FEATURE SELECTION, REGULARIZATION, HYPERPARAMETER

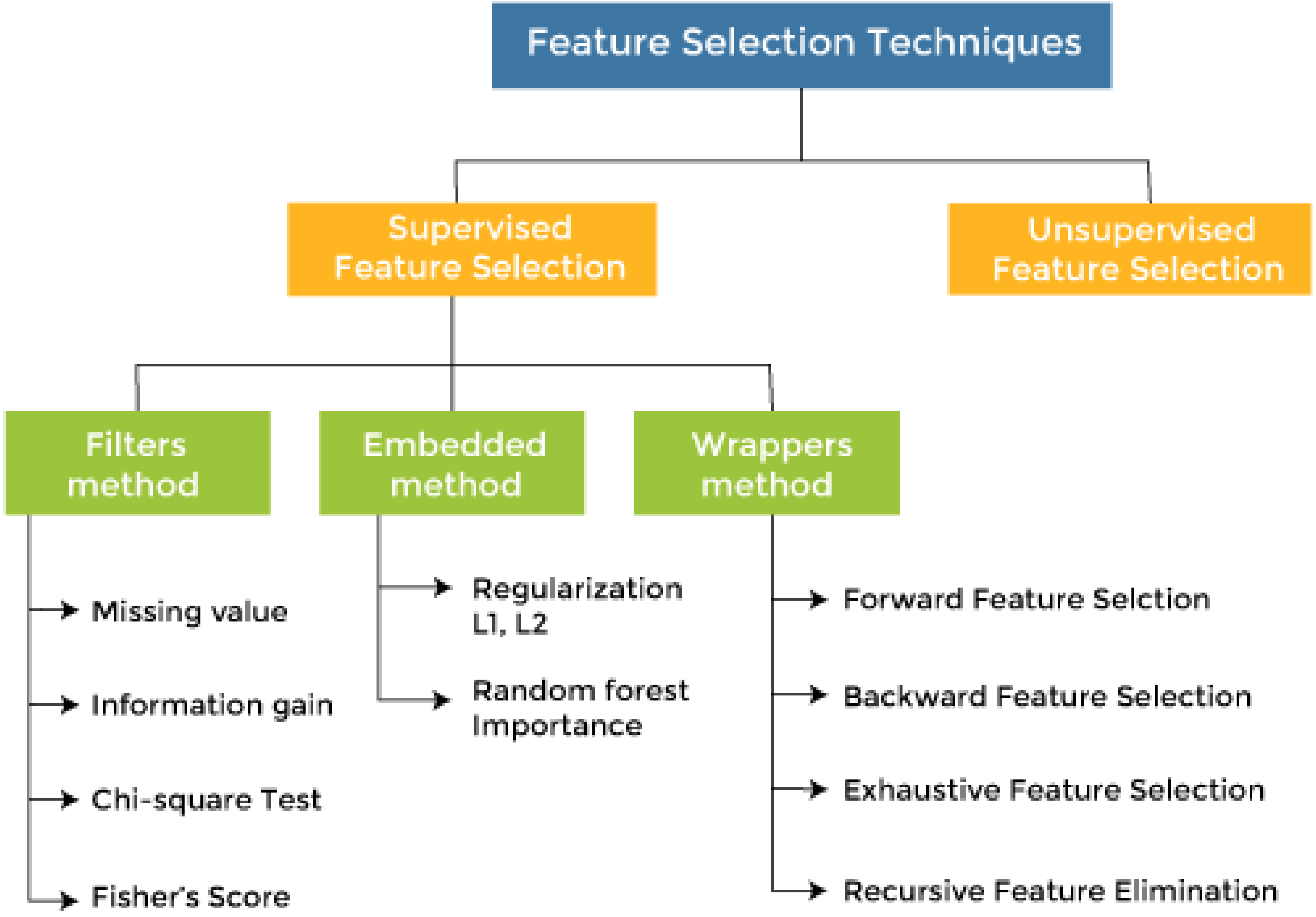
CA-2 HONORS AI
KHUSHI BADERA
3004

Feature Selection

- Feature selection is one of the important concepts of machine learning, which highly impacts the performance of the model. As machine learning works on the concept of "Garbage In Garbage Out", so we always need to input the most appropriate and relevant dataset to the model in order to get a better result.
- **A** feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection.

Technique for feature selection

- **Supervised Techniques:** These techniques can be used for labeled data and to identify the relevant features for increasing the efficiency of supervised models like classification and regression.
- **Unsupervised Techniques:** These techniques can be used for unlabeled data. For Example- K-Means Clustering, Principal Component Analysis, Hierarchical Clustering, etc.



Filter Methods

The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking. The advantage of using filter methods is that it needs low computational time and does not overfit the data.

Some common techniques of Filter methods are as follows:

- **Information Gain:** Information gain determines the reduction in entropy while transforming the dataset. It can be used as a feature selection technique by calculating the information gain of each variable with respect to the target variable.
- **Chi-square Test:** Chi-square test is a technique to determine the relationship between the categorical variables. The chi-square value is calculated between each feature and the target variable, and the desired number of features with the best chi-square value is selected.
- **Fisher's Score:** Fisher's score is one of the popular supervised technique of features selection. It returns the rank of the variable on the fisher's criteria in descending order. Then we can select the variables with a large fisher's score.
- **Missing Value Ratio:** The value of the missing value ratio can be used for evaluating the feature set against the threshold value. The formula for obtaining the missing value ratio is the number of missing values in each column divided by the total number of observations. The variable is having more than the threshold value can be dropped.

Wrapper Methods

In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations. It trains the algorithm by using the subset of features iteratively. On the basis of the output of the model, features are added or subtracted, and with this feature set, the model has trained again.

Some techniques of wrapper methods are:

- **Forward selection** - Forward selection is an iterative process, which begins with an empty set of features. After each iteration, it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not. The process continues until the addition of a new variable/feature does not improve the performance of the model.
- **Backward elimination** - Backward elimination is also an iterative approach, but it is the opposite of forward selection. This technique begins the process by considering all the features and removes the least significant feature. This elimination process continues until removing the features does not improve the performance of the model.
- **Exhaustive Feature Selection-** Exhaustive feature selection is one of the best feature selection methods, which evaluates each feature set as brute-force. It means this method tries & make each possible combination of features and return the best performing feature set.
- **Recursive Feature Elimination**- Recursive feature elimination is a recursive greedy optimization approach, where features are selected by recursively taking a smaller and smaller subset of features. Now, an estimator is trained with each set of features, and the importance of each feature is determined using coef_attribute or through a feature_importances_attribute.

Embedded Methods

Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost. These are fast processing methods similar to the filter method but more accurate than the filter method. These methods are also iterative, which evaluates each iteration, and optimally finds the most important features that contribute the most to training in a particular iteration.

Some techniques of embedded methods are:

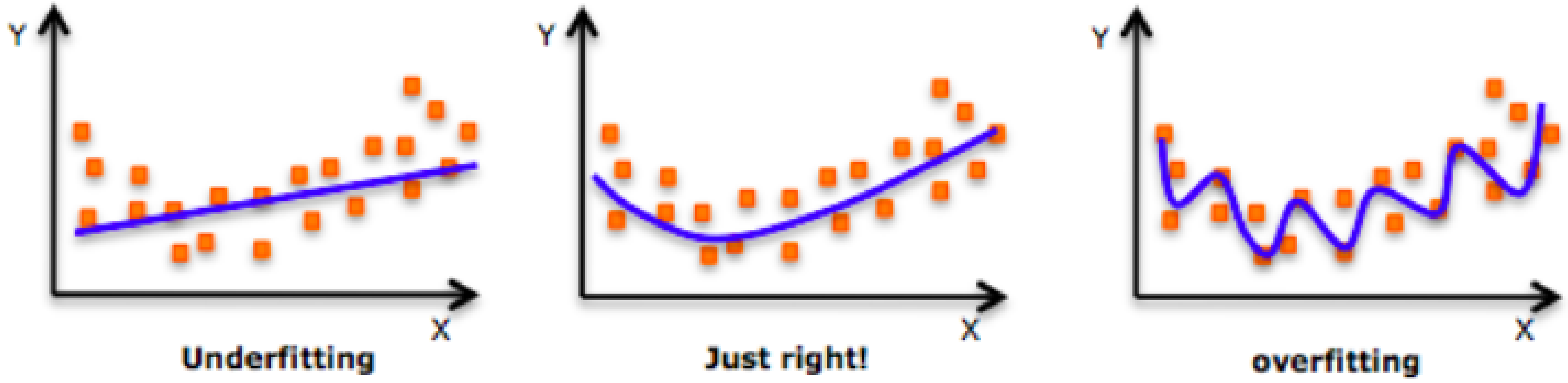
- **Regularization**- Regularization adds a penalty term to different parameters of the machine learning model for avoiding overfitting in the model. This penalty term is added to the coefficients; hence it shrinks some coefficients to zero. Those features with zero coefficients can be removed from the dataset. The types of regularization techniques are L1 Regularization (Lasso Regularization) or Ridge Regularization (L2 regularization).
- **Random Forest Importance** - Random Forest is such a tree-based method, which is a type of bagging algorithm that aggregates a different number of decision trees. It automatically ranks the nodes by their performance or decrease in the impurity (Gini impurity) over all the trees. Nodes are arranged as per the impurity values, and thus it allows to pruning of trees below a specific node. The remaining nodes create a subset of the most important features.

How to choose feature selection method?

Input Variable	Output Variable	Feature Selection technique
Numerical	Numerical	• Pearson's correlation coefficient (For linear Correlation). • Spearman's rank coefficient (for non-linear correlation).
Numerical	Categorical	• ANOVA correlation coefficient (linear). • Kendall's rank coefficient (nonlinear).
Categorical	Numerical	• Kendall's rank coefficient (linear). • ANOVA correlation coefficient (nonlinear).
Categorical	Categorical	• Chi-Squared test (contingency tables). • Mutual Information.

Regularization

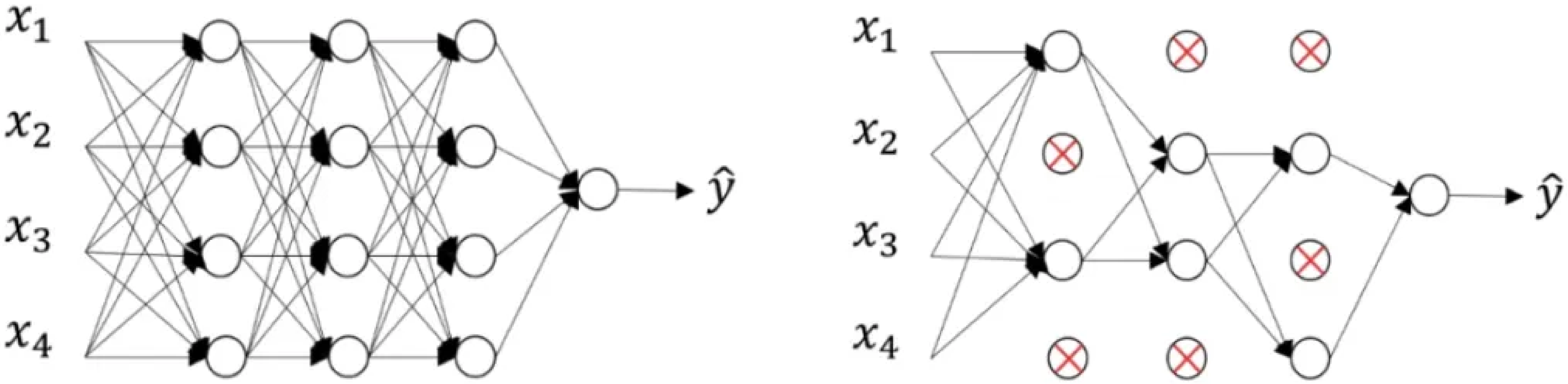
Regularization is a technique used in machine learning and deep learning to prevent overfitting and improve the generalization performance of a model. It involves adding a penalty term to the loss function during training.



Technique for Regularization

- **Lasso Regression:** A regression model which uses the L1 Regularization technique is called LASSO(Least Absolute Shrinkage and Selection Operator) regression. Lasso Regression adds the “absolute value of magnitude” of the coefficient as a penalty term to the loss function(L). Lasso regression also helps us achieve feature selection by penalizing the weights to approximately equal to zero if that feature does not serve any purpose in the model.
- **Ridge Regression:** A regression model that uses the L2 regularization technique is called Ridge regression. Ridge regression adds the “squared magnitude” of the coefficient as a penalty term to the loss function(L).

Dropout: In addition to the L2 and L1 regularization, another famous and powerful regularization technique is called the dropout regularization. The procedure behind dropout regularization is quite simple.In a nutshell, dropout means that during training with some probability P a neuron of the neural network gets turned off during training. Let’s look at a visual example.



HYPER PARAMETER

- In deep learning, hyperparameters are parameters that are not learned from the data but are set prior to training a neural network. They play a crucial role in the training process and have a significant impact on the model's performance.
- Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. The prefix ‘hyper_’ suggests that they are ‘top-level’ parameters that control the learning process and the model parameters that result from it.

Few Hyperparameter use in Deep Learning

- **Learning Rate**: The step size for adjusting model parameters during training, affecting the convergence and stability of the optimization process.
- **Number of Hidden Units**: The number of neurons or units in each hidden layer of a neural network, determining the network's capacity to model complex patterns.
- **Batch Size**: The number of data points used in each training iteration, impacting the trade-off between computational efficiency and convergence quality.
- **Activation Function**: The function applied to the output of each neuron, introducing non-linearity into the model to capture complex relationships in the data.
- **Regularization Strength**: A hyperparameter that controls the degree of regularization applied to the model, preventing overfitting by penalizing large weights.

THANK YOU!