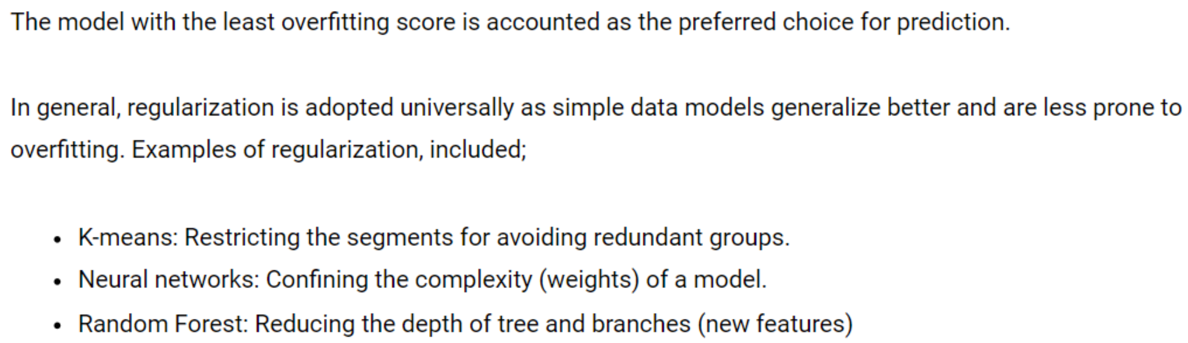
Regularization in Machine Learning

- ❖ Regularization means restricting a model to avoid overfitting by shrinking the coefficient estimates to zero.
- ❖ When a model suffers from overfitting, we should control the model's complexity.
- ❖ Technically, regularization avoids overfitting by adding a penalty to the model's loss function:

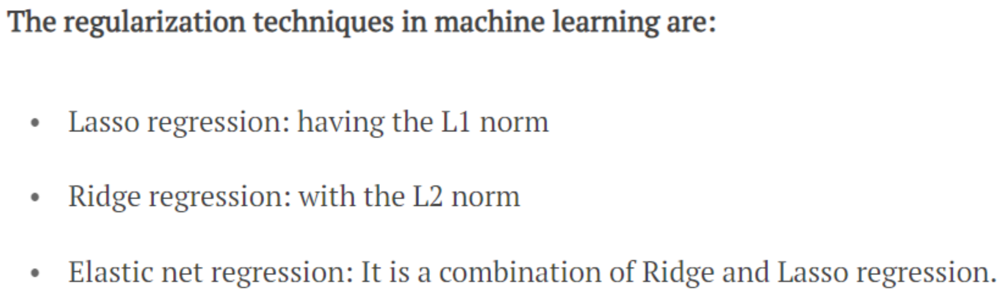
$$\text{Regularization} = \text{Loss Function} + \text{Penalty}$$



The model with the least overfitting score is accounted as the preferred choice for prediction.

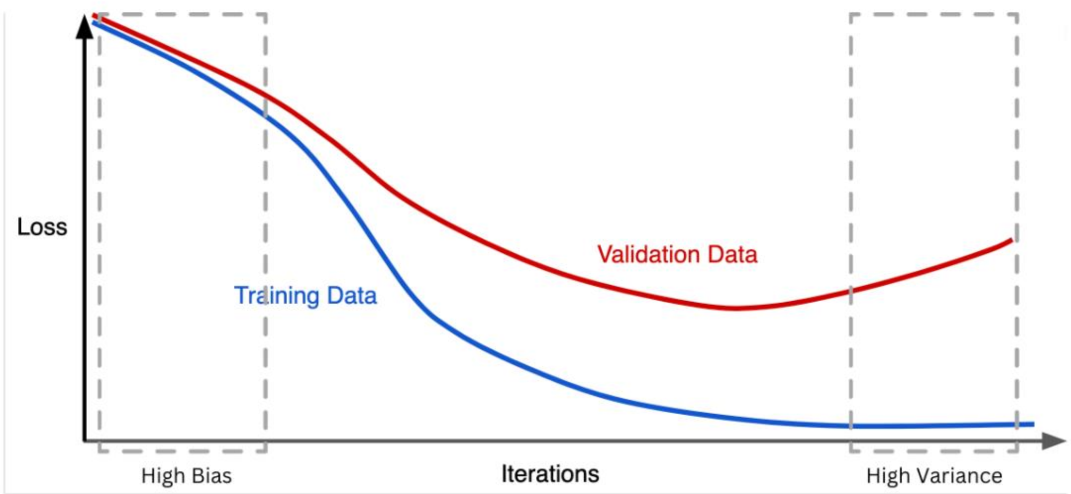
In general, regularization is adopted universally as simple data models generalize better and are less prone to overfitting. Examples of regularization, included;

- K-means: Restricting the segments for avoiding redundant groups.
- Neural networks: Confining the complexity (weights) of a model.
- Random Forest: Reducing the depth of tree and branches (new features)



The regularization techniques in machine learning are:

- Lasso regression: having the L1 norm
- Ridge regression: with the L2 norm
- Elastic net regression: It is a combination of Ridge and Lasso regression.



1. **L1 regularization:** It adds an L1 penalty that is equal to the absolute value of the magnitude of coefficient, or simply restricting the size of coefficients. For example, Lasso regression implements this method.
2. **L2 Regularization:** It adds an L2 penalty which is equal to the square of the magnitude of coefficients. For example, Ridge regression and SVM implement this method.
3. **Elastic Net:** When L1 and L2 regularization combine together, it becomes the elastic net method, it adds a hyperparameter.

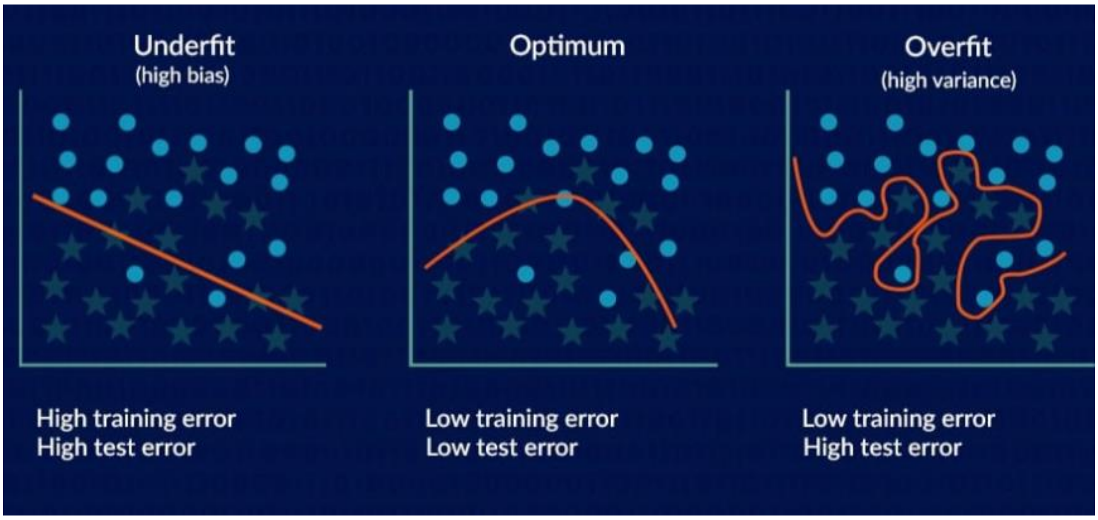
a. Ridge Regression

The Ridge regression technique is used to analyze the model where the variables may be having multicollinearity. It reduces the insignificant independent variables though it does not remove them completely. This type of regularization uses the L2 norm for regularization.

- It uses the L2-norm as the penalty.
- L2 penalty is the square of the magnitudes of beta coefficients.
- It is also known as L2-regularization.
- L2 shrinks the coefficients, however never make them to zero.
- The output of L2 regularization is non-sparse.

The cost function of the Ridge regression becomes:

$$\text{cost function} = \sum_{i=1}^{n} \left(y_{act} - y_{pred}\right)^2 + \lambda \cdot ||w||_2^2$$



b. Lasso Regression

Least Absolute Shrinkage and Selection Operator (or LASSO) Regression penalizes the coefficients to the extent that it becomes zero. It eliminates the insignificant independent variables. This regularization technique uses the L1 norm for regularization.

- It adds L1-norm as the penalty.
- L1 is the absolute value of the beta coefficients.
- It is also known as the L-1 regularization.
- The output of L1 regularization is sparse.

It is useful when there are many variables, as this technique can be used as a feature selection method by itself. The cost function for the LASSO regression is:

$$\text{cost function} \;=\; \sum_{i=1}^{n}\left(\mathrm{y}_{act}-\mathrm{y}_{pred}\right)^{2}+\lambda\cdot||w||_{1}$$

c. Elastic Net Regression

The Elastic Net Regression technique is a combination of the Ridge and Lasso regression technique. It is the linear combination of penalties for both the L1-norm and L2-norm regularization.

The model using elastic net regression allows the learning of the sparse model where some of the points are zero, similar to Lasso regularization, and yet maintains the Ridge regression properties. Therefore, the model is trained on both the L1 and L2 norms.

The cost function of Elastic Net Regression is:

$$\text{cost function} \;=\; \sum_{i-1}^{n}\left(\mathrm{y}_{act}-\mathrm{y}_{pred}\right)^{2}+\lambda_{\text{ridge}}\cdot||w||_{2}^{2}+\lambda_{\text{lasso}}\cdot||w||_{1}$$

The regularization parameters for the implementation of Elastic Net Regression are:

- λ ,and
- L1 ratio

where, λ = λ(Ridge) + λ(Lasso) and is written as:

$$L1\_ratio = \frac{\lambda_{\text{lasso}}}{\lambda_{\text{lasso}} + \lambda_{\text{ridge}}}$$

The formula for L1_ratio for all the methods becomes:

	Regularization Technique	Penalty
L1_ratio = 0	Ridge Regression	L-2
L1_ratio = 1	Lasso Regression	L-1
0< L1_ratio <1	Elastic Net Regression	Combination of L-1 and L-2

S.No	L1 Regularization	L2 Regularization
1	Panelizes the sum of absolute value of weights.	penalizes the sum of square weights.
2	It has a sparse solution.	It has a non-sparse solution.
3	It gives multiple solutions.	It has only one solution.
4	Constructed in feature selection.	No feature selection.
5	Robust to outliers.	Not robust to outliers.
6	It generates simple and interpretable models.	It gives more accurate predictions when the output variable is the function of whole input variables.
7	Unable to learn complex data patterns.	Able to learn complex data patterns.
8	Computationally inefficient over non-sparse conditions.	Computationally efficient because of having analytical solutions.

When to Use Which Regularization Technique?

The regularization in machine learning is used in following scenarios:

- Ridge regression is used when it is important to consider all the independent variables in the model or when many interactions are present. That is where collinearity or codependency is present amongst the variables.
- Lasso regression is applied when there are many predictors available and woulc want the model to make feature selection as well for us.
- When many variables are present, and we can’t determine whether to use Ridge or Lasso regression, then the Elastic-Net regression is your safe bet.