Multicollinearity

correlation $x - y \sim 0.95$

$\boxed{X_1 \quad X_2 \quad X_3} \quad y$

$\Downarrow$

$X_1 \quad X_2 \quad X_3 \quad X_4 \quad --- \quad X_{100}$

if $X_1 = X_2$

$$\left\{ \begin{matrix} 8x_1 + 2x_2 \\ 10x_1 \\ 10x_2 \\ 2x_1 + 8x_2 \end{matrix} \right\}$$

$X_1 \approx X_2 \sim y$

0.95

$X_1 \approx X_2$

$X_1 \approx (X_2 \; X_3 \; X_4)$

multi - col - linearity

multi → many

col → together

linearity → linear relationship

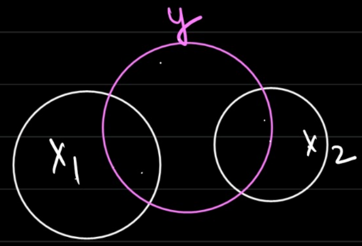
$X_1 \sim X_2 \rightarrow$ Correlation

$$\left. \begin{matrix} X_1 \approx (X_2 \; X_3) \\ X_1 \approx (X_2 \; X_3 \; X_4) \end{matrix} \right\} \text{multicollinearity.}$$

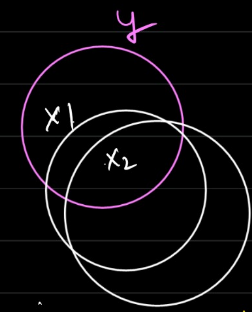
$X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad y$

Collinearity – where two features are linearly associate (high correlation) and they are used to predict target variable

multi - Collinearty – Where a feature exhibits a linear relationship with more than two variable.



No multicollinearity.



Multicollinearity

* Concern:-

→ It increases overfitting
→ Affects interpretation

$X_1 \approx X_2 \; X_3 \quad y$

$X_1 = X_2 \qquad 8X_1 + 2X_2 = y$

* Soln

① VIF and drop feature one by one with high VIF

VIF - Variance Inflation factor.

② RFE - Recursive feature elimination.

$X_1 - X_2 \rightarrow$ correlation | heatmap

$X \sim X_2 X_3 X_4 \rightarrow$ VIF

* VIF is a measure of amount of multicollinearity in regression.

$X_1 \quad X_2 \quad X_3 \quad y$

$X \approx X_2 \; X_3$

R_i^2 - %age variation in Y explained by X.

$$\boxed{VIF_i = \frac{1}{1 - R_i^2}}$$

$X_1 \; X_2 \; X_3 \; X_4 \; X_5$

Rsquare

$$VIF_{x_1} \longrightarrow x_1 \simeq \underbrace{(x_2\, x_3\, x_4\, x_5)}_{X}$$

$$\underset{y}{x_1} \longrightarrow X \longrightarrow R_i^2 \quad VIF = \frac{1}{1-R^2_{x_1}}$$

$X_1 X_2 X_3 \quad y$

Rsquare $= R^2 =$ %age variate explained in Y by X

✓ $X_1 \sim X_2 X_3$ ⟶ $VIF_{x_1} \sim X_1$ and $X_2 | X_3$ → R_{square}

✓ $X_2 \sim X_1 X_3$ $\quad \underset{y}{X_1} \quad \underset{X}{X_2 | X_3}$

✓ $X_3 \sim X_1 X_2$

$VIF = 1/1-R_i^2$

NO multicollinear

low multicolinearity

High multicollinearity

High multicollinearity

→ $VIF \geq 10$

When features have $VIF \geq 10$ then drop the feature one by one

$$\geq 10 = \frac{1}{1 - \text{Rsquare}}$$

$$1 - \text{Rsquare} = \frac{1}{10}$$

$$\text{Rsquare} = 1 - \frac{1}{10} \Rightarrow \frac{10-1}{10} = \frac{9}{10} = 0.9$$

$X - y$

$X_1 \sim (X_2 X_3 X_4)$

VIF

↓

$\underset{(y)}{X_1} \sim \underset{(X)}{X_2 X_3 X_4}$

$X_1 \sim X_2 X_3 X_4$

90% Variance in X_1 is explained $X_2 X_3 X_4$

Feature	VIF
X_1	12
✓ X_2	13
X_3	8
X_4	7

VIF $\geq$ 10 ⟹ drop the feature one by one.

* Two kinds of multicollinearity

✓ ① Data - based collinearity. — Present in data itself.

ex. latitude | longitude

✓ ② Structural multicollinearity — Caused due to new feature from existing feature.

Distance | time | speed = $\frac{\text{Distance}}{\text{time}}$

* 1000 → features

→ RFE — Recursive Feature Elimination

- It will make a model with all 1000 features
- Start dropping one by one, least important feature.

→ Untill the desired no of feature is achieved.

→ PCA