Module 2

STATISTICAL MACHINE LEARNING

- Statistics is a branch of artificial intelligence that focuses on the development of algorithms and models for making predictions and decisions based on data.
- Involves the application of statistical methods and probabilistic models to build systems that can learn from and make predictions about data.

Parametric Method

- A function or a model that has some fixed number of parameters that summarizes the data called parametric functions. In simple terms, anyone can map input into output by using these parameters.
- The set of parameters is selected by first assuming the form of function we needed to map the input and output and then, selecting a model based on the assumption (parametric methods assume data are normally distributed or mathematically tractable frequency distributions which are closely related to the normal distribution, We then estimate of the parameters of the distribution data — in the case of the normal distribution — its mean and standard deviation.) to estimate the set of parameters.
- Ex: linear regression, naive Bayes, logistic regression, neural networks, etc.
- Parametric models are interpretable, require less data, and are fast. It has some disadvantages such as it may not perform well on complex problems.

Non-Parametric Method

- Non-parametric models don't have a fixed number of parameters. When constructing a function to map the input and output, the model never assumes the form of the function to be estimated, as a result, non-parametric models can learn any form of the function to map input and output.
- Non-parametric methods are more flexible because we are not caring about the set of parameters needed. They may lead to a better model shows as no assumptions are being made about the primary function.
- Ex: Decision tree, SVM, kNN, etc.
- It has some advantages such as being more accurate in prediction, It can form any complex functions, and also, no assumptions about data are required. The main disadvantage of the Non- parametric models is very lean toward fitting and requires a larger number of data.

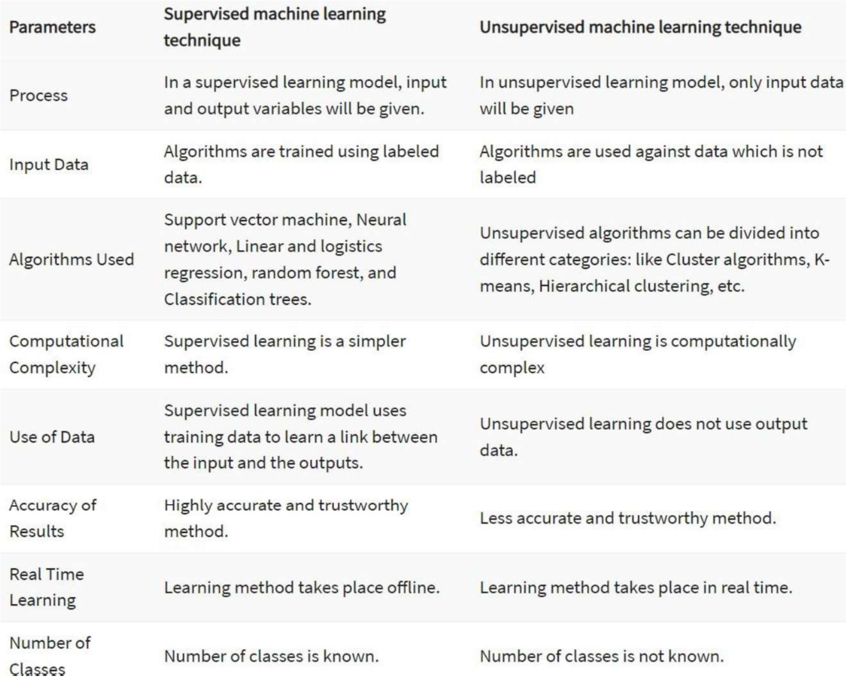
Supervised Learning

- Supervised learning is a type of machine learning where the algorithm is trained on a labeled dataset,
  i.e. the desired output is already known and provided in the dataset.
- The algorithm tries to learn a mapping between the inputs and the outputs and generalizes to unseen data.Examples of supervised learning problems are:
  - image classification
  - sentiment analysis
  - regression

Unsupervised Learning

- Is a type of machine learning where the algorithm is trained on an unlabeled dataset, i.e. the desired output is not provided.
- The algorithm tries to find patterns or structure in the data without any prior knowledge of the output.
- Examples of unsupervised learning problems are:
  - Clustering
  - Dimensionality reduction
  - Anomaly detection.

Supervised learning vs. unsupervised learning



Parameters	Supervised machine learning technique	Unsupervised machine learning technique
Process	In a supervised learning model, input and output variables will be given.	In unsupervised learning model, only input data will be given
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data which is not labeled
Algorithms Used	Support vector machine, Neural network, Linear and logistics regression, random forest, and Classification trees.	Unsupervised algorithms can be divided into different categories: like Cluster algorithms, K-means, Hierarchical clustering, etc.
Computational Complexity	Supervised learning is a simpler method.	Unsupervised learning is computationally complex
Use of Data	Supervised learning model uses training data to learn a link between the input and the outputs.	Unsupervised learning does not use output data.
Accuracy of Results	Highly accurate and trustworthy method.	Less accurate and trustworthy method.
Real Time Learning	Learning method takes place offline.	Learning method takes place in real time.
Number of Classes	Number of classes is known.	Number of classes is not known.

Key concepts in statistical machine

Semi-supervised learning: This is a hybrid of supervised and unsupervised learning, where a model is trained on a mix of labeled and unlabeled data.

Reinforcement learning: This involves training an agent to make decisions in an environment based on trial-and-error and the rewards it receives for each action it takes.

Model selection: This involves selecting the best model for a particular task based on its ability to generalize to unseen data.

Underfitting

A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data.

Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data.

Underfitting can be avoided by using more data and also reducing the features by feature selection. Underfitting refers to a model that can neither performs well on the training data nor generalize to new data.

Overfitting

A statistical model is said to be overfitted when the model does not make accurate predictions on testing data.
When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise.

The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. Overfitting is a problem where the evaluation of machine learning algorithms on training data is different from unseen data.

Quantitative variables

- Quantitative variables take on numerical values. Examples include a person's age, height, or income, the value of a house, categorical and the price of a stock.
- Regression problems require quantitative variables to predict values.

Qualitative variables

- Qualitative variablestake on values in one of K different classes, or categories.
- Examples of qualitative class variables include a person's gender (male or female), the brand of product purchased (brand A, B, or C), whether a person defaults on a debt (yes or no), or a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia)
- Qualitative response are often referred to as classification problems.

REGRESSION

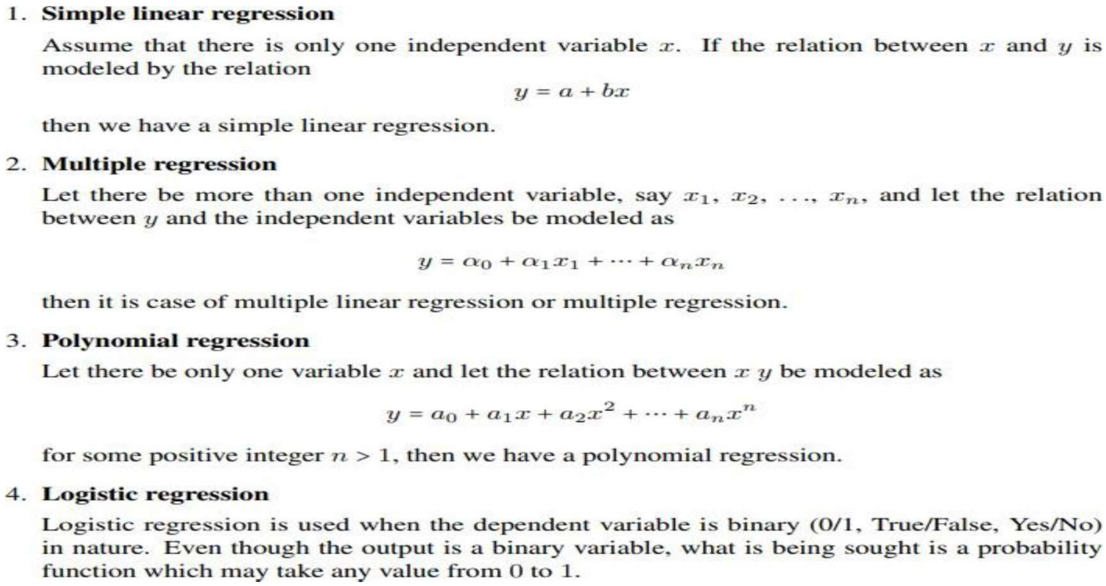
Regression is a supervised learning problem where there is an input x an output y and the task is to learn the mapping or relation between the inputs to the output. The approach in machine learning is that we assume a model, that is, a relation between x and y containing a set of parameters.

A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables and an output variable which is a real continuous variable.

- A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.
- Examples of regression problems include predicting housing prices, stock prices, or body mass index (BMI) based on relevant features such as size, location, age, etc.
- Dependent Variable: This is the variable that we are trying to understand or forecast.
- Independent Variable: These are factors that influence the analysis or target variable and provide us with information regarding the relationship of the variables with the target variable.

Different regression models

The different regression models are defined based on type of functions used to represent the relation between the dependent variable y and the independent variables



1. **Simple linear regression**

   Assume that there is only one independent variable x. If the relation between x and y is modeled by the relation
   $$y = a + bx$$
   then we have a simple linear regression.

2. **Multiple regression**

   Let there be more than one independent variable, say x_1, x_2, …, x_n, and let the relation between y and the independent variables be modeled as
   $$y = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_n x_n$$
   then it is case of multiple linear regression or multiple regression.

3. **Polynomial regression**

   Let there be only one variable x and let the relation between x y be modeled as
   $$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$
   for some positive integer $n > 1$, then we have a polynomial regression.

4. **Logistic regression**

   Logistic regression is used when the dependent variable is binary (0/1, True/False, Yes/No) in nature. Even though the output is a binary variable, what is being sought is a probability function which may take any value from 0 to 1.

LINEAR REGRESSION

Linear Regression is a predictive model used for finding the linear relationship between a dependent variable and one or more independent variables.One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

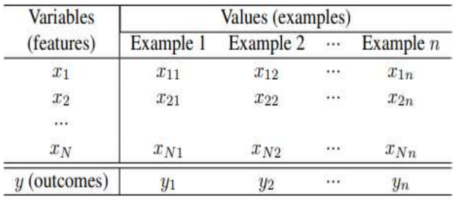
Here, we try to fit a line

$$y = \beta_0 + \beta_1 x$$

to our data. Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable.

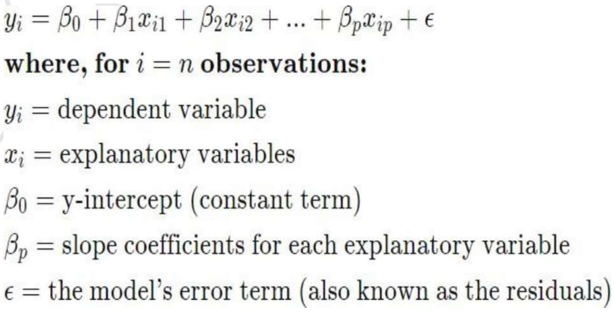
MULTIPLE LINEAR REGRESSION

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. We assume that there are N independent variables x_1, x_2,….x_N. Let the dependent variable be y. Let there also be n observed values of these variables:



Variables	Values (examples)			
(features)	Example 1	Example 2	$\cdots$	Example n
x_1	x_{11}	x_{12}	$\cdots$	x_{1n}
x_2	x_{21}	x_{22}	$\cdots$	x_{2n}
$\cdots$				
x_N	x_{N1}	x_{N2}	$\cdots$	x_{Nn}
y (outcomes)	y_1	y_2	$\cdots$	y_n

Data for multiple linear regression



$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

METHODS FOR ASSESSING MODELS

The most common methods for assessing models are:

- Residual Standard Error (RSE)
- R^2 Statistic
- Mean Squared Error (MSE)

Mean Squared Error (MSE):

Mean Squared Error is a measure of the difference between the predicted and actual values. It is calculated as the average of the squared differences between the predicted and actual values. The smaller the MSE, the better the model fits the data.

R-squared statistics:

- R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variable(s). It ranges from 0 to 1, where 1 means that the model fits the data perfectly and 0 means that the model explains none of the variance in the dependent variable. The higher the R-squared, the better the model fits the data.
- Both MSE and R-squared are commonly used to assess the goodness of fit of regression models, but they have their own strengths and limitations. MSE is more sensitive to outliers, while R- squared is more sensitive to the number of predictor variables in the model. Choosing the right evaluation metric depends on the specific goals and characteristics of the data and model.
- R-squared statistic, also known as the coefficient of determination, is a measure of the goodness-of-fit of a regression model. It provides information on how well the independent variables in a regression model explain the variability in the dependent variable.
- The R-squared statistic can be used to compare different regression models, and to determine whether additional predictor variables should be included in the model. However, it is important to keep in mind that a high R-squared value does not necessarily imply that the model is the best or the most accurate representation of the relationship between the independent and dependent variables.

To calculate R^2, we use the formula

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*

The Residual Standard Error (RSE):

The Residual Standard Error (RSE) is a statistical measure that indicates the variability of the residuals (the differences between the actual and predicted values) in a regression model. It is used to assess the goodness-of-fit of the model, and it is expressed as the standard deviation of the residuals. The RSE can be used to compare different regression models and to determine whether additional predictor variables should be included in the model. A smaller RSE value indicates a better fit of the model to the data.

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

CLASSIFICATION

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output. The best example of an ML classification algorithm is **Email Spam Detector**. The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.

Types of classification

- **Binary Classifier:** If the classification problem has only two possible outcomes, then it is called as Binary Classifier.
  Examples**:** YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc.

- **Multi-class Classifier:** If a classification problem has more than two outcomes, then it is called as Multi class Classifier.
  Example**:** Classifications of types of crops, Classification of types of music.

Types of ML classification algorithms

Linear Models

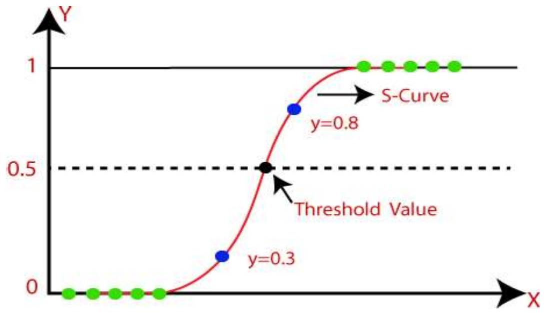
- Logistic Regression
- Support Vector Machines

Non-linear Models

- K-Nearest Neighbors
- Kernel SVM
- Naïve Bayes
- Decision Tree Classification
- Random Forest Classification

LOGISTIC REGRESSION

- Logistic regression is a Supervised Learning technique.
- It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The logistic regression is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.
- For example, spam detection in email service providers can be identified as a classification problem.This is s binary classification since there are only 2 classes as spam and not spam.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.



Sigmoid function (logit)

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the sigmoid function or the logistic function.
- Logistic regression is used to predict the likelihood of all kinds of “yes” or “no” outcomes.
- By predicting such outcomes, logistic regression helps data analysts (and the companies they work for) to make informed decisions.

Assumptions for Logistic Regression

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity: Multicollinearity is a statistical phenomenon in which two or more predictor variables in a regression model are highly correlated.

RANDOM FOREST

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks.

The algorithm for random forest

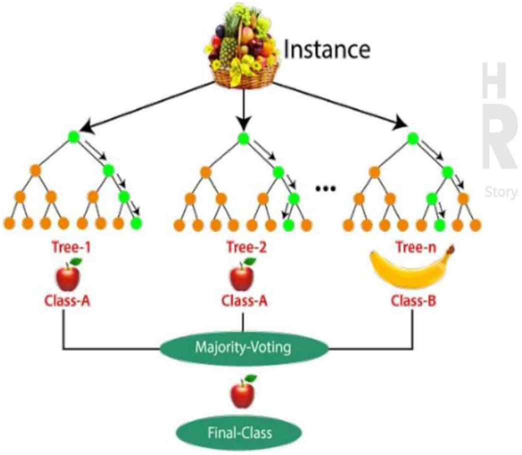
Step 1: In the Random forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression, respectively.

For example: consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket, and an individual decision tree is constructed for each sample. Each decision tree will generate an output, as shown in the figure. The final output is considered based on majority voting. In the below figure, you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.



Important Features of Random Forest

- **Diversity:** Not all attributes/variables/features are considered while making an individual tree; each tree is different.
- **Immune to the curse of dimensionality:** Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization:** Each tree is created independently out of different data and attributes. This means we can fully use the CPU to build random forests.
- **Train-Test split:** In a random forest, we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability:** Stability arises because the result is based on majority voting/ averaging

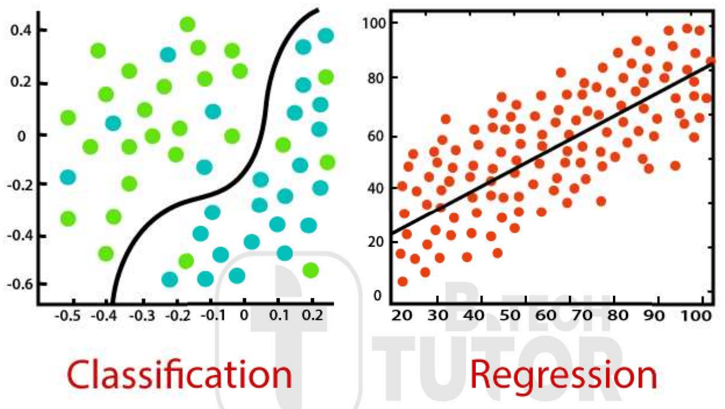
Regression and Classification in Machine Learning

- Regression and Classification algorithms are Supervised Learning algorithms. Both the algorithms are used for prediction in Machine learning and work with the labelled datasets. But the difference between both is how they are used for different machine learning problems.

- The main difference between Regression and Classification algorithms that Regression algorithms are used to predict the continuous values such as price, salary, age, etc. and Classification algorithms are used to predict/Classify the discrete values such as Male or Female, True or False, Spam or Not Spam, etc.

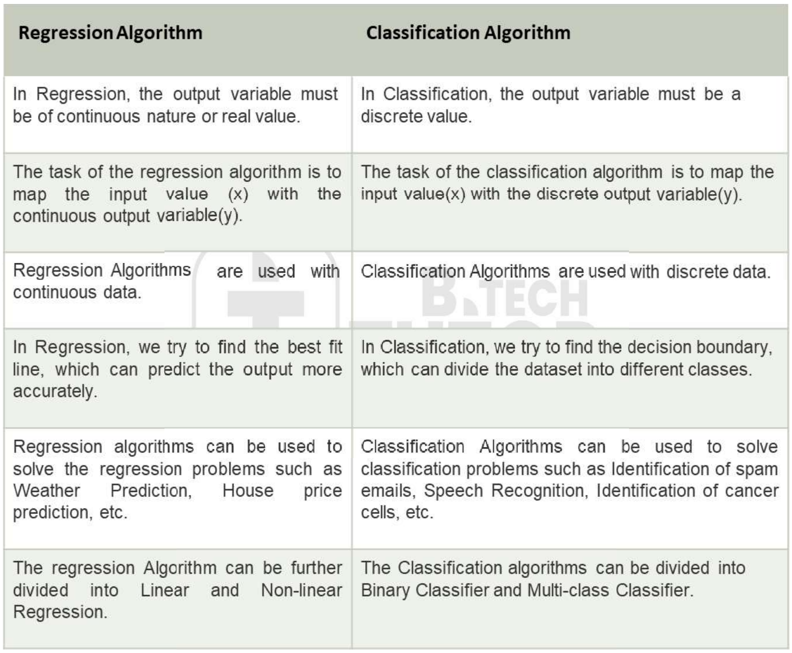
- Similarly the quantitative values/respon such as a person's age, height, or income, the value of a house, and the price of a stock etc. and these problems with a quantitative response as referred as regression problems.

- The qualitative responses such as person's marital status (married or not), the brand of product purchased (brand A, B, or C), whether a person is more likely to cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia) etc. Such problems with qualitative response are often referred to as classification problems.



Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of **Market Trends**, prediction of House prices, etc

Difference between Regression and Classification



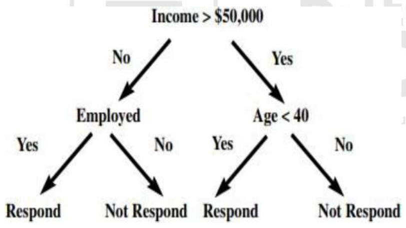
Regression Algorithm	Classification Algorithm
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.
The regression Algorithm can be further divided into Linear and Non-linear Regression.	The Classification algorithms can be divided into Binary Classifier and Multi-class Classifier.

LINEAR DISCRIMINANT ANALYSIS

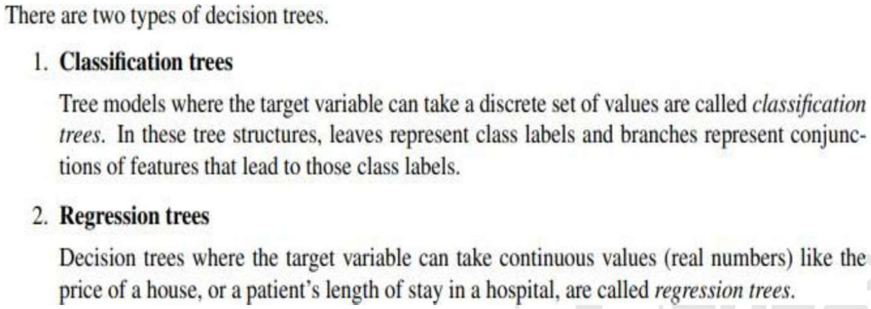
- Linear Discriminant Analysis (LDA) is a statistical technique used for classification problems in machine learning.
- LDA is a supervised learning algorithm, meaning it requires labeled training data.
- The goal of LDA is to find the best linear combination of features that separates the data into different classes with maximum separability.
- The class with the highest likelihood is then assigned as the predicted class for that data point.
- LDA is a linear method, meaning it finds the best linear boundary to separate the classes.
- It works well for datasets with linearly separable classes, and it is also computationally efficient.
- However, it may not perform well for datasets with non-linearly separable classes or a large number of features.
- LDA is often used in applications such as face recognition, text classification, and biomedical imaging analysis.

DECISION TREES

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision trees are recursive partitioning algorithms (RPAs) that come up with a tree-like structure representing patterns in an underlying data set. Figure 2 provides an example of a decision tree.



Types of Decision Tree



There are two types of decision trees.

1. **Classification trees**

   Tree models where the target variable can take a discrete set of values are called *classification trees*. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

2. **Regression trees**

   Decision trees where the target variable can take continuous values (real numbers) like the price of a house, or a patient's length of stay in a hospital, are called *regression trees*.

CLASSIFICATION TREE

The various elements in a classification tree are identified as follows:

- Nodes in the classification tree are identified by the feature names of the given data.
- Branches in the tree are identified by the values of features.
- The leaf nodes identified by are the class labels.

Stopping criteria:

- All (or nearly all) of the examples at the node have the same class.
- There are no remaining features to distinguish among the examples.
- The tree has grown to a predefined size limit.

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.

The complete process can be better understood using the below algorithm:

- Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
- Step-2: Find the best attribute in the dataset using *Attribute Selection Measure (ASM).*
- Step-3: Divide the S into subsets that contains possible values for the best attributes.
- Step-4: Generate the decision tree node, which contains the best attribute.
- Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Feature selection measures

- **Entropy**
  The degree to which a subset of examples contains only a single class is known as purity, and
  any subset composed of only a single class is called a pure class.
  Informally, entropy is a measure of "impurity" in a dataset.

  Let the data segment S has only two class labels, say, "yes" and "no". If p is the proportion of examples having the label "yes" then the proportion of examples having label "no" will be $1 - p$. In this case, the entropy of S is given by:

$$\text{Entropy}(S) = -p\log_2(p) - (1-p)\log_2(1-p).$$

- **Information gain**
  Information Gain is the expected reduction in entropy caused by partitioning the examples according to a given attribute.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \times \text{Entropy}(S_v).$$

Decision Tree Algorithm

1. Place the "best" feature (or, attribute) of the dataset at the root of the tree.

2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for a feature.

3. Repeat Step 1 and Step 2 on each subset until we find leaf nodes in all the branches of the tree.

Issues in decision tree learning

1. Avoiding overfitting of data When we construct a decision tree, the various branches are grown just deeply enough to perfectly classify the training examples. This leads to difficulties when there is noise in the data or when the number of training examples are too small. In these cases, the algorithm can produce trees that overfit the training examples.

   We say that a hypothesis *overfits* the training examples if some other hypothesis that fits the training examples less well actually performs better over the entire distribution of instances, including instances beyond the training set.

   Approaches to avoiding overfitting

   - The main approach to avoid overfitting is pruning.
   - Pruning is a technique that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances.
   - Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.
   - We may apply pruning earlier, that is, before it reaches the point where it perfectly classifies the training data.
   - We may allow the tree to overfit the data, and then post-prune the true

Reduced error pruning

- In reduced-error pruning, we consider each of the decision tress to be a candidate for pruning.
- Pruning a decision node consists of removing the subtree rooted at that node, making it a leaf node, and assigning it the most common classification of the training examples affiliated to that node.
- Nodes are removed only if the resulting pruned tree performs no worse than the original over validation set.
- Nodes are pruned iteratively, always choosing the node whose removal most increases the accuracy over the validation set. Pruning of nodes is continued until further pruning decreases the accuracy over the validation set.

2. Problem of missing attributes

- The following are some of the methods used to handle the problem of missing attributes.
- Deleting cases with missing attribute values
- Replacing a missing attribute value by the most common value of that attribute
- Assigning all possible values to the missing attribute value
- Replacing a missing attribute value by the mean for numerical attributes
- Assigning to a missing attribute value the corresponding value taken from the closest t cases, or replacing a missing attribute value by a new value

REGRESSION TREE

A regression problem is the problem of determining a relation between one or more independent variables and an output variable which is a real continuous variable and then using the relation to predict the values of the dependent variables.

Regression problems are in general related to prediction of numerical values of variables. Trees can also be used to make such predictions. A tree used for making predictions of numerical variables is called a prediction tree or a regression tree.

An algorithm for constructing regression trees

Starting with a learning sample, three elements are necessary to determine a regression tree:

1. A way to select a split at every intermediate node
2. A rule for determining when a node is terminal
3. A rule for assigning a value for the output variable to every terminal node

The decision tree in the image follows a top-down, greedy approach. It starts from a single node, which contains all the data points. Then, it makes a decision at each node by splitting the data points based on a single feature. The feature that results in the greatest reduction in a cost function (entropy in the case of classification) is chosen for the split. This process continues recursively until a stopping condition is met, such as reaching a leaf node where all data points belong to the same class (classification) or have similar values for the target variable (regression).

The text in the image outlines the steps involved in building a decision tree:

1. Start with a single node containing all data points. Calculate the information gain (entropy) for this node.
2. If all the points in the node have the same value for all the independent variables, stop. This node is a leaf node, and it will represent the class or value that all the data points in the node belong to.
3. Otherwise, search over all possible binary splits of all variables to find the one that will reduce the information gain (entropy) as much as possible. This is the best split, as it will create two child nodes that are more homogeneous than the parent node.
4. If the largest decrease in information gain is less than a predefined threshold, or if one of the resulting nodes would contain too few data points (less than a minimum number of samples, q), stop and assign the majority class (or average value for regression) to the node. This is to prevent overfitting the data.
5. Otherwise, create two new child nodes based on the chosen split.
6. In each new child node, go back to step 1.

Advantages and Disadvantages of Trees

Decision trees for regression and classification have a number of advantages over the more classical approaches:

- Trees are very easy to explain to people. In fact, they are even easier to explain than linear regression!
- Some people believe that decision trees more closely mirror human decision-making than do the regression and classification approaches.
- Trees can be displayed graphically, and are easily interpreted even by a non-expert (especially if they are small).
- Trees can easily handle qualitative predictors without the need to create dummy variables.

The disadvantages of trees are:

- Trees generally do not have the same level of predictive accuracy
- Additionally, trees can be very non-robust. In other words, a small change in the data can cause a large change in the final estimated tree.

ENSEMBLE MODELS

Ensemble methods aim at improving the predictive performance of a given statistical learning or model fitting technique. The general principle of ensemble methods is to construct a linear combination of some model fitting method, instead of using a single fit of the method.

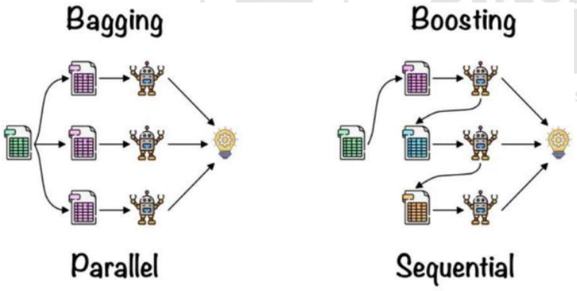
An ensemble is itself a supervised learning algorithm because it can be trained and then used to make predictions. Ensemble methods combine several decision trees classifiers to produce better predictive performance than a single decision tree classifier. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner, thus increasing the accuracy of the model. When we try to predict the target variable using any machine learning technique, the main causes of the difference in actual and predicted values are **noise, variance, and bias**. Ensemble helps to reduce these factors (except noise, which is irreducible error). The noise-related error is mainly due to noise in the training data and can't be removed..

However, the errors due to bias and variance can be reduced.

Ensemble uses two types of methods:

1.Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

2.Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.



Advantages of Ensemble Algorithm

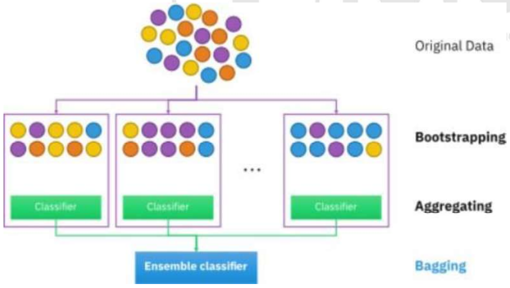
- Ensemble is a proven method for improving the accuracy of the model and works in most of the cases.
- Ensemble makes the model more robust and stable thus ensuring decent performance on the test cases in most scenarios.
- You can use ensemble to capture linear and simple as well nonlinear complex relationships in the data. This can be done by using two different models and forming an ensemble of two.

Disadvantages of Ensemble Algorithm

- Ensemble reduces the model interpret-ability and makes it very difficult to draw any crucial business insights at the end
- It is time-consuming and thus might not be the best idea for real-time applications
- The selection of models for creating an ensemble is an art which is really hard to master

BAGGING

Bagging, also known as Bootstrap Aggregation, is the ensemble technique used by random forest.Bagging chooses a random sample/random subset from the entire data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently, which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting, is known as aggregation.



The bagging algorithm, which has three basic steps:

1. Bootstrapping: Bagging leverages a bootstrapping sampling technique to create diverse samples. This resampling method generates different subsets of the training dataset by selecting data points at random and with replacement. This means that each time you select a data point from the training dataset, you are able to select the same instance multiple times. As a result, a value/instance repeated twice (or more) in a sample.
2. Parallel training: These bootstrap samples are then trained independently and in parallel with each other using weak or base learners.
3. Aggregation: Finally, depending on the task (i.e. regression or classification), an average or a majority of the predictions are taken to compute a more accurate estimate. In the case of regression, an average is taken of all the outputs predicted by the individual classifiers; this is known as soft voting. For classification problems, the class with the highest majority of votes is accepted; this is known as hard voting or majority voting.

The algorithm for random forest

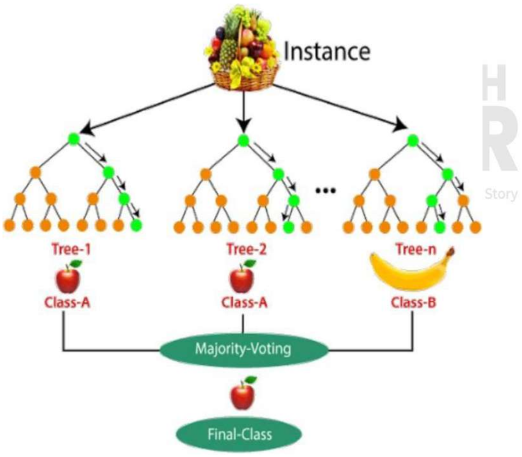
Step 1: In the Random forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.

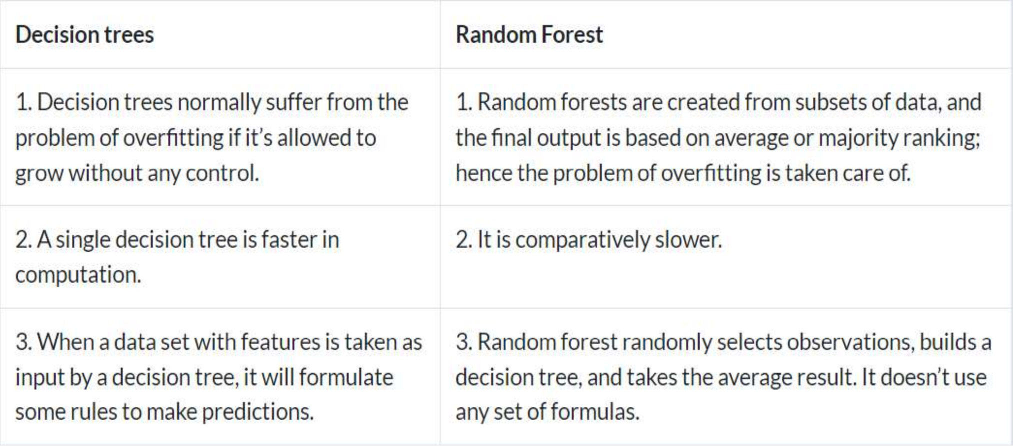
Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression, respectively.

For example: consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket, and an individual decision tree is constructed for each sample. Each decision tree will generate an output, as shown in the figure. The final output is considered based on majority voting. In the below figure, you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.





Decision trees	Random Forest
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data, and the final output is based on average or majority ranking; hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation.	2. It is comparatively slower.
3. When a data set with features is taken as input by a decision tree, it will formulate some rules to make predictions.	3. Random forest randomly selects observations, builds a decision tree, and takes the average result. It doesn't use any set of formulas.

IMPORTANT QUESTIONS

- Is regression a supervised learning technique? Justify your answer. Compare it with classification giving examples.
- Explain random forest ensemble method with an example.
- What is decision tree? Explain the working of decision tree
- What are ensemble methods? Explain the bagging technique.
- Differentiate supervised and unsupervised learning techniques with examples.
- Discuss linear discriminant analysis.
- Compare parametric and non-parametric methods in statistical learning

THANK YOU!