Page :
Date :

⇒ Regression :

→ Overdetermined and Underdetermined systems :

Let $Ax = b$, where $A_{m \times n}$ is coefficient matrix, $x \in R^n$ be unknown vector and $b \in R^m$ is the right hand vector

① If $m >> n$, then the system is called overdetermined system (no. of observations is more than no of unknown variables).

② If $m << n$, then the system is called underdetermined system (no. of observations is less than no. of unknown variables). It will have infinitely many solution.

→ Least square Approximation: Given $Ax = b$, where $A \in R^{m \times n}$ and $m >> n$. Here, in least square approximation, we solve

$$\arg\min_x \|Ax - b\|_2^2.$$

↓

Minimize the Euclidean distance between Ax and b.

Ex :

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \\ a_{31}x_1 + a_{32}x_2 &= b_3 \end{aligned}\right| \rightarrow \text{Goal} \rightarrow \|Ax - b\|_2^2 \} \text{minimize.}$$

$$= (a_{11}x_1 + a_{12}x_2 - b_1)^2 + (a_{21}x_1 + a_{22}x_2 - b_2)^2 + (a_{31}x_1 + a_{32}x_2 - b_3)^2$$

$$= E$$

$$\Rightarrow \frac{\partial E}{\partial x_1} = 0, \quad \frac{\partial E}{\partial x_2} = 0$$

Page :
Date :

Method $Ax = b$

$A^TAx = A^Tb$

$x = (A^TA)^{-1}A^Tb$

$\boxed{x = A^+b}$ ⟶ least square solution

Here, $A^+ = (A^TA)^{-1}A^T$ is called Pseudoinverse (right) of A and

$X = A^+b$ is the least square solution $AX = b$

→ Minimum Normed Solution (MNS): Given $AX = b$, where $A \in R^{m \times n}$ and $m << n$. In this case, we are having $n-m$ free variables. Assigning any arbitrary values to these variables lead to a solution of $AX = b$. Therefore, we will have infinitely many solutions of the system $AX = b$. A minimum normed solution is that which minimizes the $\|X\|$ among these infinite solutions.

$$\min_X \|X\|, \text{ subject to } AX = b \quad — (I)$$

$$\Rightarrow \min_X \|AX - b\|_2^2 + \|X\|_2^2 \quad — (II)$$

Here we can't go with right pseudoinverse because A^TA is $n \times n$ matrix and $\text{rank}(A^TA) \leq m$, means it is rank deficient, so $(A^TA)^{-1}$ does not exist but AA^T is a $m \times m$ matrix and $\text{rank}(AA^T) = m$, we will have full rank matrix. Hence the inverse of AA^T will exist.

So we will have

$$A^{\dagger} = A^{T}(AA^{T})^{-1}$$ as left pseudo-inverse

so, $$X = A^{\dagger}b \quad \text{— (iii)}$$

↳ left pseudoinverse.

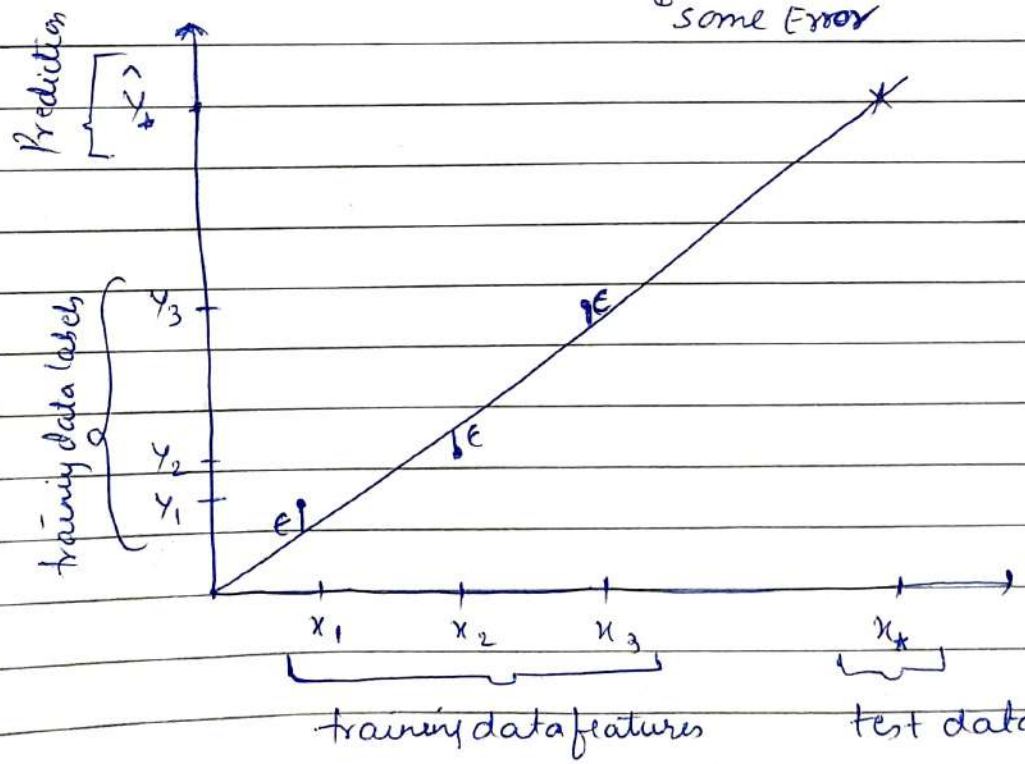
solving (iii) will give you MNS.

→ Regression refers to the problem of learning the ~~bar~~ ~~relationship~~ relationships between some (qualitative or quantitative) input variables

$$X = [x_1, x_2, \cdots, x_n]^{T}$$ and a quantitative output variable y. Means, regression is about learning model f such that

$$y = f(X) + \epsilon$$

↓ some Error



→ Relationship can be linear or non-linear

Date :

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p + \epsilon$$

↓ Noise term

y — output variable (a scalar)

a_i → parameters

x_i → input variables (each a scalar)

Our dataset looks like:

x_1	x_2	x_3	x_4	y
x_1^1	x_2^1	x_3^1	x_4^1	y_1
x_1^2	x_2^2	x_3^2	x_4^2	y_2
⋮	⋮	⋮	⋮	

$$y_1 = \alpha_0 + \alpha_1 x_1^1 + \alpha_2 x_2^1 + \alpha_3 x_3^1 + \alpha_4 x_4^1$$

⋮

$$y_n = \alpha_0 + \alpha_1 x_1^n + \alpha_2 x_2^n + \alpha_3 x_3^n + \alpha_4 x_4^n$$

In Matrix form, we can write

$$\begin{bmatrix} 1 & x_1^1 & x_2^1 & x_3^1 & x_4^1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^n & x_2^n & x_3^n & x_4^n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$A_{n \times 5}$ $\quad \alpha_{5 \times 1}$ $\quad y_{n \times 1}$

Here $n \gg 5$, so it is overdetermined matrix or system

So; $A\alpha = y$. we can use the method of least square to solve this

Page :
Date :

Here we have

$$A\alpha = Y$$

$$\underbrace{A^T A}_{5\times 5}\alpha = \underbrace{A^T Y}_{5\times 1}$$

$$\boxed{\alpha = (A^T A)^{-1} A^T Y}$$

Calculus approach:

$$E = (y_1 - (\alpha_0 + \alpha_1 x_1^1 + \alpha_2 x_2^2 + \alpha_3 x_3^2))^2 + \cdots + (y_n - (\alpha_0 + \alpha_1 x_1^n + \alpha_2 x_2^n + \alpha_3 x_3^n + \alpha_4 x_4^n))^2$$

$$\frac{\partial E}{\partial \alpha_0} = 0, \frac{\partial E}{\partial \alpha_1} = 0, \frac{\partial E}{\partial \alpha_2} = 0, \frac{\partial E}{\partial \alpha_3} = 0, \frac{\partial E}{\partial \alpha_4} = 0$$

9t is same as the previous approach.

* Fitting Evaluation: MSE & RMSE:

Mean square error $= \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2$, where y_i^* is the predicted value of y_i^*

RMSE (Root Mean Squared Error)

$$= \sqrt{MSE}$$

Goodness of fitting: The residual for the i^{th} observation is given by:

$$y_i - \hat{y}_i = y_i - (\hat{a}_0 + \hat{a}_1 x_{i1} + \cdots + \hat{a}_p x_{ip})$$

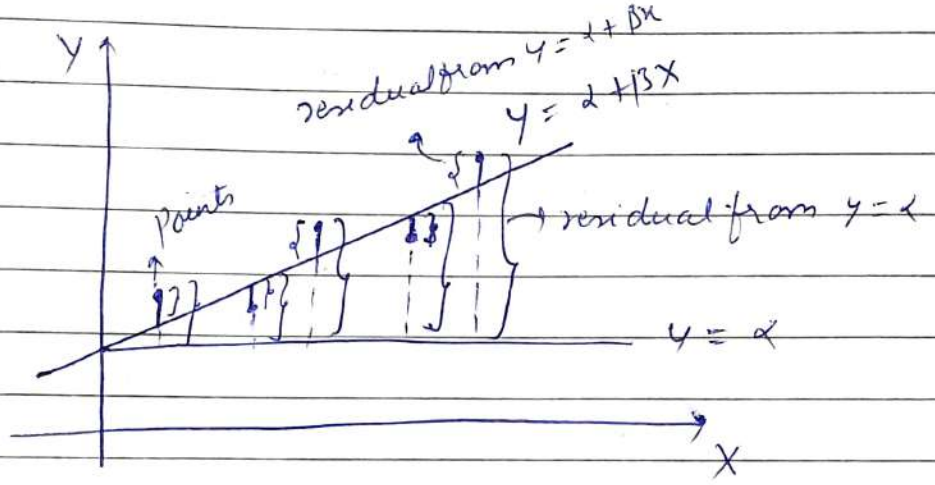
Page :
Date :

Consider the linear regression models as

$y = \alpha + \beta x$ and $y = \alpha$, then

R^2 is defined as

$$R^2 = 1 - \frac{\text{sum of square residuals from model with } \alpha \& \beta}{\text{sum of square residuals from model with } \alpha \text{ only}}$$



→ If Numerator = Denominator, then $R^2 = 0$ → model is not useful

→ If Numerator = 0 → model fits all the points perfectly and $R^2 = 1$

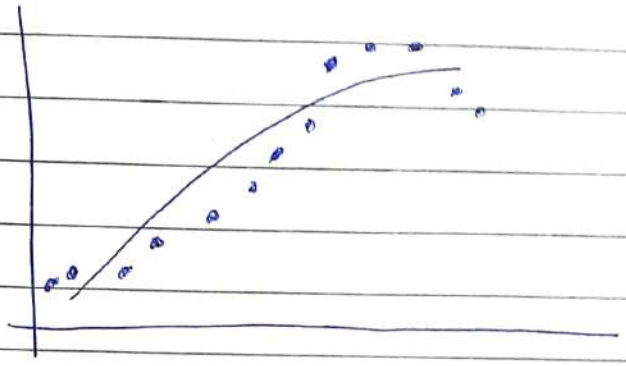
→ We see always have $0 \leq R^2 \leq 1$

Same is applicable for multiple variables

→ How large does R^2 need to be considered as "good"? This depends on the content, there is no absolute answer.

Page :
Date :

⇒ Polynomial Regression: $Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$



if you go for higher degree polynomial the line will fit the points perfectly. This may lead to overfitting.

⇒ Ridge Regression:

→ To combat overfitting we regularize our models. When we add a regularization term we call it ridge regression.

Here,

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

regularization term

$$= RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

Here RSS is our residual square

↓

In ridge regression we minimize this whole term.

$\lambda \geq 0$ → is a hyperparameter. or a tuning parameter.

This is also called l_2 regularization. The above is same as

$$\min_{x} \|Ax - b\|_2^2 + \lambda \|x\|_2^2$$

Page :
Date :

→ The effect of this equation is to add a shrinkage penalty of the form

$$\lambda \sum_{j=1}^{p} \beta_j^2$$

where the tuning parameter λ is a positive value

→ This has the effect of shrinking the estimated beta coefficient towards zero. It turns out that such a constraint should improve the fit, because shrinking the coefficient can significantly reduce their variance.

→ Selecting the good value for λ is critical.

⇒ Lasso Regression.

→ One significant problem of ridge regression is that the penalty term will never force any of the coefficient to be exactly zero.

→ Thus, the final model will include all p predictors, which creates a challange in model interpretation.

→ The lasso works in a similar way to ridge regression, except it uses a different penalty term that shrinks some of the coefficients exactly to zero.

→ The lasso coefficients minimizes the quantity :

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \qquad \text{where } \lambda > 0$$

This is called l_2 penalty. It forces some of the coefficient to be exactly zero when λ is sufficiently large. This means lasso performs variable / feature selection.