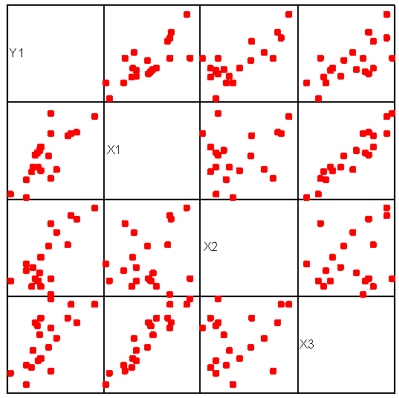
Multivariate Regression



1. Notation:

- m: number of training examples
- n: number of features
- x: input variable
- y: output variable
- $x^{(i)}$: input features of ith training example
- $x_j^{(i)}$: value of feature j in ith training example
- $y^{(i)}$: output of ith training example
- $(x^{(i)}, y^{(i)})$: ith training example

x_0	x_1: size	x_2: # of bedrooms	x_3: # of floors	x_4: age of house	y
1	2104	5	1	45	460
1	1416	3	2	40	232

- This example shows a training data with:
  - 2 training examples
  - 4 features (by convention, there will always be feature 0 & it will always be 1)
  - $x_3^{(2)}$: feature 3 of the second training example, which is 2

2. Hypothesis:

- Multivariate linear regression equation:

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

  - Note: we can write $h_\theta(x)$ as following:

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

Multivariate Regression

$$h_\theta(x) = (\theta)^T x = (\theta_0 \quad \theta_1 \quad \dots \quad \theta_n)\begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$$

- Polynomial regression equation (some examples):

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1^1 + \theta_2 x_2^2 + \cdots + \theta_n x_n^n$$

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1^{\frac{1}{2}} + \theta_2 x_2^{\frac{1}{4}} + \cdots + \theta_n x_n^{\frac{1}{2n}}$$

3. **<u>Cost Functions:</u>**
   - Squared error (SE) cost function:

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

$$= \frac{1}{2m}\sum_{i=1}^{m}\left((\theta)^T x^{(i)} - y^{(i)}\right)^2$$

$$= \frac{1}{2m}\sum_{i=1}^{m}\left(\left(\sum_{j=0}^{n}\theta_j x_j^{(i)}\right) - y^{(i)}\right)^2$$

4. **<u>Gradient Descent Algorithm Using SE Loss Function for Multivariate Regression:</u>**

Repeat until convergence {

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)\left(x_j^{(i)}\right) \qquad (for\ j = 0, \dots, n)$$

}

- Note:

$$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)\left(x_j^{(i)}\right)$$

5. **<u>Feature Scaling</u>**
   - Make sure that the features are on similar scale (i.e. Data normalization)
   - Feature scaling helps making the contour plot of the cost function less skewed. Thus, helping the gradient descent move through a much direct path
   - Here are some formulas used for features scaling:

Multivariate Regression

$$x_i = \frac{x_i - \bar{x}}{\max(x)}$$

$$x_i = \frac{x_i - \bar{x}}{\max(x) - \min(x)}$$

$$x_i = \frac{x_i - \bar{x}}{s}$$

- Note: we don't apply feature scaling to feature 0 (i.e. x_0) because by convention, it will always be 1

6. Normal Equation

- A method to solve for θ analytically
- We can apply the following formula to solve for θ:

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

- Example:

x_0	x_1: size	x_2: # of bedrooms	x_3: # of floors	x_4: age of house	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

Construct a matrix X and Y from the data:

$$X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix} \qquad Y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

From here, we can solve for θ

- Note:
  - Using normal equation method does not require us to do feature scaling.
  - There are some pros and cons of normal equation method

Gradient descent	Normal equation
+ Need to choose α + Need many iterations + Works well even when n is large	+ No need to choose α + Don't need to iterate + Slow when n is large because it needs to compute the inverse of $X^T X$ ($\sim O(n^3)$) + $X^T X$ may not be invertible

Multivariate Regression

- If the matrix is not invertible, we can try:
  - Pseudoinverse the matrix (?)
  - Remove redundant features (linearly dependent features)
  - Have fewer features (Could affect the correctness)