ADVANCED REGRESSION TOPICS

NORMAL EQUATION

- Training Data Set can be written as following:

Sl. No.	$\boldsymbol{x_0}$	$\boldsymbol{x_1}$	$\cdots$	$\boldsymbol{x_k}$	**y**
1	1	$x_1^{(1)}$	$\cdots$	$x_k^{(1)}$	$y^{(1)}$
2	1	$x_1^{(2)}$	$\cdots$	$x_k^{(2)}$	$y^{(2)}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
m	1	$x_1^{(m)}$	$\cdots$	$x_k^{(m)}$	$y^{(m)}$

- There are total k many features and m many training examples.
- Notice that we have added one extra feature column $\boldsymbol{x_0}$ with all values 1 in the left.
- The training samples can now be written as $\left\langle \boldsymbol{x}^{(i)}, y^{(i)} \right\rangle_{i=1}^{m}$

  Where $\boldsymbol{x}^{(i)} = \left[x_0^{(i)}, x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)}\right]^T$ is the vector of dimension $k + 1$.

- Now the equation $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ \ldots + \theta_k x_k$ can be written in vector form as $y = \boldsymbol{\Theta}^T \boldsymbol{x}$.

- Where $\boldsymbol{\Theta} = [\theta_0\ , \theta_1, \theta_2, \ldots, \theta_k]^T$ is the vector of parameters of the model. $\boldsymbol{\Theta}$ is of dimension $k + 1$.

NORMAL EQUATION

- In linear regression we are trying to estimate the model parameter vector from the given set of data. Let the estimated parameter vector be $\widehat{\boldsymbol{\Theta}}$ and the corresponding predicted values be $\widehat{\boldsymbol{y}}$. Then in vector-matrix notation:

$$\widehat{\boldsymbol{y}} = \begin{bmatrix}\hat{y}^{(1)}\\ \hat{y}^{(2)}\\ \vdots\\ \hat{y}^{(m)}\end{bmatrix} = \begin{bmatrix}1 & x_1^{(1)} & \cdots & x_k^{(1)}\\ 1 & x_1^{(2)} & \cdots & x_k^{(2)}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & x_1^{(m)} & \cdots & x_k^{(m)}\end{bmatrix}\begin{bmatrix}\hat{\theta}_0\\ \hat{\theta}_1\\ \vdots\\ \hat{\theta}_k\end{bmatrix} \quad or \;\; \widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\Theta}}\,,$$

Where $\boldsymbol{X}$ is the matrix:

$\boldsymbol{x_0}$	$\boldsymbol{x_1}$	$\cdots$	$\boldsymbol{x_k}$
1	$x_1^{(1)}$	$\cdots$	$x_k^{(1)}$
1	$x_1^{(2)}$	$\cdots$	$x_k^{(2)}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$
1	$x_1^{(m)}$	$\cdots$	$x_k^{(m)}$

- The mean square error cost function in vector-matrix notation is following:

$$J(\widehat{\boldsymbol{\Theta}}) = \frac{1}{2m}\left(\boldsymbol{X}\widehat{\boldsymbol{\Theta}} - \boldsymbol{y}\right)^T\left(\boldsymbol{X}\widehat{\boldsymbol{\Theta}} - \boldsymbol{y}\right)$$

where $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\Theta}}$ is the vector of predicted values and y is vector of the actual values.

NORMAL EQUATION

- Simplified cost function is:

$$J(\hat{\boldsymbol{\Theta}}) = \frac{1}{2m}(\boldsymbol{X}\hat{\boldsymbol{\Theta}} - \boldsymbol{y})^T(\boldsymbol{X}\hat{\boldsymbol{\Theta}} - \boldsymbol{y}) = \frac{1}{2m}\left((\boldsymbol{X}\hat{\boldsymbol{\Theta}})^{\boldsymbol{T}} - \boldsymbol{y}^{\boldsymbol{T}}\right)(\boldsymbol{X}\hat{\boldsymbol{\Theta}} - \boldsymbol{y}) =$$

$$\frac{1}{2m}\left((\boldsymbol{X}\hat{\boldsymbol{\Theta}})^{\boldsymbol{T}}\boldsymbol{X}\hat{\boldsymbol{\Theta}} - \boldsymbol{y}^{\boldsymbol{T}}(\boldsymbol{X}\hat{\boldsymbol{\Theta}}) - (\boldsymbol{X}\hat{\boldsymbol{\Theta}})^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{y}\right) = \frac{1}{2m}\left(\hat{\boldsymbol{\Theta}}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\hat{\boldsymbol{\Theta}} - \boldsymbol{2}\hat{\boldsymbol{\Theta}}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{y}\right)$$

Because $\boldsymbol{y}^{\boldsymbol{T}}(\boldsymbol{X}\hat{\boldsymbol{\Theta}})$ and $(\boldsymbol{X}\hat{\boldsymbol{\Theta}})^{\boldsymbol{T}}\boldsymbol{y}$ are scalars and $\boldsymbol{y}^{\boldsymbol{T}}(\boldsymbol{X}\hat{\boldsymbol{\Theta}}) = (\boldsymbol{X}\hat{\boldsymbol{\Theta}})^{\boldsymbol{T}}\boldsymbol{y} = \hat{\boldsymbol{\Theta}}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$

- After differentiating $J(\hat{\boldsymbol{\Theta}})$ with respect to $\hat{\boldsymbol{\Theta}}$ and setting the derivative to zero:

$$\frac{\partial J(\hat{\boldsymbol{\Theta}})}{\partial \hat{\boldsymbol{\Theta}}} = \frac{1}{m}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\hat{\boldsymbol{\Theta}} - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}) = 0 \quad or \quad \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\hat{\boldsymbol{\Theta}} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \quad or$$

$$\boxed{\hat{\boldsymbol{\Theta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}}\,, assuming\ \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\ is\ invertible$$

Normal Equation

NORMAL EQUATION

- Gradient Descent or Normal Equation which one is preferable?

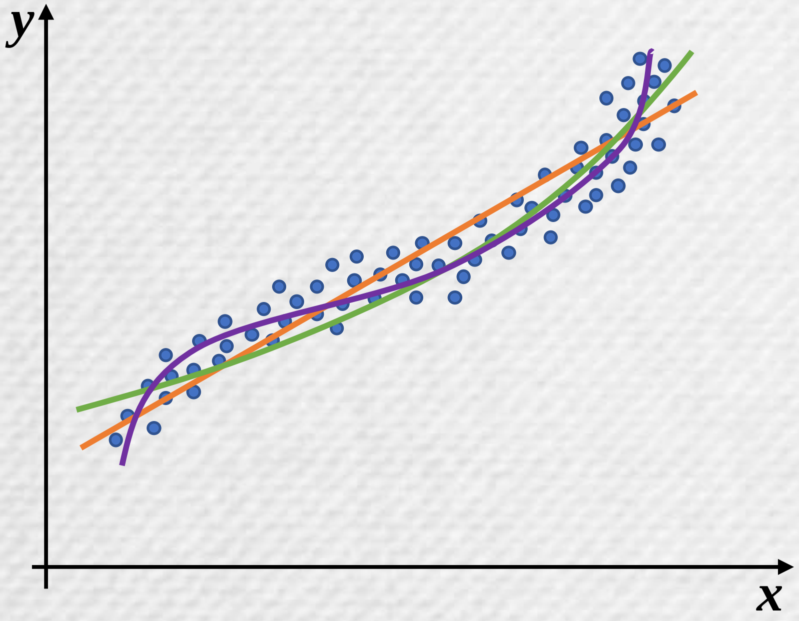
Though normal equation directly gives solution without iteration like GD, it has many drawbacks. Like, for large dataset computing $\left(\boldsymbol{X^T X}\right)^{-1}$ is a costly operation. Moreover, if $\boldsymbol{X^T X}$ is non-invertible we can’t use normal equation directly as above.

The workaround in the case when $\boldsymbol{X^T X}$ is non-invertible is to use pseudo-inverse.

Hence, gradient descent is more popular and good choice for solving linear regression problem.

POLYNOMIAL REGRESSION

- Consider the following example:



- We can fit a straight line through the datapoints of the form

$$y = \theta_0 + \theta_1 x$$

- But we can do better, if we fit second order polynomial of the form:

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

- Or if we fit third order polynomial of the form:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

- In general we can fit a n^{th} order polynomial through the datapoints:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ \dots + \theta_n x^n$$

POLYNOMIAL REGRESSION

- Smaller the value of $\boldsymbol{n}$, the complexity of the model is less but the model may not fit the dataset appropriately. So we have to choose $\boldsymbol{n}$ accordingly such that we get reasonably good fit with less complexity.

- We can convert the polynomial regression problem into multiple linear regression problem just by assigning:
  $\boldsymbol{x_1 = x, x_2 = x^2, x_3 = x^3, \dots, x_n = x^n}$ and then constructing multiple linear regression model $\boldsymbol{y = \theta_0 + \sum_{i=1}^{n} \theta_i x_i}$

- For more than one predictor variables the polynomial regression becomes more complicated. For two predictor variables $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ the generalized form of second order polynomial is:
  $\boldsymbol{y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2}$

COEFFICIENT OF DETERMINATION

To determine the “goodness” of the fit in a linear regression model we use a quantitative measure. That is *“Coefficient of Determination”* (R^2). It is defined as follows.

- Let there are $\boldsymbol{m}$ number of datapoints. $\boldsymbol{y} = [y_1, y_2, y_3, \dots, y_m]^T$ is the vector of the actual values of target variable and $\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, \hat{y}_3, \dots, \hat{y}_m]^T$ is the vector of predicted values of the target variable.

- Let, $\bar{y}$ is the mean of the target variable. Then the *Total Sum of Squares (TSS)* is defined as follows:

$$TSS = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

- TSS is proportional to the variance of the target variable.

COEFFICIENT OF DETERMINATION

- We already know the *Residual Sum of Squares (RSS)*

$$RSS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

- Now the *Fraction of Unexplained Variance (FUV)* is defined as:

$$FUV = \frac{RSS}{TSS}$$

- Coefficient of Determination (R^2) also called *Fraction of Explained Variance (FEV)* is defined as:

$$R^2 = 1 - FUV = 1 - \frac{RSS}{TSS}$$

COEFFICIENT OF DETERMINATION

Properties of Coefficient of Determination:

- Coefficient of Determination (R^2) lies between 0 to 1
- Closer the value of R^2 to 1, Regression model fits better to our dataset and can better explain the observed variability of the target variable.
- Smaller value of R^2 implies that the regression model is not that good.
- It can be shown that for bivariate dataset

$$R^2 = Square\ of\ the\ correlation\ coefficient\ between\ the\ predictor\ and\ target\ variable$$