The Normal Equation:

May 13, 2021

The normal equation is a method to solve for the θ that minimizes $J(\theta)$ in one step, as opposed to the numerous steps of g.d.

e.g. $J(\theta) = a\theta^2 + b\theta + c$, $\theta \in \mathbb{R}$

$\frac{d}{d\theta} J(\theta) = 0 \rightarrow$ find min

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

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

$\frac{\partial}{\partial \theta_j} J(\theta) = 0$ for $j = 0, 1 \ldots n$

Then solve for $\theta_0, \theta_1 \ldots \theta_n$

e.g.

	x_0	x_1	x_2	x_3	x_4	y
$x^{(1)}$	1	2104	5	1	45	460
$x^{(2)}$	1	1416	3	2	40	232
$x^{(3)}$	1	1534	3	2	30	315
$x^{(4)}$	1	852	2	1	36	178

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

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

* Solving system of linear equations formula

* If you're using the normal equation, you don't need to use feature scaling.

Benefits of Normal Equation:
- don't need to choose α
- don't need to iterate

Drawbacks:
- doesn't work well for large # of features

↳ runtime is $O(n^3)$, very slow

↳ G.D. is $O(kn^2)$

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

Reasons:

1. Redundant features (lin dep cols)
   ↳ e.g. x_1 = size in ft^2
   x_2 = size in m^2

   * Invertible iff full rank
   ↳ Invertible Matrix Theorem

2. Too many features ($m \leq n$) for too little data
   - delete some features or use regularization