmatrix}$$

Prediction of first h_θ

Predictions of 2nd h_θ

Commutativity: Properties of Matrices

$3 \times 5 = 5 \times 3$ "Commutative"

Let A and B be matrices. Then in general,

$A \times B \neq B \times A$. (not commutative.)

E.g. $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$

$\neq$

$\begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \end{bmatrix}$

$\underset{m \times n}{A} \times \underset{n \times m}{B}$

$A \times B$ is $m \times m$

$B \times A$ is $n \times n$

Associativity: Properties of Matrices

$3 \times 5 \times 2$

$3 \times 10 = 30 = 15 \times 2$

$3 \times (5 \times 2) = (3 \times 5) \times 2$

"Associative"

$A \times (B \times C)$

$(A \times B) \times C$

$$A \times B \times C.$$

Let $D = B \times C$. Compute $A \times D$.

Let $E = A \times B$. Compute $E \times C$.

$A \times (B \times C)$

$(A \times B) \times C$

Same answer.

Identity : Properties of Matrices

Identity Matrix

1 is identity.

$\boxed{1 \times z = z \times 1 = z}$ for any z

Denoted I (or $\boxed{I_{n \times n}}$).

Examples of identity matrices:

$[1]$ 1×1

$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ 2 x 2

$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ 3 x 3

$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ 4 x 4

Informally:

$\begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & \ddots \\ & & & & 1 \end{bmatrix}$

For any matrix A,

$$\underset{m \times n}{A} \cdot \underset{n \times n}{\boxed{I}} = \underset{m \times m}{\boxed{I}} \cdot \underset{m \times n}{A} = \underset{m \times n}{A}$$

$I_{n \times n}$

Note:

$AB \neq BA$ in general

$AI = IA$ ✓

Inverse: Properties of Matrix

I = "identity"

$3\,\underbrace{\boxed{(3^{-1})}}_{\frac{1}{3}} = 1$

$12 \times \underbrace{(12^{-1})}_{\frac{1}{12}} = 1$

$0\,(0^{-1})$ undefined

Not all numbers have an inverse.

Matrix inverse: square matrix (#rows = #columns) A^{-1}

If A is an m x m matrix, and if it has an inverse,

$$\rightarrow \quad A(A^{-1}) = A^{-1}A = I.$$

E.g.

$$\underbrace{\begin{bmatrix} 3 & 4 \\ 2 & 16 \end{bmatrix}}_{\substack{2\times 2 \\ A}} \underbrace{\begin{bmatrix} 0.4 & -0.1 \\ -0.05 & 0.075 \end{bmatrix}}_{A^{-1}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_{2\times 2}$$

$A^{-1}A$

Matrices that don't have an inverse are "singular" or "degenerate"

Transpose: Properties of Matrix

Example: $A = \begin{bmatrix} 1 & 2 & 0 \\ 3 & 5 & 9 \end{bmatrix}$ 2×3

$B = A^T = \begin{bmatrix} 1 & 3 \\ 2 & 5 \\ 0 & 9 \end{bmatrix}$ 3×2

Let A be an m x n matrix, and let $B = A^T$.

Then B is an n x m matrix, and

$$B_{ij} = A_{ji}.$$

$B_{12} = A_{21} = 2$

$B_{32} = 9$ $\quad$ $A_{23} = 9.$

MULTIPLE FEATURES

Size (feet^2) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178

MULTIPLE FEATURES

Size (feet²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

$m = 47$

Notation:

- n = number of features $\quad n=4$
- $x^{(i)}$ = input (features) of i^{th} training example.
- $x_j^{(i)}$ = value of feature j in i^{th} training example.

$$x^{(2)} = \begin{bmatrix} 1416 \\ 3 \\ 2 \\ 40 \end{bmatrix}$$

$$x_3^{(2)} = 2$$

Representation of Hypothesis

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

For convenience of notation, define $\boxed{x_0 = 1.}$ $\quad (x_0^{(i)} = 1)$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

$$h_\theta(x) = \theta_0 \overset{=1}{x_0} + \theta_1 x_1 + \cdots + \theta_n x_n$$

$$= \boxed{\theta^T x.}$$

$$\underbrace{\begin{bmatrix} \theta_0 & \theta_1 & \cdots & \theta_n \end{bmatrix}}_{\theta^T} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$(n+1) \times 1$ matrix

$\theta^T x$

x

Multivariate linear regression.

Hypothesis: $h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

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

Cost function:

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

Hypothesis: $h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

Parameters: $\cancel{\theta_0, \theta_1, \ldots, \theta_n}$ $\underline{\theta}$ n+1 - dimensional vector

VECTOR

Cost function:

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

$J(\theta)$

Gradient descent:

Repeat $\{$

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

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

Gradient descent:

Repeat $\{$

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

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

VECTOR

Gradient Descent

Previously (n=1):

Repeat {

$$\theta_0 := \theta_0 - \alpha \underbrace{\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})}_{\frac{\partial}{\partial \theta_0} J(\theta)}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$$

(simultaneously update θ_0, θ_1)

}

New algorithm $(n \geq 1)$:

Repeat {

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

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

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

}

$$\theta_0 := \theta_0 - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_1^{(i)}$$

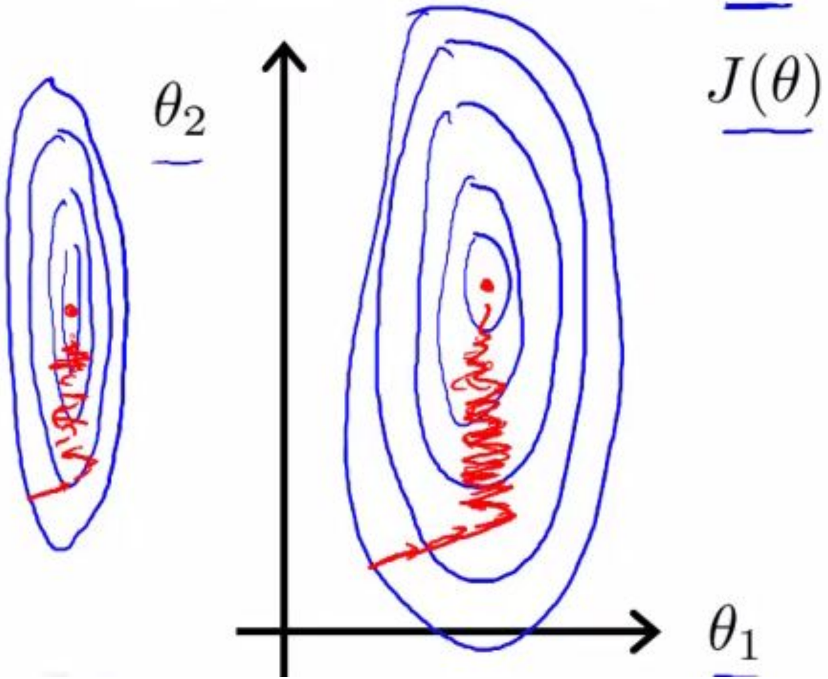
$$\theta_2 := \theta_2 - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_2^{(i)}$$

Feature Scaling

Idea: Make sure features are on a similar scale.

E.g. x_1 = size (0-2000 feet^2) $\leftarrow$

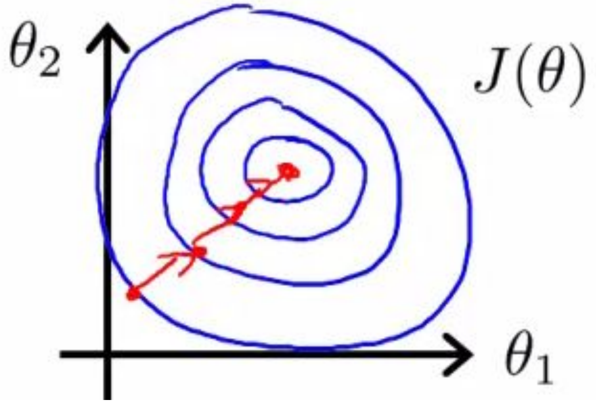
x_2 = number of bedrooms (1-5) $\leftarrow$



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

$$\rightarrow x_2 = \frac{\text{number of bedrooms}}{5}$$

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



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. $\boxed{x_1 = \frac{size - 1000}{2000}}$ Average size = 100

$\boxed{x_2 = \frac{\#bedrooms - 2}{5}}$ 1-5 bedrooms

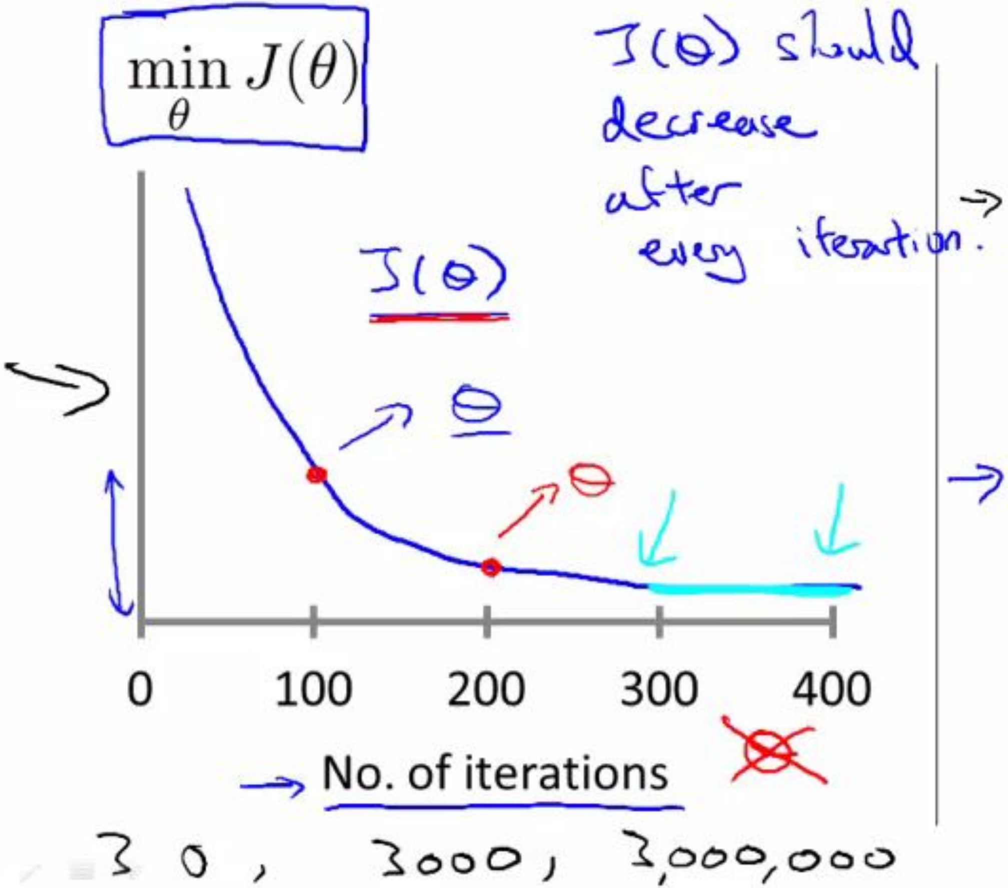
$\boxed{-0.5 \le x_1 \le 0.5}$, $\boxed{-0.5 \le x_2 \le 0.5}$

$x_1 \leftarrow \frac{x_1 - \mu_1}{s_1}$ ← avg value of x_1 in training set

s_1: range (max - min) (or standard deviation)

$x_2 \leftarrow \frac{x_2 - \mu_2}{s_2}$

Making sure gradient descent is working correctly.

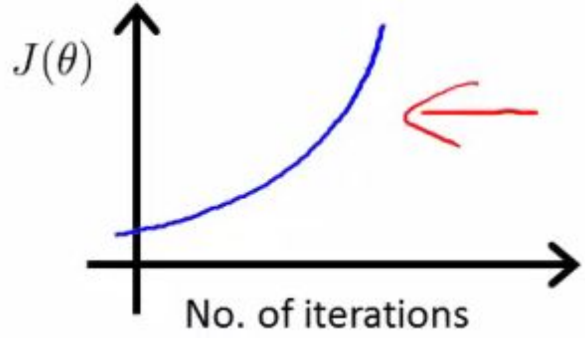


→ Example automatic convergence test:

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

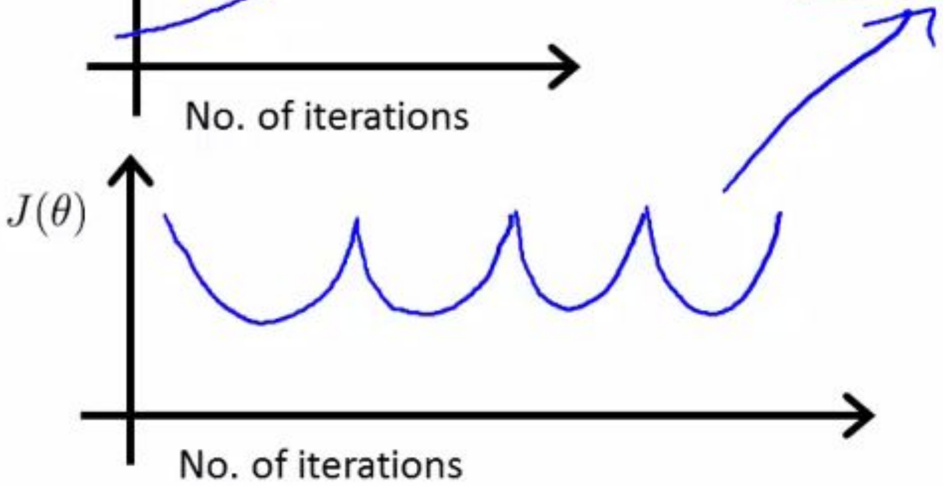
ε

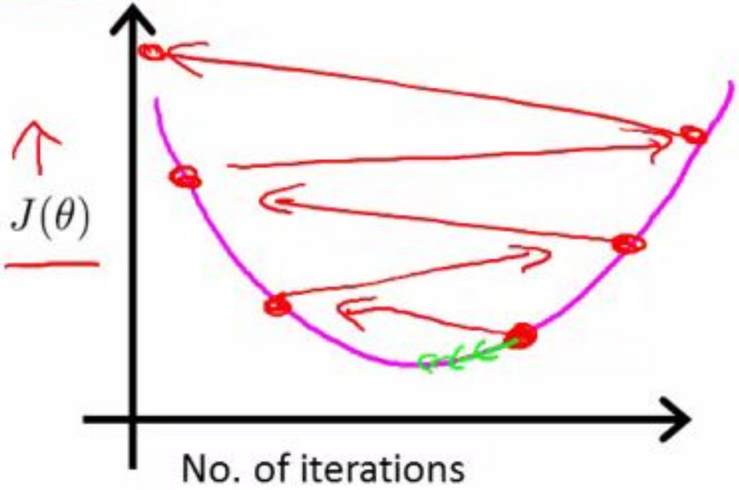
Making sure gradient descent is working correctly.



Gradient descent not working.

$\boxed{\text{Use smaller } \alpha.}$

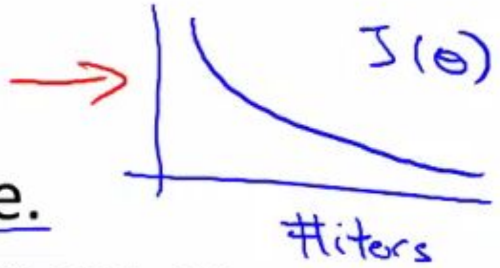




- For sufficiently small α, $J(\theta)$ should decrease on every iteration.
- But if α is too small, gradient descent can be slow to converge.

Summary:

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



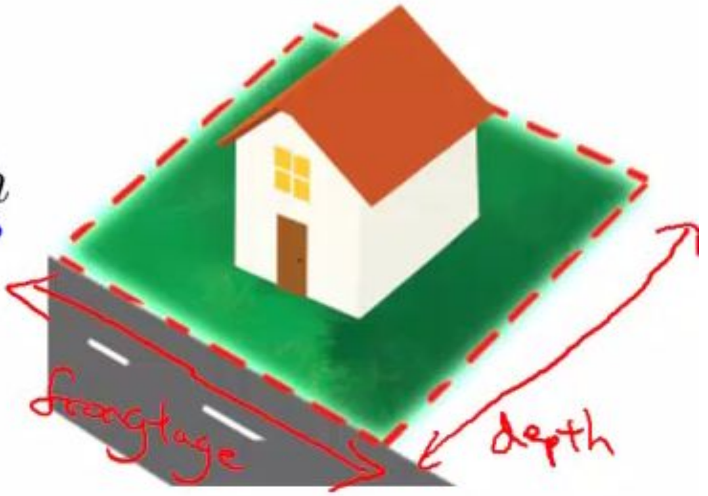
To choose α, try

$$\dots, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, \dots$$

3x ≈3x 3x ≈3x

Housing prices prediction

$$h_\theta(x) = \theta_0 + \theta_1 \times \underbrace{frontage}_{\underline{x_1}} + \theta_2 \times \underbrace{depth}_{\underline{x_2}}$$



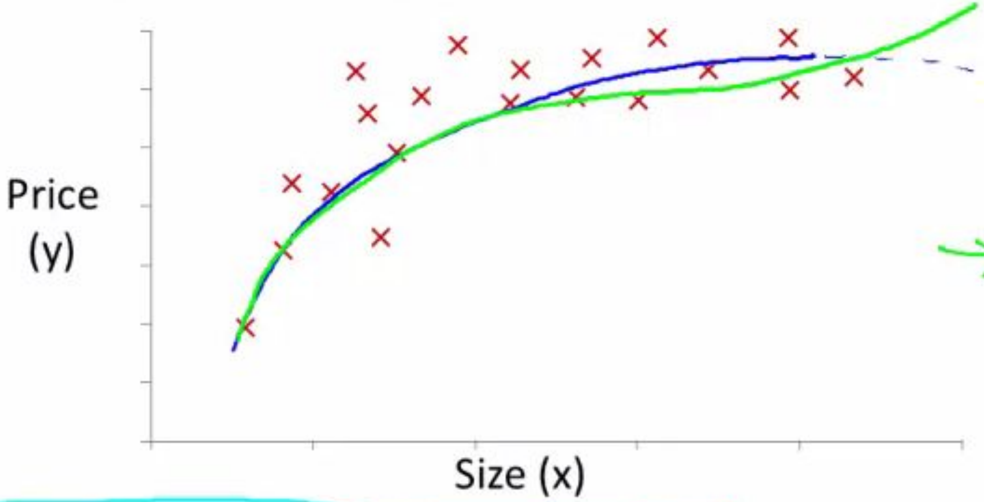
<u>Area</u>

$$x = \underline{frontage * depth}$$

$$h_\theta(x) = \theta_0 + \theta_1 x$$

← land area

Polynomial regression



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

$$\boxed{\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3}$$

$$\boxed{h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3}$$

$$= \theta_0 + \theta_1 (size) + \theta_2 (size)^2 + \theta_3 (size)^3$$

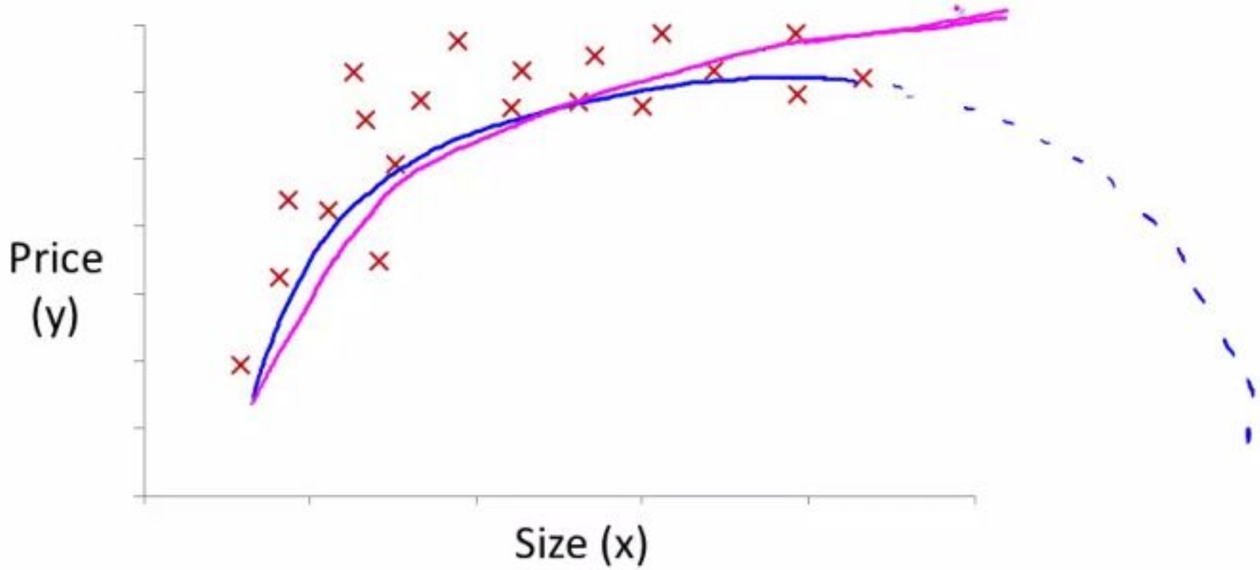
$$x_1 = (size)$$
$$x_2 = (size)^2$$
$$x_3 = (size)^3$$

Size: 1-1000

Size^2: 1-1000,000

Size^3: 1-10^9

Choice of features



$$\rightarrow \quad h_\theta(x) = \theta_0 + \theta_1(size) + \theta_2(size)^2$$

$$\rightarrow \quad \boxed{h_\theta(x) = \theta_0 + \theta_1(size) + \theta_2\sqrt{(size)}}$$

NORMAL EQUATION

In $\min J(\theta_0, \theta_1)$, solve for θ_0, θ_1 exactly, without needing iterative algorithm (gradient descent).

Examples: $m = 4.$

x_0	Size (feet²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
1	2104	5	1	45	460
1	1416	3	2	40	232
1	1534	3	2	30	315
1	852	2	1	36	178

$$X = \underbrace{\begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix}}_{m \times (n+1)} \qquad y = \underbrace{\begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}}_{m\text{-dimensional vector}}$$

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

m **examples** $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$ **;** n **features.**

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

$$X = \begin{bmatrix} \text{---} & (x^{(1)})^T & \text{---} \\ \text{---} & (x^{(2)})^T & \text{---} \\ & \vdots & \\ \text{---} & (x^{(m)})^T & \text{---} \end{bmatrix}$$

(design matrix)

$m \times (n+1)$

E.g. If $x^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \end{bmatrix}$

$$X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_2^{(1)} \\ \vdots & \vdots \\ 1 & x_m^{(1)} \end{bmatrix} \qquad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

$m \times 2$

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

$(X^T X)^{-1}$ is inverse of matrix $X^T X$.

Set $A = X^T X$

$\boxed{(X^T X)^{-1}} = A^{-1}$

Octave: `pinv(X'*X)*X'*y`

X' X^T

pinv$(X^T * X) * X^T * y$

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

$\min_\theta J(\theta)$

Feature Scaling

$0 \leq x_1 \leq 1$

$0 \leq x_2 \leq 1000$

$0 \leq x_3 \leq 10^{-5}$

m training examples, n features.

Gradient Descent	Normal Equation
• Need to choose α.	• No need to choose α.
• Needs many iterations.	• Don't need to iterate.
• Works well even when n is large.	• Need to compute $(\boxed{X^T X})^{-1}$ $n \times n$ $O(n^3)$
	• Slow if n is very large.

m **training examples,** n **features.**

Gradient Descent	Normal Equation
• Need to choose α.	• No need to choose α.
• Needs many iterations.	• Don't need to iterate.
• Works well even when n is large.	• Need to compute $(\boxed{X^T X})^{-1}$ $n \times n$ $O(n^3)$
	• Slow if n is very large.

- n = 10 ^ 6(1 million features)

- n = 100
- n = 1000
- n = 10000

What if $X^T X$ is non-invertible?

- Redundant features (linearly dependent).

  E.g. $x_1 =$ size in feet^2

  $x_2 =$ size in m^2

- Too many features (e.g. $m \leq n$).
  - Delete some features, or use regularization.

Hinton’s Closing Prayer

Our father who art in n-dimensions

hallowed by the backprop,

thy loss be minimized,

thy gradients unvarnished,

on earth as it is in Euclidean space.

Give us this day our daily hyperparameters,

and forgive us our large learning rates,

as we forgive those whose parameters diverge,

and lead us not into discrete optimization,

but deliver us from local optima.

For thine are dimensions,

and the GPUs, and the glory,

forever and ever. Dropout.

From buZZrobot