# Lecture 2

Saturday, April 18, 2020 8:33 PM

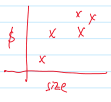
April 18th

Linear Regression and Gradient Descent

Example: housing dataset

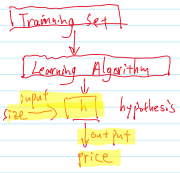
given size and price

| Size | Price (thousands) |
|---|---|
| 2104 | 400 |
| 1416 | 232 |
| 1534 | 315 |
| 852 | 178 |
| ⋮ | ⋮ |




Goal: fit a straight

How?




How to represent $h$?

$h(x) = \theta_0 + \theta_1 x$

You can have more features.

| Size | # of rooms | Price |
|---|---|---|
| ⋮ | ⋮ | ⋮ |

more generally, $x_1$: size, $x_2$: # bedrooms

$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

$h(x) = \sum_{j=0}^{2} \theta_j x_j$

where $x_0 = 1$

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$ $X = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$

$\theta$: parameters

the goal of the learning algorithms is to choose parameter $\theta$ that allows you to make good predictions about price of houses.

More Notations:

$m$ = # of training examples (# of rows from dataset)

$X$ = "inputs" / features

$y$ = "output" / target variables

$(x, y)$ = training example

$(x^{(i)}, y^{(i)})$ = $i^{th}$ training example

example: using above datasets

$x_1^{(1)} = 2104$

$x_1^{(2)} = 1416$

$n$ = # of features

How to choose $\theta$?

choose $\theta$ such that $h(x) \approx y$ for training example

notation for representing $h$ depends on input $x$ and parameter $\theta$

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

We want to minimize the equation $(h_\theta(x) - y)^2$

$$\underset{\theta}{\text{minimize}} \underbrace{\frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}_{J(\theta)}$$

minimize $J(\theta)$ over $\theta$

use gradient decent to minimize

start with some initial value $\theta$ (say $\theta = \vec{0}$)

Keep changing $\theta$ to reduce $J(\theta)$

One implementation:

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

($:=$ assignment notation; $\alpha$ learning rate; $\frac{\partial}{\partial \theta_j}$ partial derivative)

$$\frac{\partial J(\theta)}{\partial \theta_j} = 2\left[\frac{1}{2}\left(h_\theta(x) - y\right)\right] \cdot \underbrace{\frac{\partial}{\partial \theta_j}\left(h_\theta(x) - y\right)}_{\text{chain rule}}$$

$$= h_\theta(x) - y \cdot \frac{\partial}{\partial \theta_j}\left[\theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n - y\right]$$

all the terms become zero except the $j$th term, which leave us with $\theta_j x_j$ and their derivative give us $x_j$

$$= \left(h_\theta(x) - y\right) \cdot x_j$$

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

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

Loop {

for i = 1 to m, {

$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$ (for every j)

}

}

Stochastic gradient descent

looking at random training example at each iteration and look for improvement in the next iteration

would be helpful if m, # of training is large.

don't have to scan entire dataset (compares to batch gradient descent)

---

# The normal equations

Notations:

$\nabla_\theta J(\theta)$ derivative of $J(\theta)$ with respect to $\theta$ $= \begin{bmatrix} \frac{\partial J}{\partial \theta_0} \\ \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \end{bmatrix}$

$\theta \in \mathbb{R}^{n+1}$

example: $A \in \mathbb{R}^{2\times 2}$

$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$

$f(A)$ $\quad f: \mathbb{R}^{2\times 2} \to \mathbb{R}$ $\quad \nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} \end{bmatrix}$

function

if $f(A) = A_{11} + A_{12}^2$

then $f\left(\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}\right) = 5 + 6^2 \longrightarrow \nabla_A f(A) = \begin{bmatrix} 1 & 2A_{12} \\ 0 & 0 \end{bmatrix}$

$\nabla_\theta J(\theta) \overset{\text{set}}{=} 0$

More Notation: if A is a nxn square matrix

then the trace of A is defined to be its diagonal entries:

$\operatorname{tr} A = \sum_{i=1}^{n} A_{ii}$

some properties of $\operatorname{tr} A$:

$\operatorname{tr} A = \operatorname{tr} A^T$

$\operatorname{tr}(A+B) = \operatorname{tr} A + \operatorname{tr} B$

$\operatorname{tr} aA = a \operatorname{tr} A$

$\operatorname{tr} ABC = \operatorname{tr} ACB$

$\nabla_A \operatorname{tr} AA^T C = CA + C^T A$

if $f(A) = \operatorname{tr} AB$ then $\nabla_A f(A) = B^T$

Express

$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h(x^{(i)}) - y^{(i)}\right)^2$ in Matrix vector notation

$X\theta = \begin{bmatrix} — (x^{(1)})^T — \\ — (x^{(2)})^T — \\ \vdots \\ — (x^{(m)})^T — \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} x^{(1)T}\theta \\ x^{(2)T}\theta \\ \vdots \\ x^{(m)T}\theta \end{bmatrix} = \begin{bmatrix} h_\theta(x^{(1)}) \\ \vdots \\ h_\theta(x^{(m)}) \end{bmatrix}$

$\vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix}$ $\quad \begin{bmatrix} h_\theta(x^{(1)}) - y^{(1)} \end{bmatrix}$ $\quad z^T z = \sum z^2$

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} \qquad X\theta - y = \begin{bmatrix} h_\theta(x^{(1)}) - y^{(1)} \\ \vdots \\ h_\theta(x^{(m)}) - y^{(m)} \end{bmatrix} \qquad z^T z = \sum_i z^2$$

thus: $J(\theta) = \frac{1}{2}\sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2$

$$= \frac{1}{2}(X\theta - y)^T (X\theta - y)$$

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(X\theta - y)^T (X\theta - y)$$

$$= \frac{1}{2}\nabla_\theta (\theta^T X^T - y^T)(X\theta - y)$$

$$= \frac{1}{2}\nabla_\theta \left[\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y\right]$$

$$= \frac{1}{2}\left[X^T X\theta + X^T X\theta - X^T y - X^T y\right]$$

$$= X^T X\theta - X^T y \overset{\text{set}}{=} \vec{0}$$

$$X^T X\theta = X^T y \quad \text{"Normal equation"}$$

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

$\nabla_{A^T} f(A) = (\nabla_A f(A))^T$

$\nabla_A \operatorname{tr} ABA^T C = CAB + C^T AB^T$

$\nabla_{A^T} \operatorname{tr} ABA^T C = B^T A^T C^T + BA^T C$

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(X\theta - \vec{y})^T (X\theta - \vec{y})$$

$$= \frac{1}{2}\nabla_\theta \left(X\theta^2 - 2X\theta\vec{y} + \vec{y}^2\right) \to \frac{1}{2}\nabla_\theta (\theta^T X^T X\theta - \theta^T X^T \vec{y} - \vec{y}^T X\theta + \vec{y}^T \vec{y})$$