RGPV CS-601 ML Unit-1 Detailed Notes pt

Machine Learning (Rajiv Gandhi Proudyogiki Vishwavidyalaya)

ML - Unit : 1

* Machine Learning

The ability of a computer to perform tasks without being explicitly programmed to do so is ML.

ML is a subset of AI that focuses on development of comp. algo. that improve automatically through experience and use of data. In simpler terms, ML enables comp. to learn from data & make decisions without being explicitly programmed to do so.

ML is all about creating & implementing algo. that facilitate these decisions & predictions and can imitate intelligent human behaviour. These algo. are designed to improve their performance over time, becoming more effective & accurate as they process more data.

To understand ML better, let us consider an example.
If we want a comp. to recognize images of dogs, we don't provide it with specific instructions on what a dog looks like. Instead, we give it thousands of images of dogs & let the ML algo. figure out the common patterns & features that define a dog. Over time, as the algo. processes

more images, it gets better at recognising a dog, even when presented with images it has never seen before.

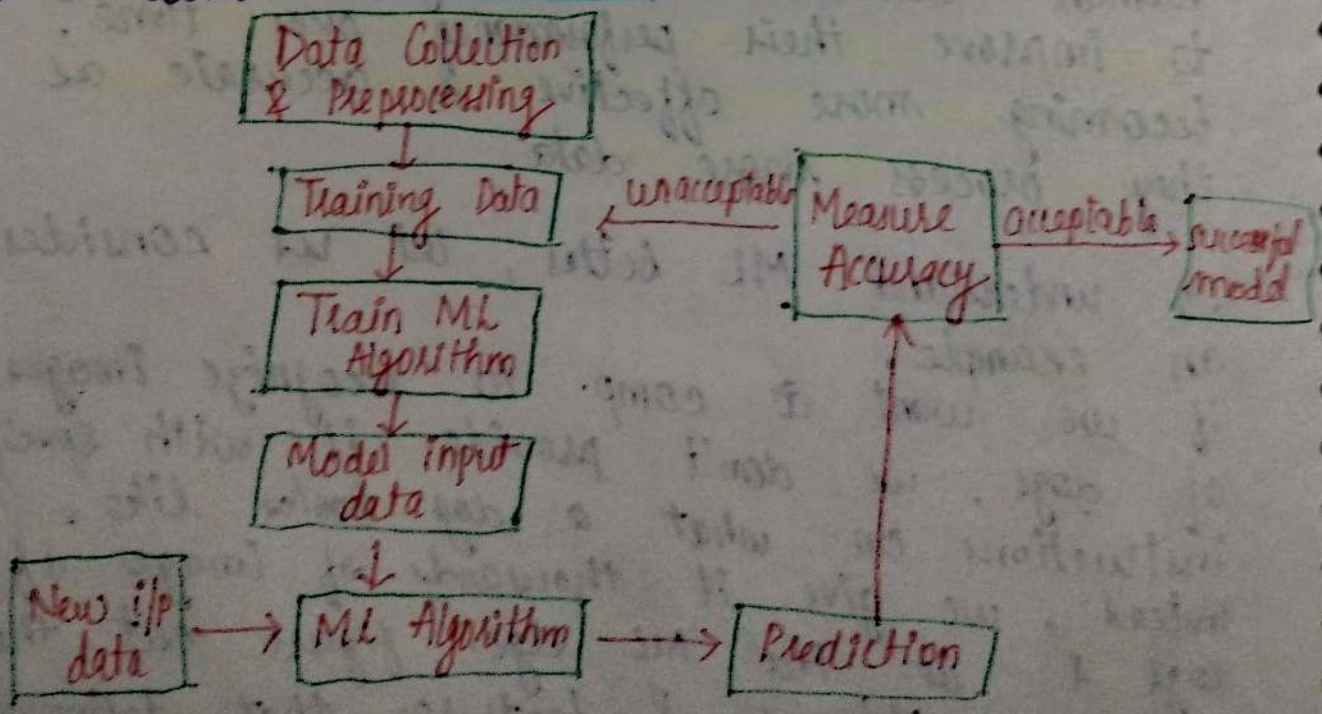
Thus, we can conclude that in ML, the comp. is given a task to perform & some data, but it's up to the comp. to figure out, how to accomplish that task.

Also, it must be noted that, ML is the only way humans have managed to reach AI. #All data

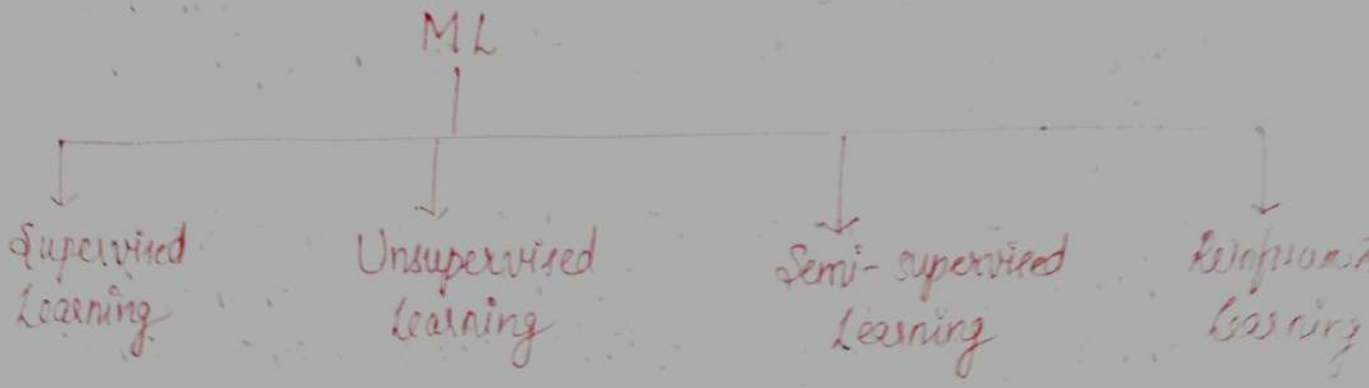
Key Components of any ML process

1. Data
2. Model
3. Objective fn
4. Optimization algo.

How does ML work?



Types of ML



Applications of ML in various industries

1. Healthcare

- wearable devices & ~~sentonds~~ sensors such as wearable fitness trackers, smart health watches, etc. help monitor users' health data to assess their health in ~~to~~ real time.
- ML algo. allow medical experts to predict a patient's lifespan with ↑ accuracy.
- Drug discovery – ML helps speeding up the process. e.g. Pfizer uses IBM's Watson to analyze massive volumes of disparate data for drug discovery.
- Personalized treatment – companies like Genentech have collaborated with GNS Healthcare to leverage ML to analyze individual genes to provide targeted therapies.

2. Finance

mainly ML is used to detect fraudulent activity

e.g.
- Citibank has partnered with Feedzai, a fraud detection company to handle online & in-person banking frauds.
- Paypal uses several ML tools to differentiate b/w legitimate & fraud transactions.

3. Retail

Retail websites use ML to recommend items based on users' purchase / search history, :

They also implement ML for marketing campaigns, customer insights, merchandise planning, & price optimization & client retention.

e.g. • Amazon, Flipkart, etc.
• Netflix, Youtube.

4. Travel

Rides offered by Uber, Ola & even self-drive cars have an extensive & robust ML backend.

ML is also implemented to analyze customer reviews through sentiment analysis.
ML is also used for campaign monitoring, brand monitoring, compliance monitoring, etc.

e.g. Uber's dynamic pricing

Uber uses a ML model called 'Geosurge' to manage dynamic pricing parameters. It uses real-time predictive modeling on traffic patterns, supply & demand to decide fare.

5. Social Media

ML is pivotal in driving SM platforms from personalizing news feeds to delivering user-specific ads.

eg.
- FB auto-tagging feature employs image recognisition to identify your & friend's face & automatically tags them.
- LinkedIn employs ML to tell user where they should apply, whom they should connect with & how their skills rank compared to others.

Other applications of ML

1. Chatbots.
2. AI assistants & search engines.
3. Self service tools like Google maps.
4. Augmented Reality (AR)
5. Automobile industry e.g. self-driving cars
6. TinyML - it allows IoT edge devices to run ML driven processes.
   e.g. wake up commands like 'Hey Siri' falls under TinyML.

ML Tools

1. Python
2. R
3. Tensorflow
4. Scikit Learn
5. Keras
6. Pytorch.

* Scope of ML

ML has a vast scope & is used in various different fields.
Its scope is ↑ day by day as more & more companies are adopting ML technologies for advancement of their businesses.

1. **Data Analysis** – due to excessive production of data, we need a method that can be used to structure, analyze & draw useful insights from data. This is where ML comes in. It uses data to solve problems & find solutions.

2. **Improved Decision Making**: By making use of various ML algo, companies can make better business decisions.
   e.g. ML is used for sales forecasting; predict downfalls in stock market, identify risks & anomalies, etc.

3. **Uncover patterns & trends** in data that humans would not be able to detect.
   e.g. financial institutions use ML to detect frauds & prevent money laundering.

4. **Image & speech recognition** : ML algo. can analyze images & speech data & recognize patterns, making it possible to identify objects, people & even emotions accurately.
   e.g. self-drive cars use ML to identify road signs & objects on roads.

* Limitations of ML

1. Data Acquisition ML requires massive datasets to work effectively and this ~~must~~ data must be inclusive, unbiased & of good quality. This particularly poses problems for small businesses, which may not have access to large data.

2. Bias and Discrimination ML algo. learn from historical data, which may contain biases & thus influence the predictions.

   e.g. • according to a research, facial recognition doesn't perform accurately on people with darker skin tones, which causes false +ve or false -ve rates to higher for people with dark races.

3. Overfitting & Underfitting
   A model that is too simple to capture data complexities is an underfitting model. Underfitting represents inability of model to learn training data effectively which results in poor performance both on training & testing data.

   Overfitting occurs when a model learns noise in the training data rather than actual underlying patterns. It can lead to poor generalization.

4. Computational Resources Training complex ML models may be computationally intensive, expensive and require substantial resources to be successfully trained.

5. Lack of causalty ML models may not shed light on the underlying causal links in the data because correlation doesn't always imply causation. This may reduce our capacity for precise prediction when causalty is crucial.

6. Interpretability Some complex ML models like deep NN are challenging to interpret making it difficult to explain their decisions.

7. Ethical Considerations ML models can have major social, ethical & legal repercussions when used to make judgements that affect peoples lives.
e.g. Racial

* Regression Analysis

Regression analysis is a set of statistical methods used for the estimation of relationships b/w a dependent variable (target) and one or more independent variables (predictor)

Regression analysis is used for forecasting, time-series modelling and finding the causal effect rel^{n}ship b/w the variables.

It is an imp. tool for modelling and analyzing data. Here we fit a curve/line to the data points, in such a manner that the differences b/w the distances of data points from the curve or line is minimized.

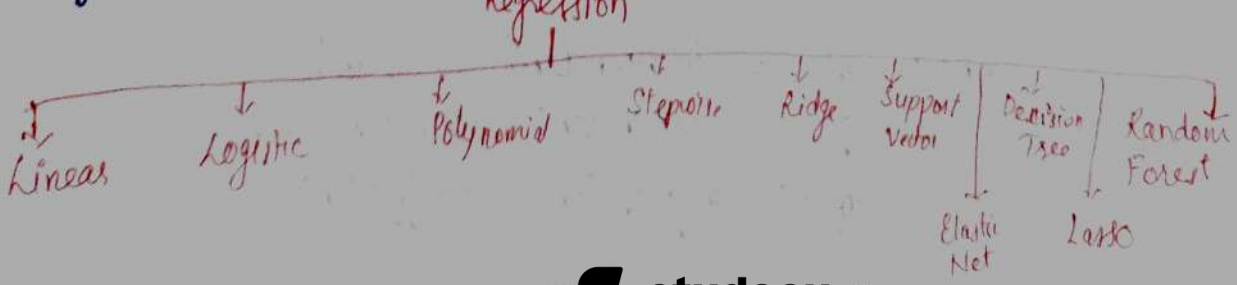
Thus, a regression model is a powerful tool in ML & is used for predicting continuous values based on rel^{n}ship b/w target & predictor.

Types of Regression Techniques

There are various kinds of regression techniques, however, these are mostly driven by 3 metrics:

i. No. of independent variables
ii. Type of dependent variable
ii. Shape of regression line.

~~Based on these metrics~~, most commonly used regression techniques are:

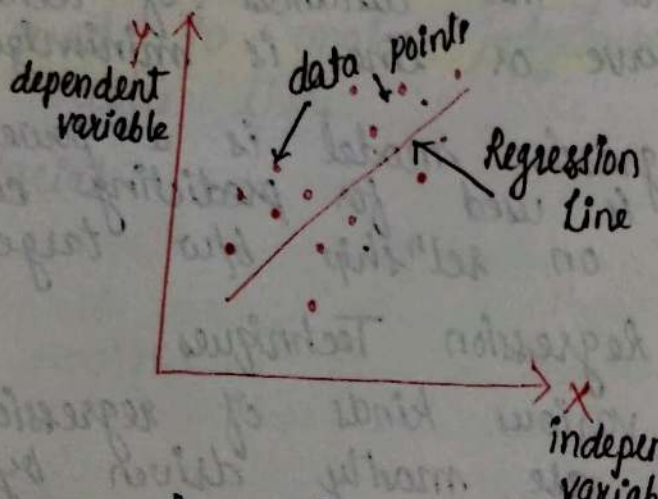


0 Linear Regression

Linear regression is an algo. that predicts the relationship b/w two variables by assuming a linear correlation b/w the independent and dependent variables.

It seeks the optimal line that minimizes the sum of squared differences b/w predicted and actual values.

Linear regression is applied in various domains like economics & finance & it seeks to analyze and forecast data trends.



Linear regression line is a straight line that goes through plots on the graph while being as close as it can be to all of them simultaneously.

The eqⁿ for linear regression is:

$$y(x) = P_0 + P_1 x$$

where $x \rightarrow$ independent variable
$y \rightarrow$ dependent variable
$P_1 \rightarrow$ regression coefficient
$P_0 \rightarrow$ bias term

NOTE Best fit line is a line that fits the given scatter plot in the best way.

Mathematically, the best fit line is obtained by minimising the Residual Sum of Squares (RSS)

Types of Linear Regression
1. Simple
2. Multiple

Benefits of Linear Regression

1. Simplicity and Interpretability
2. Prediction : allows you to predict future values based on existing data.
3. Scalability : not computationally heavy, hence fits well where scaling is essential.
4. Optimal for online settings : because of ease of computation.
5. Widespread applicability : used in various fields
6. Foundation for other techniques

Assumptions of Linear Regression

1. Linearity of Residuals : There ~~need~~ must be a linear relationship b/w target & predictor.
2. Independence of Residuals : The error terms should not be dependent on one another. There should be no correlation b/w the residual terms. The nice of this phenomenon is called autocorrelation
3. Normal distribution of residuals : The mean of residual terms should follow a normal distribution with a mean equal to zero or close to zero. This is done in order to check whether the selected line is actually the line of best fit or not.

If error terms are non-normally distributed, it suggests that there are few unusual data points that must be studied closely to make a better model.

4. The equal variance of residuals
The error terms must have constant variance. This phenomenon is known as ~~homos~~ Homoscedasticity.

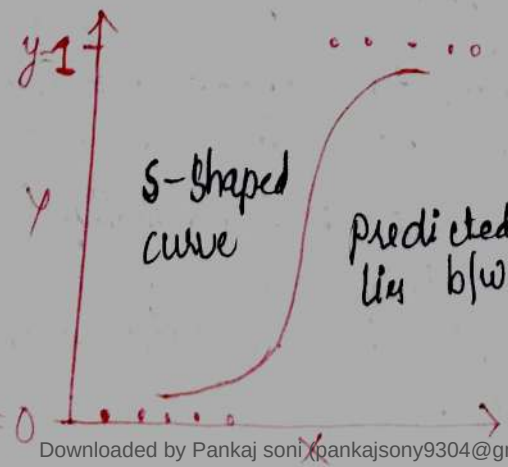
The case of non-constant variance is Heteroscedasticity.

0 Logistic Regression

Logistic regression is a supervised ML algo. that accomplishes binary classification tasks by predicting the probability of an outcome, event or observation.

Logistic regression model delivers a binary or dichotomous outcome limited to two possible outcomes : yes/no, 0/1, True/False. Hence, this model is useful during decision making process.

Logistic regression analyzes the relationship b/w one or more independent variables & classify data into discrete classes. It is extensively used in predictive modelling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.



Logistic regression eqn

$$y = \frac{e^{(b_0 + b_1 x)}}{1 + e^{(b_0 + b_1 x)}}$$

where $x \rightarrow$ i/p value
$y \rightarrow$ predicted o/p
$b_0 \rightarrow$ bias term
$b_1 \rightarrow$ regression coeffici

NOTE Logistic regression's dependent variable follows Bernoulli distribution.

Types of Logistic regression

1. Binary 2. Multinomial 3. Ordinal

Advantages of Logistic Regression

1. Easier to implement than other ML models.
2. Suitable for linearly seperable datasets
3. Provides valuable insights.

Logistic Regression Assumptions

1. The dependent/response variable is binary or dichotomous.
2. Little or no multicollinearity b/w predictor variables.
3. Linear relationship of independent variables to log odds.
4. Requires sufficiently large sample size.
5. No extreme outliers
6. Should have independent observations.

NOTE Log odds refer to the ways of expressing probabilities. However, log odds are different from probabilities. Odds refer to the ratio of ~~succes~~ success to failure, while probability refers to the ratio of success to everything that can occ

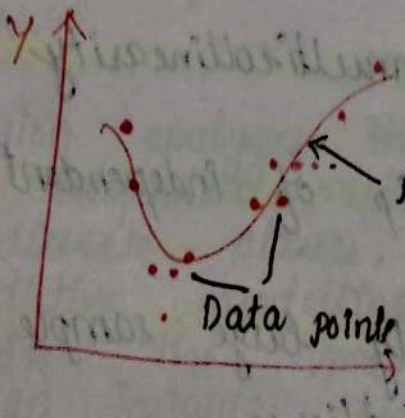
Polynomial Regression

In polynomial regression, we describe the relationship b/w independent variable x and dependent variable y using an n^{th} degree polynomial in x.

As we know that a simple linear regression algorithm only works when relationship b/w data is linear. This means that if rel'ship b/w data is non-linear, ~~regress~~ simple regression analysis would fail.

Hence, to overcome this problem, polynomial regression is introduced.

Polynomial regression helps to identify curvilinear rel^{n}ship b/w independent & dependent variable.



Polynomial regression eq^n

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \dots + b_n x_1^n$$

NOTE

Polynomial Regression predicts the best-fit line that matches the pattern of the data (curve)

Types of Polynomial Regression

1. Linear – degree 1
2. Quadratic – degree 2
3. Cubic – degree 3

(it goes on, on the basis of degree)

Advantages of Polynomial regression

1. It can accomodate a wide range of functions.

Practical Application of Polynomial regression

1. Used to figure out what isotopes are tnt in sediments.
2. Utilized to look at spread of various illnesses across a population.
3. Research on creation of synthesis.

Polynomial Regression Assumptions.

1. The behaviour of a dependent variable can be explained by a linear or curvilinear, additive rel"ship b/w the dependent variable & set of k independent variables.

2. The independent variables lack any interrelationship.

3. The errors are independent, normally distributed with mean zero & a constant variance.

Stepwise Regression

Stepwise regression is method of fitting a regression model by iteratively adding or removing independent variables, and testing for statistical significance after each iteration.

It is used when there are multiple independent variables. A special feature of stepwise regression is that the independent variables are chosen automatically, without intervention.

Statistical values like R-square and t-stats are used to identify the right independent variables.

Approaches of Stepwise Regression

1. Forward selection: begins with no variables in the model, tests each variable as it is added to the model, then keeps those that are deemed most statistically significant. The process is repeated until optimal results are obtained.
2. Backward elimination starts with a set of independent variables, deleting one at a time, then testing to see if the removed variable is statistically significant.
3. Bidirectional elimination is a combination of first 2 methods that tests which variable should be included or excluded.

Features of Stepwise Regression

1. Independent variables are chosen automatically.
2. It is used when data has high dimensionality. This is because its goal is to maximize the prediction ability of model with minimum no. of variables.

Ridge Regression

Ridge regression is a technique which is used when the data suffers from multicollinearity.

When data suffers from multicollinearity, even though least-squares are unbiased, their variances are large, this results in predicted values being far away from the actual values. By adding a degree of bias to the regression estimates, ridge regression reduces standard errors.

This regression technique performs L2 regularization, hence it is aka L2 Regression.

Assumptions of Ridge Regression

The assumptions of ridge regression are same as that of linear regression i.e.

i. Linearity of residuals
ii. Independence of residuals
iii. Constant variance.

However, as ridge regression does not provide confidence limits, so normality of errors need not to be assumed.

Features of Ridge Regression

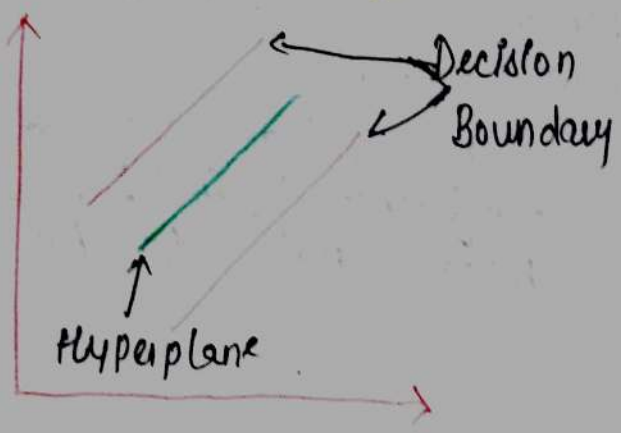
1. It shrinks the parameters, ∴ mostly used to prevent multicollinearity
2. It reduces model complexity by coefficient shrinkage
3. Use L2 regularization
4. It prevents overfitting

○ Support Vector Regression

SVR is a type of ML algo. used for regression analysis. The goal of SVR is to find a func" that approximates the relationship b/w the i/p variables & a continuous target variable, while minimizing the prediction errors.

Unlike Support Vector Machines (SVM) used for classification tasks, SVR seeks to find a hyperplane that best fits the data points in a continuous space. This is achieved by mapping the i/p variables to a high dimensional feature space while also minimizing the prediction errors.

Our best fit line is the hyperplane that has the max. number of points.



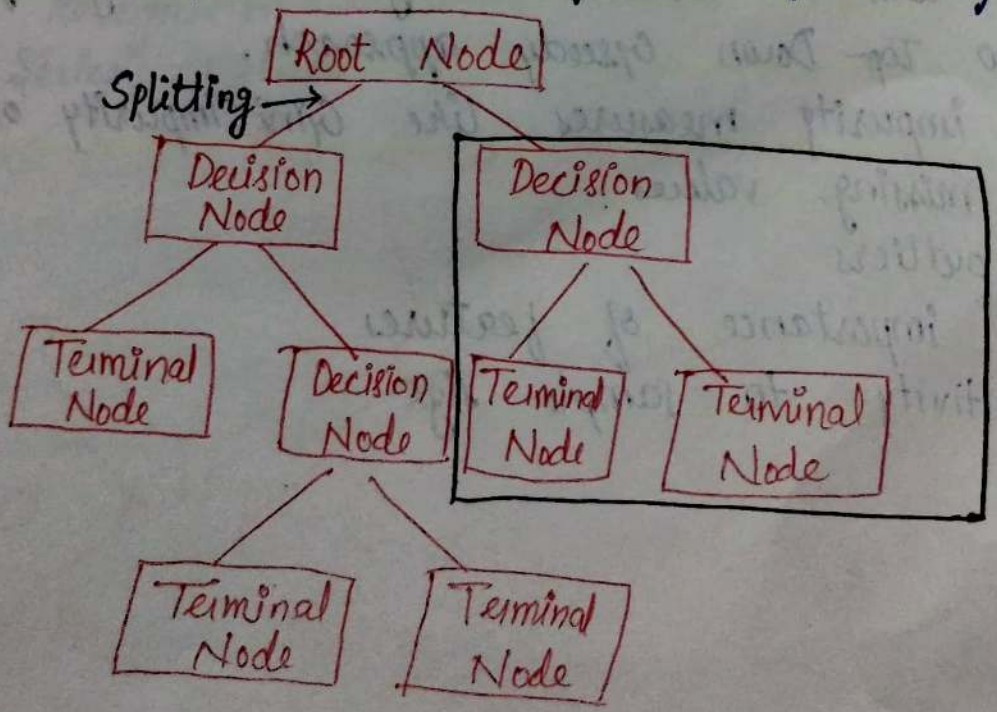
Features of SVR

1. It is particularly used for regression tasks where linear regression may not be sufficient due to complex relationships or non-linear data.
2. It is an extension of the SVM
3. It can handle both linear & non-linear data.

0 Decision Tree Regression

A decision tree is a non-parametric supervised learning algo. for classification & regression tasks. It has a hierarchial tree structure consisting of a root node, branches, internal nodes & leaf nodes.

It is used in decision support that depicts decisions & their potential outcomes, incorporating chance events, resource expenses & utility.



Structure of a Decision Tree

Decision trees are a bunch of if-else statement in layman terms.

Decision Tree Regression observes features of an obj. and trains the model in the structure of a tree to predict data in the future to produce meaningful continuous o/p.

How decision tree algo. works?

1. Starting at the root
2. Asking the best questions
3. Branching out
4. Repeat the process.

Decision Tree Assumptions

1. Decision Tree makes binary splits.
2. They use a recursive partitioning process.
3. Feature Independence
4. They aim to create homogeneous subgroups.
5. Follow Top-Down Greedy approach.
6. Use impurity measures like Gini Impurity or Entropy.
7. No missing values
8. No outliers
9. Equal importance of features
10. Sensitivity to sample size.

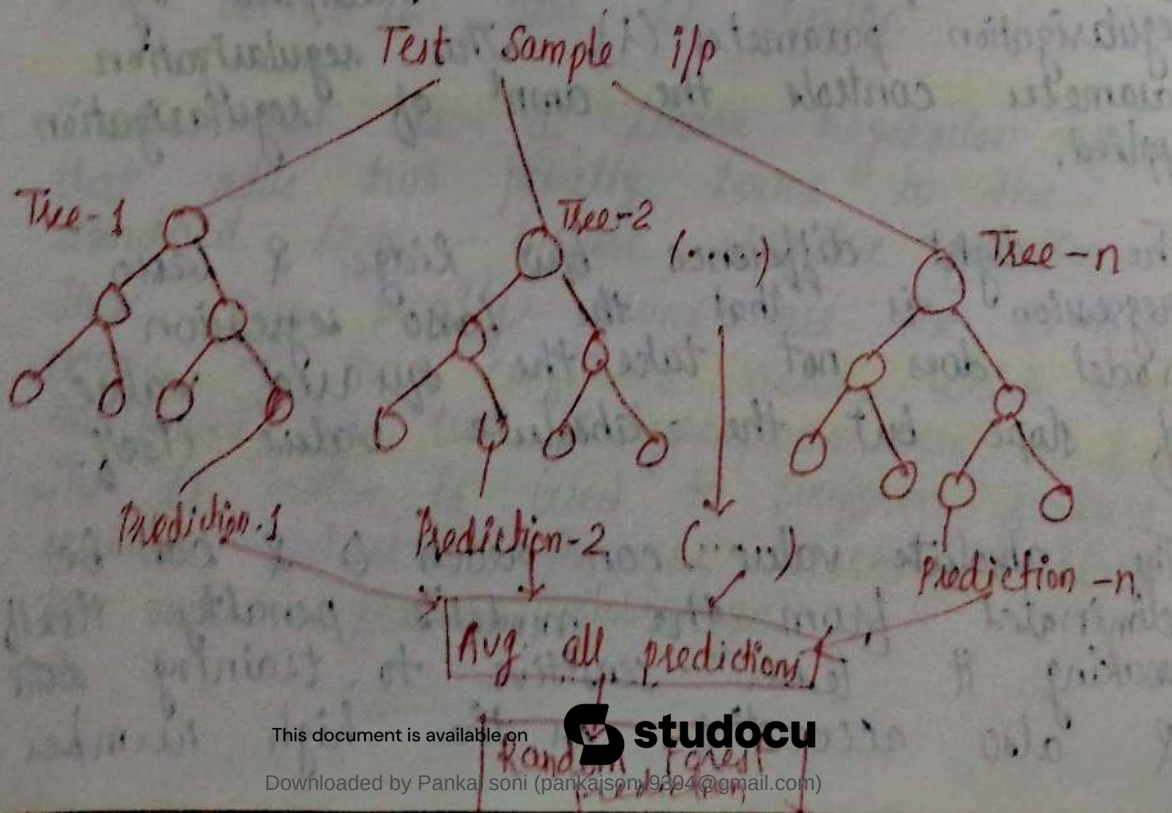
Random Forest Regression

→ ML algo. developed by Leo Breiman & Adele Cutler
→ combines o/p of multiple decision trees to reach a single result.
→ handles both classification & regression tasks.
→ can handle complex datasets & mitigate overfitting.

→ **Most imp feature:**
It can handle dataset containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification.

Random Forest Applications

1. Customer Churn Prediction
2. Fraud Detection
3. Stock Price Prediction
4. Medical Diagnosis
5. Image Recognition
6. Time-Series Analysis



Lasso Regression

Lasso Regression is an acronym in the term LASSO which stands for-

L: Least
A: Absolute
S: Shrinkage and
S: Selection
O: Operation

Lasso Regression is a slight modification of Ridge regression.

It is a regularization technique that applies a penalty to prevent overfitting and enhance the accuracy of statistical models. It uses L1 regularization, ∴ it is aka ~~L1~~ L1 regression.

Lasso regression is frequently used in ML to handle high dimensional data as it facilitates automatic feature selection. It does this by adding a penalty term to residual sum of squares (RSS), which is then multiplied by regularization parameter (λ). This regularization parameter controls the amnt of regularization applied.

The slight difference b/w Ridge & Lasso regression is that the Lasso regression model does not take the squared value of slope but the absolute value itself.

The absolute value can reach 0 & can be eliminated from the model's penalty itself making it lesser sensitive to training data & also accounting for the high number

of variables in your data & using only those which are actually helpful.

~~Lasso regression is one of the~~
The rest is same as in Ridge regression
— The higher the λ, the higher the penalty.

Lasso regression is good when you want to penalize your data to not overfit the training data & also if you want to exclude some unnecessary variables from your model.

o Elastic Net Regression

Elastic Net Regression is a powerful ML that combines the features of both Lasso and Ridge regression. Ridge utilizes L2 penalty and Lasso uses an L1 penalty. With elastic net, you don't have to choose between these two models, because elastic net uses both L2 and L1 penalty.

Elastic net is a linear regression algo. that adds two penalty terms to the standard least - square objective function. These two penalty terms are the L1 & L2 norms of the coefficient vector, which are multiplied by two hyperparameters, α & λ. The L1 norm is used to perform feature selection whereas the L2 norm is used to perform feature shrinkage.

Advantages of Elastic Net Regression

1. Feature selection results in a model with fewer variables which are easier to interpret & less prone to overfitting
2. More robust than other linear regression techniques such as Ridge & Lasso because it combines strength of both techniques.
3. Better performance

Applications

1. Bioinformatics
2. Finance
3. Marketing
4. Image Processing