24/11/25

Ensemble Techniques:

Ensemble techniques are used to overcome the underfitting & overfitting problem, but with this we have two more for regression is Lasso & Ridge Regression.

* Actually Lasso & Ridge are called as regularization techniques.

Lasso Regression: L1 Regularization.

Lasso Regression adds a penalty to the linear regression loss function to reduce model complexity & prevent overfitting.

$$Loss = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{} |w_j|$$

λ - hyper parameter $w \rightarrow$ weightage

Adding λ can shrink some coefficients exactly to zero, Effectively performing feature selection. It helps to identify the most important predictors.

* It also works as feature selector.

When lasso?

- Whenever you want a simpler model & want to remove irrelevant features use lasso.
- It will be selecting the most impactful variables for prediction.

Ridge Regression (L2 Regularization Regression):

Ridge Regression adds a penalty to the linear regression loss function to reduce model complexity & prevent overfitting.

* It adds the sum of squares of the coefficients (weights) to the loss function.

$$Loss = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} w_j^2$$

λ - hyper parameter

w - weightage.

* λ is regularization parameter, it controls penalty strength.
 w model coefficients.

* It will shrink coefficients towards zero but never exactly zero.

* It keeps all features but reduces their influence.

Example:

we have 10 records   1 2 3 10

important :   1 2 4 6 7 8 10 → 95%

not so important 3 5 9   → 5%. 1 → 0
                                    ↓
                                  lasso

* When 95% of data is enough for me i go with lasso, when i want all the data then go with Ridge

* When we have many correlated features (multicollinearity)

Boston Housing dataset

Column descriptions:

CRIM → Crime rate — how much crime happens in the area.

ZN → Residential land — % of land used for large houses.

INDUS → Industrial Area — % of land used for factories & industries.

CHAS → 0 — House is not near river
1 — House is near river.

NOX → Air pollution level — Amount of nitrogen oxide in air
(higher = more pop polluted)

RM → Avg. no of rooms in house.

AGE → Old houses % — How many old houses (built before 1940) are in the area

DIS → distance to workplaces — How far the houses are from business job centres.

RAD → Road Access — How easily the area can be accessed from highw

TAX → property tax rate — The amount of tax paid for property.

PTRATIO → Student-teacher ratio — no. of students per teacher in schoo nearby.

B → population composition index — indicates the proportion of diff community groups.

LSTAT → Low-status population % — % of people from lower-income groups in the area.

MEDV → Median House Value — The price of the houses.

Target