

Penalization Methods

Regularization is a method used to make complex models simpler by penalising coefficients to reduce their magnitude, variance in the training set and in turn, reduce overfitting in the model. Regularization occurs by shrinking the coefficients in the model towards zero such that the complexity term added to the model will result in a bigger loss for models with a higher complexity. There are two types of regression techniques such as Ridge and Lasso regression.

Ridge Regression

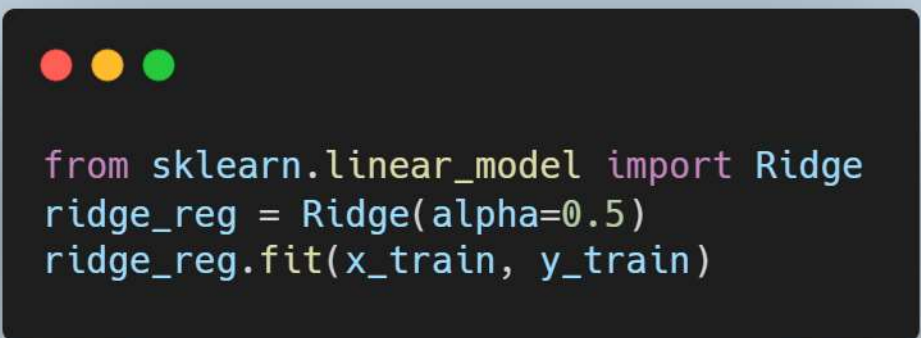
Also known as L2 Regularisation, this is a technique that uses a penalty term to shrink the magnitude of coefficients towards zero without eliminating them. The shrinkage prevents overfitting caused by the complexity of the model or collinearity. It includes the square magnitude of the coefficients to the loss function as the penalty term. If the error is defined as the square of residual, when a L2 regularization term is added, the result is the equation below.

$$\text{Loss with L2 regularisation} = \text{Error}(y, \hat{y}) + \lambda \sum_i^n w_i^2$$

where: w = weights of coefficients

λ = regularization parameter

As lambda increases, the penalty increases causing more coefficients to shrink in the same vein, if lambda is zero, it results in the loss function.



```
from sklearn.linear_model import Ridge
ridge_reg = Ridge(alpha=0.5)
ridge_reg.fit(x_train, y_train)
```

Feature Selection, The LASSO Regression and Elastic Net

- Feature Selection and Lasso Regression

Some datasets can be high dimensional with a very high number of features and some of them not contributing towards predicting the response variable. As a result, it becomes more computationally expensive to train a model and can also introduce noise causing the model to perform poorly. The process of selecting significant features that contribute the most in obtaining high performing models is known as feature selection. Lasso regression (Least Absolute Shrinkage and Selection Operator) reduces overfitting of the dataset by penalising the coefficients such that some coefficients are shrunk to zero and, indirectly performs feature selection by selecting only a subset of features leaving only relevant variables that minimize prediction errors. By using L1 regularisation, it includes the absolute value of the magnitude to the loss function. The application of L1 regularisation (Lasso regression) results in simpler and sparse models that allow for better interpretation. Although lasso regression helps prevent overfitting, one major limitation is that it does not consider other factors when eliminating predictors. For example, it arbitrarily eliminates a variable from a correlated pair which might not be a good rational from a human perspective. When a L1 regularization term is added, the result is the equation below.

$$\text{Loss with L1 regularisation} = \text{Error}(y, \hat{y}) + \lambda \sum_i^n |w_i|$$

```
from sklearn.linear_model import Lasso
lasso_reg = Lasso(alpha=0.001)
lasso_reg.fit(x_train, y_train)

#comparing the effects of regularisation
def get_weights_df(model, feat, col_name):
    #this function returns the weight of every feature
    weights = pd.Series(model.coef_, feat.columns).sort_values()
    weights_df = pd.DataFrame(weights).reset_index()
    weights_df.columns = ['Features', col_name]
    weights_df[col_name].round(3)
    return weights_df

linear_model_weights = get_weights_df(linear_model, x_train, 'Linear_Model_Weight')
ridge_weights_df = get_weights_df(ridge_reg, x_train, 'Ridge_Weight')
lasso_weights_df = get_weights_df(lasso_reg, x_train, 'Lasso_weight')

final_weights = pd.merge(linear_model_weights, ridge_weights_df, on='Features')
final_weights = pd.merge(final_weights, lasso_weights_df, on='Features')
```

Features	Linear_Model_Weight	Ridge_Weight	Lasso_weight
Surface Area	-208535169648.890289	-0.062275	0.000000
Relative Compactness	-64.612863	-0.283471	-0.027719
Orientation	-0.023080	0.003369	0.000000
Glazing Area Distribution	0.203771	0.029088	0.021431
Overall Height	4.170480	0.442467	0.463482
Glazing Area	19.931923	0.212449	0.206132
Wall Area	208535169648.863983	0.103061	0.200087
Roof_Area	417070339297.606750	-0.163192	-0.000000

- Elastic Net Regression

This is simply a combination of the L1 and L2 penalties from ridge and lasso regression. This method arose from the need to overcome the limitations of lasso regression. It regularizes and performs feature selection simultaneously by initially finding the optimal values of the coefficients as in ridge then performs a shrinkage.