10.04.24 | Lecture 13

Let us have a system of linear equation

$$AX = b$$

$A_{m \times n} \rightarrow$ coeff. mat

$X \in \mathbb{R}^n \rightarrow$ unknown vector

$b \in \mathbb{R}^m \rightarrow$ right hand vector.

① If $m >> n$ [No. of rows >> no. of colum]

system is overdetermined system.

no. of observations is more than no. of unknown variables

② if $m << n$ [No. of colum >> no. of rows]

System is underdetermined

No. of observations are less than the no. of variable

Overdetermined soln. $m >> n$

$$\begin{bmatrix} 1 & 7 & 5 \\ 3 & 9 & 4 \\ 2 & 7 & 2 \\ 8 & 4 & 6 \\ 9 & 2 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \\ 9 \\ 10 \\ 15 \end{bmatrix}$$

$5 \times 3 \quad 3 \times 1 \quad 5 \times 1$

$5 =$ ns + 3 unknown.

exact soln. rarely.

approx. soln. → least square

Undetermined system

$$\begin{bmatrix} 7 & 8 & 2 & 6 & 9 \\ 5 & 4 & 2 & 9 & 7 \\ 3 & 8 & 2 & 7 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 3 \\ 8 \\ 2 \end{bmatrix}$$

3×5 $\quad$ 5×1 $\quad$ 3×1

∞ no. of solns.

One can arbitrary fix 2 & find other solns.

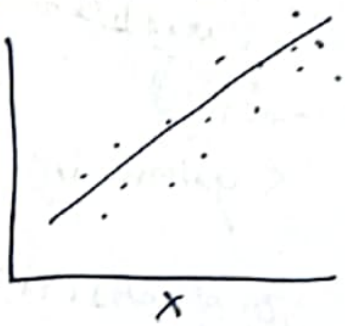
Overdetermin

Simple example you all did in first year B.S. Lab.

Line fitting

↳ Many points

→ Many ways to draw straight lines to fit all the data points.



→ least square - find line that fit best

↳ Linear regression

Best fitting line is called least squares line or the regression line.

→ Minimal residual error

$Ax = b$ where $A \in \mathbb{R}^{m \times n}$ and $m \gg n$

$\underset{x}{\text{argmin}} \|Ax - b\|_2^2$ Euclidean norm minimize

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \\ a_{31}x_1 + a_{32}x_2 &= b_3 \end{aligned}\right\} \Rightarrow \begin{aligned} &(a_{11}x_1 + a_{12}x_2 - b_1)^2 + \\ &(a_{21}x_1 + a_{22}x_2 - b_2)^2 + \\ &(a_{31}x_1 + a_{32}x_2 - b_3)^2 = E \end{aligned}$$

$\frac{\partial E}{\partial x_1} = 0 \; ; \frac{\partial E}{\partial x_2} = 0$

2 linear eqn solve & get the value

$$Ax = b \qquad A_{m \times n} \rightarrow A^T_{n \times n}$$

$$A^T A x = A^T b$$

$$x = (A^T A)^{-1} A^T b$$

$\hookrightarrow A^+ b$ pseudo inverse

$A^+ = (A^T A)^{-1} A^T$ is called pseudo inverse (left) of A

$x = A^+ b$ is the least square soln. of $Ax = b$ overdetermined system

Example

$$\begin{pmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \\ 0 \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} \quad A^T A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 4 & 10 \end{bmatrix}$$

$(A^T A)^{-1} A^T b$ (left pseudo inverse)

$$\begin{bmatrix} 0.7143 & -0.2857 \\ -0.2857 & 0.2143 \end{bmatrix} \begin{bmatrix} 5 \\ 0 \end{bmatrix} = \begin{bmatrix} 3.57 \\ -1.43 \end{bmatrix}$$

Minimum Normed Solution (MNS)

Consider the linear system $AX = b$

where A is $m \times n$ s.t. $m << n$.

"$n-m$" free variables.

Assigning any arbitrary values to these variables lead to a solution of $AX = b$

∴ one can have infinitely many solutions of the system $AX = b$.

→ Minimum Normed Solution is that which minimize the $\|X\|$ among these infinite solutions.

The minimization problem can be solved as.

$$X^* = A^T (AA^T)^{-1} b$$

Here, $A^+ = A^T (AA^T)^{-1}$ is ~~pseudo~~ the right pseudo-inverse of A.

$AA^T \rightarrow m \times m$ full rank

$$\boxed{X = A^+ b}$$

:a

$$x_1 - 2x_2 + 5x_3 = 9$$
$$-x_1 + x_2 + 2x_3 = 1$$

$$A = \begin{bmatrix} 1 & -2 & 5 \\ -1 & 1 & 2 \end{bmatrix}$$

$$AA^T = \begin{bmatrix} 30 & 7 \\ 7 & 6 \end{bmatrix}$$

right pseudo inverse

$$A^+ = A^T (AA^T)^{-1} b$$

$$= \begin{bmatrix} 1 & -1 \\ -2 & 1 \\ 5 & 2 \end{bmatrix} \begin{bmatrix} 0.3588 \\ -0.2519 \end{bmatrix} = \begin{bmatrix} 0.6107 \\ -0.9695 \\ 1.2901 \end{bmatrix}$$

Minimum Normed Soln.

Regression

Regression refers to the problem of learning the relationships b/w input variables!

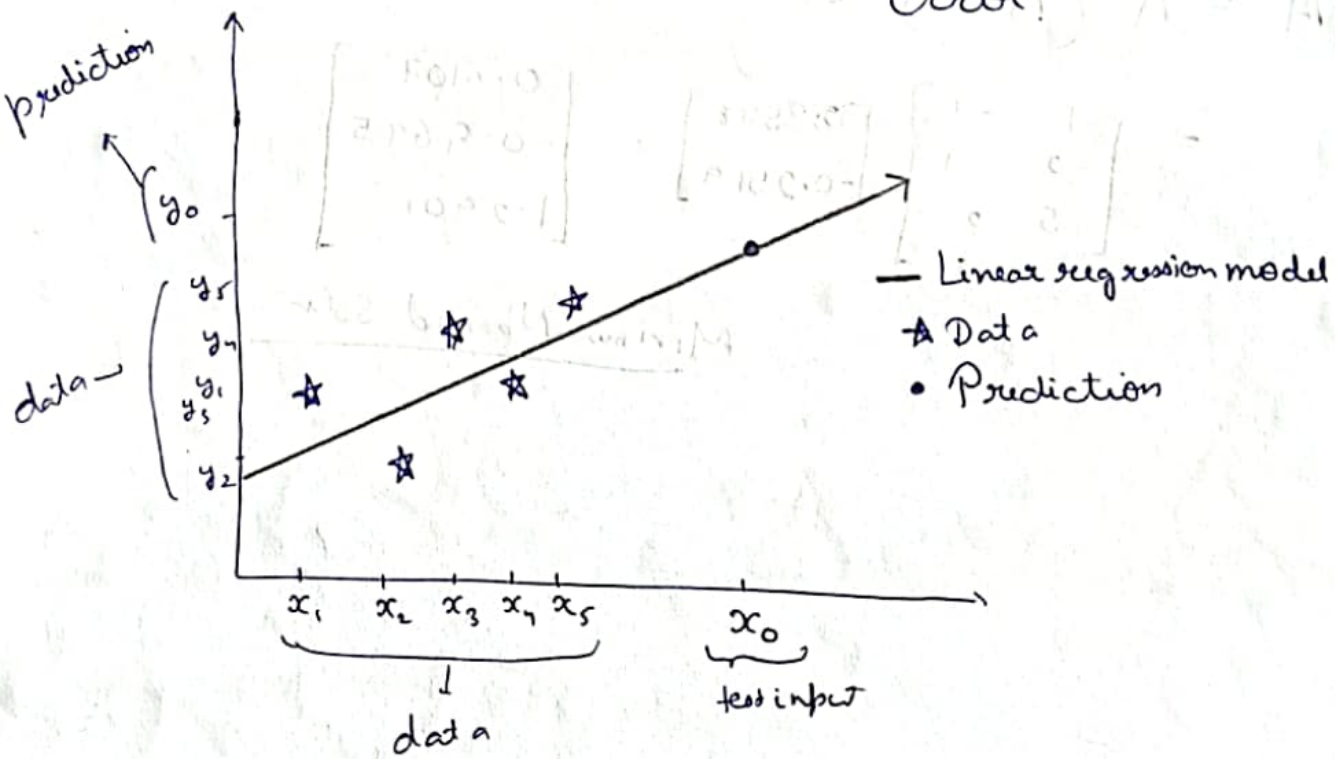
$$X = [x_1, x_2, x_3, \ldots, x_n]^T$$

and quantitative output variable y.

In simple words; regression is about learning a model f such that

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

ϵ → error



$$y = a_0 + a_1 x$$

Simple linear regression - Linear fit

1

Linear / Multiple Linear Regression

It is the simplest regression technique.

The linear regression model describe output variable y (a scalar) as a combination of the input variable $x_1, x_2, \ldots, x_d$ (each scalar) plus a noise term ϵ; i.e.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_d x_d + \epsilon$$

where a_i are parameters of the regression model which we needs to estimate based on the given data

X_1	X_2	X_3	Y	
x_1^1	x_2^1	x_3^1	y_1	$y_1 = a_0 + a_1 x_1^1 + a_2 x_2^1 + a_3 x_3^1$
x_1^2	x_2^2	x_3^2	y_2	$y_2 = a_0 + a_1 x_1^2 + a_2 x_2^2 + a_3 x_3^2$
x_1^3	x_2^3	x_3^3	y_3	$y_3 = a_0 + a_1 x_1^3 + a_2 x_2^3 + a_3 x_3^3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_1^d	x_2^d	x_3^d	y_d	$y_d = a_0 + a_1 x_1^d + a_2 x_2^d + a_3 x_3^d$

N no. of = n & 4 unknown over determined system

$$\begin{bmatrix} 1 & x_1^1 & x_2^1 & x_3^1 \\ 1 & x_1^2 & x_2^2 & x_3^2 \\ 1 & x_1^3 & x_2^3 & x_3^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^d & x_2^d & x_3^d \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

A

$d \times 4$

$A\,a = y$

As $m >> n$

def: pseudo inverse

$$a = (A^T A)^{-1} A^T y$$

Fitting Evaluation:

				Predicted
x_1	x_2	x_3	y	y^*
x_1	x_2	x_3	y_1	y_1^*
x_1'	x_2'	x_3'	y_2	y_2^*
⋮	⋮	⋮	⋮	⋮
x_1^n	x_2^n	x_3^n	y_n	y_n^*

Residual error defined

MSE = Mean Square Error

$$= \frac{1}{n} \sum_{i=1}^{n} (y_i - y_i^*)^2$$

RMSE = Root Mean Square Error

$$= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - y_i^*)^2}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \{ |y_i - y_i^*| \}$$

Mean Absolute Error.

2 | Goodness of fitting R^2

Standard way to ~~on~~ see how accurately one can predict value. R^2 is used.

- Total variation $TV = \sum_i |y_i - \bar{y}|^2$ (y_i → actual values, $\bar{y}$ → mean of actual values)

  where $\bar{y}$ is average of all the y values in one data. → some places called total Sum of Squares (SST or SS_{tot})

- Residual variation.

$$RV = \sum_i |y_i - f(x_i)|^2$$

(y_i → actual value, $f(x_i)$ → predicted value from regression model)

→ some places called Residual Sum of Squares (SSR or SS_{res})

$$R^2 = 1 - \frac{RV}{TV}$$

→ Normally $0 \leq R^2 \leq 1$

→ Technically R^2 can be -ve if ones model is extremely bad. Worse than horizontal line (mean) at predicting data.

→ $R^2 = 0$; if one just defined the fitted function to return the average of y as a constant value.

$$f(x) = \bar{y}$$

→ $R^2 = 1$ → model fits all the points perfectly

Some one can ask: what values of R^2 is considered "good"? The answer is not straightforward.

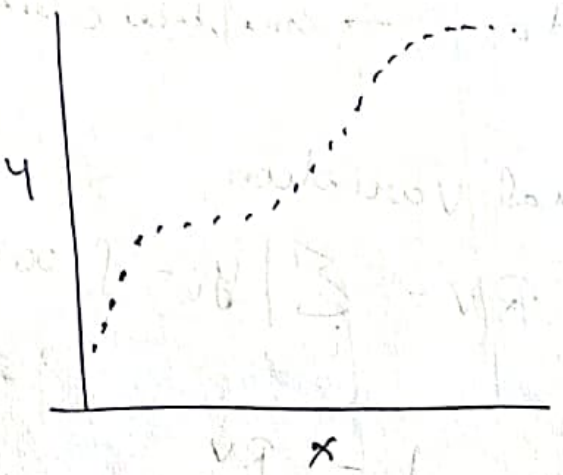
→ It depends on the situation. For hard to

predict Y variables, smaller value may be "good".

Polynomial Regression

Sometimes linear ~~regre~~ relation are not sufficient to capture the true pattern going on in the data with a single dependent variable x.

Exaff



One then has to build a model of form

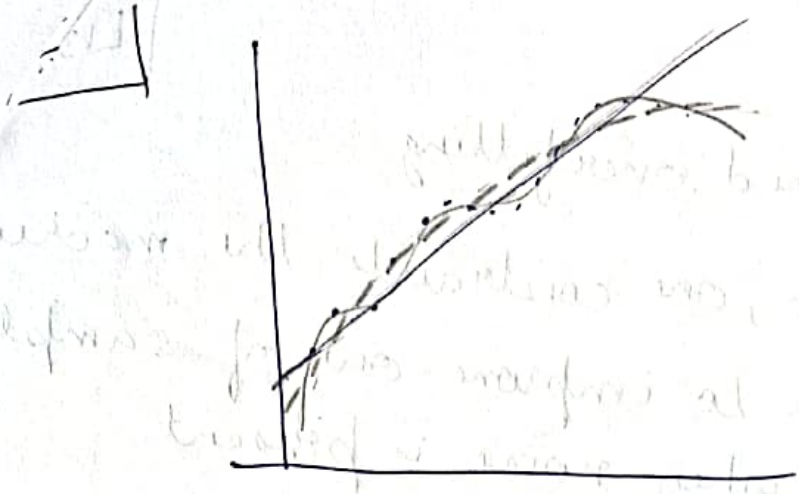
$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$$

or more generally for some polynomial of degree p.

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_p x^p$$

$$= \alpha_0 + \sum_{i=1}^{p} \alpha_i x^i$$

With higher order polynomial one can fit the data well but there is a risk of over fitting.



$$y = \underbrace{\alpha_0 + \alpha_1 x + \alpha_2 x^2}_{\text{sufficient}} + \underbrace{\alpha_3 x^3 + \alpha_4 x^4 + \alpha_5 x^5 + \alpha_6 x^6}_{\text{overfitting}}$$

$R^2 = 1$ but overfitting

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$X_p \qquad \alpha \qquad y$

$$\alpha = \underbrace{(X_p^T X_p)^{-1} X_p^T}\, y$$

left pseudo inverse

~~right pseudo inverse~~

Adjusted R-squared

A high value of R-square is always better for the model. Hence adding more independent variables to the regression model increases R-square value. This will inflate R-square & lead to overfitting.

One can avoid this by using adjusted R-square value.

It penalize the model for including independent variables.

Adjusted R-squared takes into account the number of predictors and only increase if the added variables improve the model more than would by expected by chance.

$$R^2_{adj} = 1 - \left(\frac{(1-R^2)(n-1)}{n-p-1} \right)$$

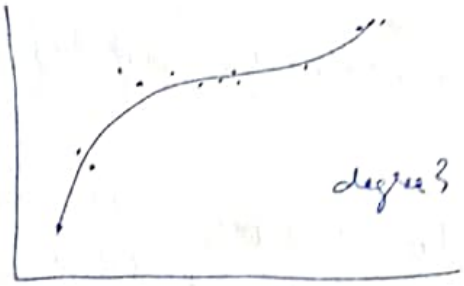
R^2 → Rsquared.

n → no. of observation (sample size)

p → no. of independent variables (predictor) in the model

Variance poor variance of complicated model, High variance significa. the model pass through most of data point, thereby leading to overfitting the data.

Underfit High bias
low Variance



Correct fit
low bias
low Variance

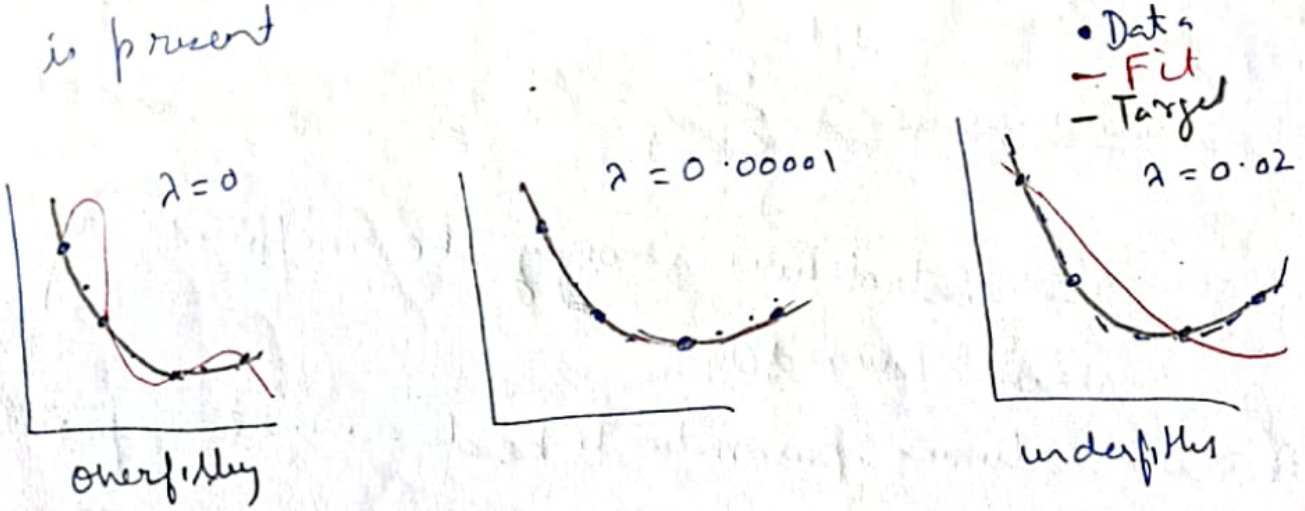
Overfit
low bias
High Variance

Regularized regression

In linear regression algorithm, coefficient of each independent variable are selected in such a way that the loss function is minimized. Too large coefficient results in overfitting the model on training dataset. Such a model will not generalize well on the unseen data.

∴ to overcome this limitation, we perform regularization that penalizes the large coefficients

→ One constraints the machine learning algorithm to improve out-of-sample error, especially when noise is present



λ → regularization parameter

out-of-sample → how well model generalizes the unseen data

in-sample error → how well it fit the training data.

Ridge regression.

L2 regularization

Extension of linear regression

In the loss function, a penalty parameter is added. The penalty is equivalent to the square of the magnitude of coefficients.

In linear regression, procedure estimates the alpha coefficients using the values that minimizes

$$RSS = \sum_{i=1}^{n} \left(y_i - \alpha_0 - \sum_{j=1}^{p} \alpha_j x_i^j \right)^2$$

Residual sum of squares.

In ridge regression, coefficients are estimated by minimizing a slightly different quantity.

$$\sum_{i=1}^{n} \left(y_i - \alpha_0 - \sum_{j=1}^{p} \alpha_j x_i^j \right)^2 + \lambda \sum_{j=1}^{p} \alpha_j^2$$

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

→ regularization term

λ → control how strongly the coeffiant are shrunk toward 0.

$\lambda \geq 0$ is a tuning parameter to be determined separately

Matrix form

$$J(\alpha) = (y - X\alpha)^T (y - X\alpha) + \lambda \alpha^T \alpha$$

y = vector of true target values

$X \rightarrow$ matrix of polynomial features

$\alpha \rightarrow$ vector of coefficient (to be learned)

$\lambda \alpha^T \alpha \rightarrow$ L2 penalty.

Soln. of Ridge Regression.

$$\frac{\partial J}{\partial \alpha}$$

$$\rightarrow \alpha = (X^T X + \lambda I)^{-1} X^T y$$

$$= (A^T A + \lambda I)^{-1} A^T y$$

This has the effect of shrinking the estimated α coefficient towards zero

It turns out that such a constraint should improve the fit ∵ shrinking the coefficient can significantly reduce their variance

Selecting a good value of λ is critical.

LASSO (Least Absolute Shrinkage and Selection Operator

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 predictor which creates challenge in model-interpretation.

An effective variant of this regularization is known as Lasso.

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, helping with feature selection

$$\sum_{i=1}^{n} \left(y_i - \alpha_0 - \sum_{j=1}^{p} \alpha_j x_i^j \right)^2 + \lambda \sum_{j=1}^{p} |\alpha_j|$$

$$RSS + \lambda \sum_{j=1}^{p} |\alpha_j|$$

Lasso uses a ~~L1 regu~~ L_1 penalty instead of L_2,

called L_1 regularization

λ is regularization parameter that control the amount of regularization.

Higher λ means stronger regularization, leading to more coefficients becoming zero.

→ LASSO also performs variable / feature selection as some of α will be zero.

Matrix form

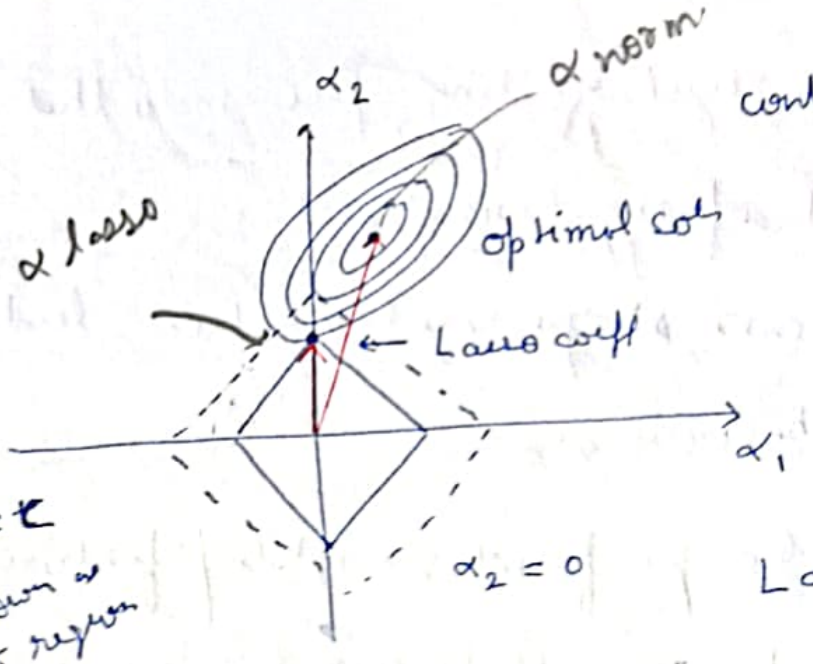
$$J(\alpha) = (y - X\alpha)^T (y - X\alpha) + \lambda \sum_{j=1}^{k} |\alpha_j|$$

Unlike Ridge regression, Lasso doesn't have a closed form solution ∵ the L1 regularization term is not differentable at zero.

Instead, numerical optimization techniques like coordinate descent or LARS (Least Angle Regression) are used to solved Lasso regularization

General idea's →

Iteratively update coefficient α in a way that minimize the objective function; applying a "soft threshold" operation that can shrink some coefficient to zero



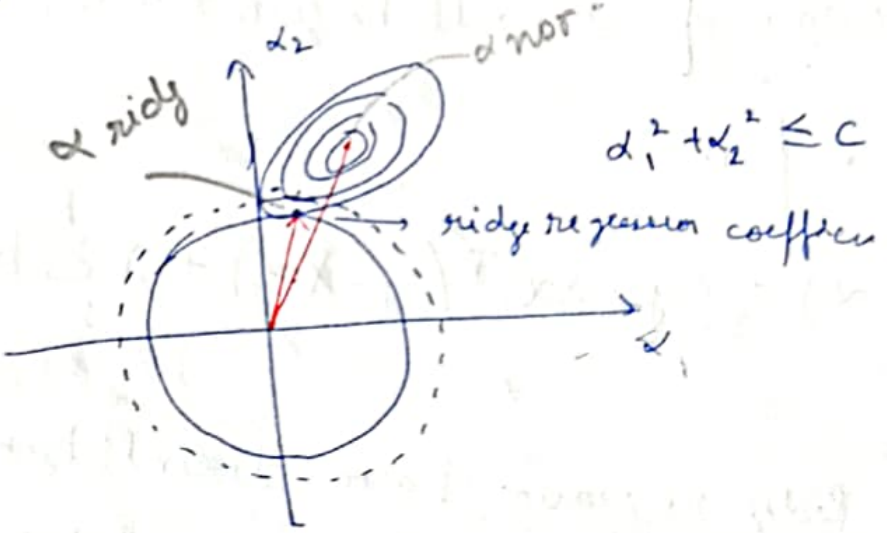
contours of RSS
↓
as it moves away from the minimum.

$|\alpha_1| + |\alpha_2| \leq t$

Penalty term shown as the constraint region

$\alpha_2 = 0$

Lasso



As λ increases, the c values ↓ so the size of diamond & circle shrink. A large λ imposes a stronger penalty forcing optimal solutions on the axes more frequently (for Lasso).

$$RSS + \lambda \sum_{j=1}^{p} |\alpha_j|^q$$