Advanced Regression Topics

Sourav Karmakar

souravkarmakar29@gmail.com

Outline

- Normal Equation of Linear Regression
- Polynomial Regression
- Coefficient of Determination (R-squared)
- Adjusted R-squared

Normal Equation

- Training Data Set can be written as following:

Sl. No.	x_0	x_1	⋯	x_k	y
1	1	$x_1^{(1)}$	⋯	$x_k^{(1)}$	$y^{(1)}$
2	1	$x_1^{(2)}$	⋯	$x_k^{(2)}$	$y^{(2)}$
⋮	⋮	⋮	⋯	⋮	⋮
m	1	$x_1^{(m)}$	⋯	$x_k^{(m)}$	$y^{(m)}$

- 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 $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle_{i=1}^{m}$

  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 $k+1$.

$\vec{v}_1 = [1, 2, 3, 4]^T$ — $\vec{v}_1$ is of dimension 4 (4×1)

$\vec{v}_2 = [1.5, 2.5, 1, 3]^T$

$\vec{v}_1^T \vec{v}_2 = [1, 2, 3, 4] \begin{bmatrix} 1.5 \\ 2.5 \\ 1 \\ 3 \end{bmatrix} = 1 \times 1.5 + 2 \times 2.5 + 3 \times 1 + 4 \times 3$

$= 1.5 + 5 + 3 + 12 = 21.5$

Dot / inner product.

$\vec{v}_1^T \vec{v}_1 = [1, 2, 3, 4] \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = 1^2 + 2^2 + 3^2 + 4^2$

$\vec{v}_1^T \vec{v}_1 = \|\vec{v}_1\|^2$

$\|\vec{v}_1\| = \sqrt{1^2 + 2^2 + 3^2 + 4^2}$

magnitude of $\vec{v}_1$

$\vec{v}^T \vec{v} = \|\vec{v}\|^2$ $\qquad \underset{m\times 1}{\hat{y}} = X\hat{\theta}$ $\qquad X_{m\times(k+1)} \cdot \hat{\theta}_{(k+1)\times 1}$

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

$$= \frac{1}{2m}\left[\left(\hat{y}^{(1)} - y^{(1)}\right)^2 + \left(\hat{y}^{(2)} - y^{(2)}\right)^2 + \cdots + \left(\hat{y}^{(m)} - y^{(m)}\right)^2\right]$$

$$= \frac{1}{2m}\|\hat{y} - y\|^2$$

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

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

$\hat{y}$	y	$(\hat{y} - y)$
$\hat{y}^{(1)}$	$y^{(1)}$	$\hat{y}^{(1)} - y^{(1)}$
$\hat{y}^{(2)}$	$y^{(2)}$	$\hat{y}^{(2)} - y^{(2)}$
$\hat{y}^{(3)}$	$y^{(3)}$	$\hat{y}^{(3)} - y^{(3)}$
$\vdots$	$\vdots$	$\vdots$
$\hat{y}^{(m)}$	$y^{(m)}$	$\hat{y}^{(m)} - y^{(m)}$

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 $\hat{\boldsymbol{y}}$. Then in vector-matrix notation:

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

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

x_0	x_1	$\cdots$	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 $\hat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\Theta}}$ is the vector of predicted values and y is vector of the actual values.

Normal Equation

- Simplified cost function is:

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

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

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

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

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

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

Normal Equation

Normal Equation

- Gradient Descent or Normal Equation which one is preferable and Why?

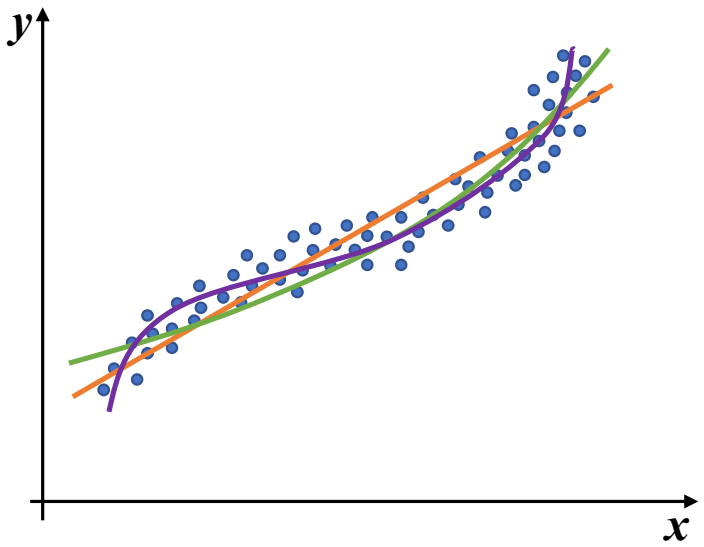
Though normal equation directly gives solution without iteration like GD, it has many drawbacks. Like, for large dataset computing $\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}$ is a costly operation. Moreover, if $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is non-invertible we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

The workaround in the case when $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ 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:

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

Polynomial Regression

- Smaller the value of n, the complexity of the model is less but the model may not fit the dataset appropriately. So we have to choose 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:
  $x_1 = x, x_2 = x^2, x_3 = x^3, \dots, x_n = x^n$ and then constructing multiple linear regression model $y = \theta_0 + \sum_{i=1}^{n} \theta_i x_i$

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

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

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

* **Note:** Negative values of coefficient of determination is not mathematically impossible but in that case the linear regression performs really poor and in that case we can't interpret coefficient of determination as square of correlation coefficient.

Adjusted R-Squared

Problem with R-squared:

Even if we are adding redundant variables to the data, the value of R-squared does not decrease. It either remains the same or increases with the addition of new independent variables, though some of the independent variables might not be useful in determining the target variable.

- The **adjusted R-squared** is a modified version of **R-squared** that has been **adjusted** for the number of predictors in the model.

$$R^2_{adj} = 1 - \frac{(m-1)}{(m-k-1)}(1-R^2)$$

m : Number of data points used to fit the model
k : Number of features/predictors used in the model
R^2 : Coefficient of determination of the model

Thank You