Lecture 2

Normal equation

Provides a way to minimize the cost function $J(\theta)$ by choosing θ analytically.

Remember that...

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 = \frac{1}{2m}\sum_{i=1}^{m}(y_i - h_\theta(x_i))^2 = \frac{1}{2m}\left(y - \theta^T x\right)^2 = \frac{1}{2m}\left(y - \theta^T x\right)' \; ($$

So,

$$\min_\theta J(\theta) = \min_\theta \frac{1}{2m}\left(y - \theta^T x\right)' \left(y - \theta^T x\right) = \min_\theta \left(y - \theta^T x\right)' \left(y - \theta^T x\right) \tag{2}$$

(in the last step, the constant doesn't make difference as we're minimizing in relation to θ).

Note: $\theta^T x = x\theta$

Problem:

$$\min_\theta J(\theta) = \min_\theta (y - x\theta)' (y - x\theta) = \min_\theta \|y - x\theta\|^2 = \min_\theta \|y - h_\theta(x)\|^2 \tag{3}$$

Distributing the product in the second equality of (3), we get:

$$J(\theta) = (y - x\theta)' (y - x\theta) = y' y - y' x\theta - \theta' x' y + \theta' x' x\theta \tag{4}$$

Derivating in relation to θ...

$$\frac{\partial}{\partial\theta} J(\theta) = -y' x - x' y + 2x' x\theta = 0 \;\; \Rightarrow \;\; 2x' y = 2x' x\theta \;\; \Rightarrow \;\; \theta = (x' x)^{-1} x' y \tag{5}$$

So, the normal equation formula is: $\theta = (x' x)^{-1} x' y$

Examples: $m = 4.$

	Size (feet²)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
x_0	x_1	x_2	x_3	x_4	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

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

$$\boxed{\theta = (X^TX)^{-1}X^Ty}$$

p.s.: There is no need to do feature scaling with the normal equation.

> The following is a comparison of gradient descent and the normal equation:

Gradient Descent	Normal Equation
Need to choose alpha	No need to choose alpha
Needs many iterations	No need to iterate
$O(kn^2)$	$O(n^3)$, need to calculate inverse of X^TX
Works well when n is large	Slow if n is very large

What if $(X'X)$ is noninvertible?

Note: When implementing the normal equation in octave we want to use the 'pinv' function rather than 'inv.'

If X^TX is noninvertible, the common causes might be having :

- Redundant features, where two features are very closely related (i.e. they are linearly dependent)
- Too many features (e.g. $m \leq n$). In this case, delete some features or use "regularization" (to be explained in a later lesson).

Solutions to the above problems include deleting a feature that is linearly dependent with another or deleting one or more features when there are too many features.

References

[1] Machine Learning - Stanford University (https://www.coursera.org/learn/machine-learning)