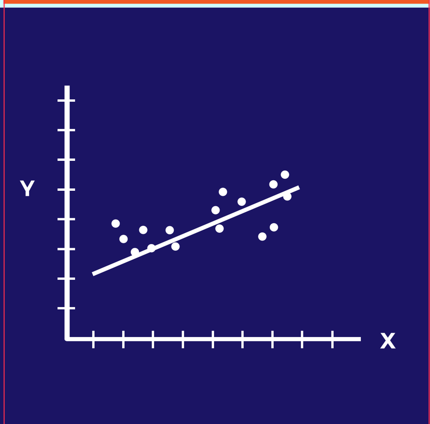
Machine Learning Algorithms Cheatsheet

Supervised Learning

Linear Regression



Equation

The linear regression equation is written as:

$$y = b_0 + b_{1x1} + b_{2x2} + ... + b_{n*xn}$$

where

y is the dependent variable, b_0 is the intercept, b_1, b_2, ..., b_n are the coefficients, and x1, x2, ..., xn are the independent variables.

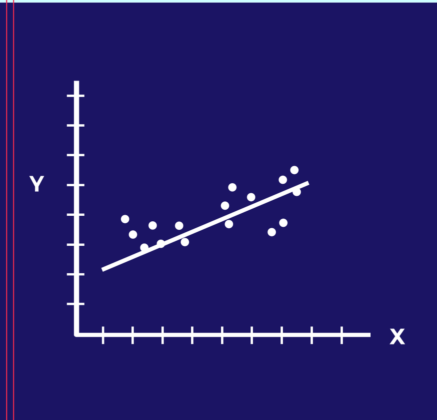
Concepts

- Residuals
- Ordinary Least Squares (OLS)
- Overfitting and Underfitting
- multicollinearity
- homoscedasticity
- Mean squared error (MSE)
- root mean squared error (RMSE)
- mean absolute error (MAE)
- R-squared value

Scenarios

- It is used to identify the linear relationship between the dependent variable and the independent variables.
- Mostly used in cases where data has minimal noise and outliers.
- It is suitable for predicting continuous numeric values.
- It can serve as a baseline model for comparison with more complex algorithms.
- It can help identify the most influential features in the prediction task.

Lasso Regression



Equation

The equation for Lasso Regression can be represented as follows:

$$\text{minimize: RSS} + \lambda * \Sigma|\beta|$$

where

RSS represents the residual sum of squares, which measures the difference between the predicted values and the actual target values.

λ is the regularization parameter that controls the strength of the penalty term.

Σ|β| is the sum of the absolute values of the regression coefficients β, which serves as the penalty term to encourage sparsity and shrinkage of less important variables.

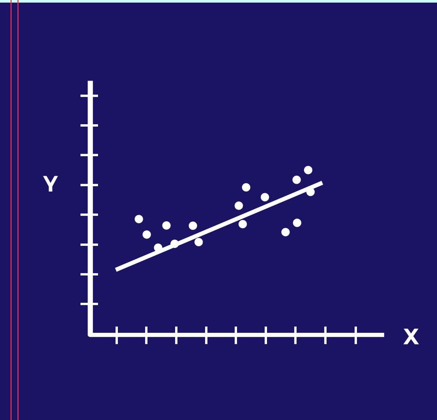
Concepts

- Linear regression
- Regularization
- Residual sum of squares (RSS)
- Regression coefficients
- Penalty term
- Feature selection
- Shrinkage
- Sparsity

Scenarios

It is commonly used in scenarios where there are a large number of features, and the goal is to identify the most relevant variables while simultaneously controlling model complexity and preventing overfitting.

Ridge Regression



Equation

The equation for Ridge Regression can be summarized as:

$$\text{minimize: RSS} + \lambda * \Sigma(\beta^2)$$

where

RSS represents the residual sum of squares, which measures the difference between the predicted values and the actual target values.

λ is the regularization parameter that controls the strength of the penalty term.

Σ(β2) is the sum of the squared regression coefficients β, which serves as the penalty term to control the magnitude of the coefficients and reduce the impact of multicollinearity.

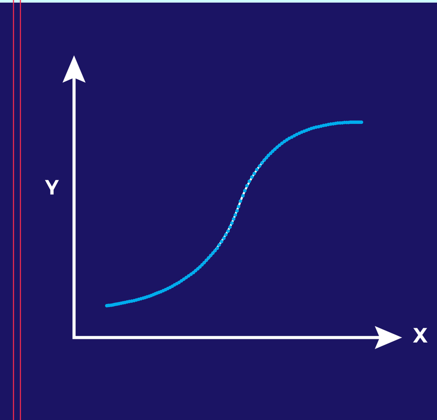
Concepts

- Linear regression
- Regularization
- Residual sum of squares (RSS)
- Regression coefficients
- Penalty term
- Multicollinearity
- Model complexity
- Shrinkage

Scenarios

It is commonly used when dealing with high-dimensional data and multicollinearity issues, as it provides a balance between model fit and stability by shrinking the coefficients towards zero without eliminating any variables entirely.

Logistic Regression



Equation

$$p = 1 / (1 + e^{(-z)})$$

where

p represents the probability of the binary outcome (e.g., the probability of belonging to class 1).
z is the linear combination of the independent variables and their corresponding coefficients.
e is the base of the natural logarithm.

The linear combination (z) is calculated as:

$$z = b_0 + b_{1x1} + b_{2x2} + ... + b_{n*xn}$$

where
b0 is the intercept term (the coefficient for the constant term).
b1, b2, ..., bn are the coefficients f or the independent variables.
x1, x2, ..., xn are the values of the independent variables.

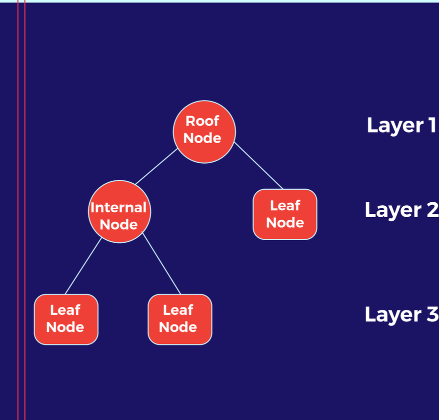
Concepts

- Decision boundary
- threshold value
- Maximum Likelihood Estimation
- Precision
- Recall
- F1Score
- AUC-ROC curve
- Logistic Function
- Log Odds and Logit Function
- Binary Classification

Scenarios

- It helps in understanding the impact of predictors on the outcome.
- It can be used for determining the relative importance of features.
- It performs well with datasets that have minimal noise and outliers.
- It can handle medium to large datasets efficiently
- It is commonly used in research studies and exploratory data analysis.
- It is well-suited for binary classification tasks

Decision Trees



Equation

Decision trees do not have a specific mathematical equation like linear regression or logistic regression. Instead, they make decisions based on a hierarchical structure of nodes and branching conditions.

Concepts

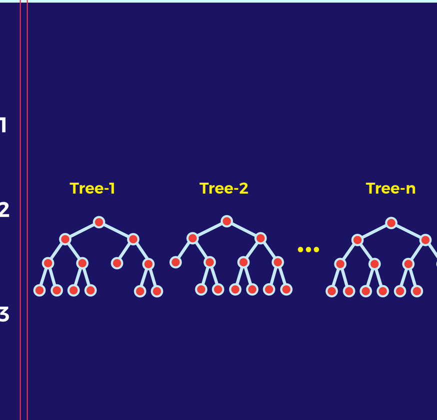
Splitting Criteria like

- Gini impurity and entropy
- Information Gain, Pruning
- Leaf Nodes
- Ensemble Methods

Scenarios

- It is well suited for classification tasks
- It is suited for data where the relationships between the input variables and the target variable are nonlinear or involve complex interactions.
- It can be used for feature selection: By examining the splits and hierarchy of the tree, less important features can be identified
- It can offer interpretability due to their hierarchical structure and clear splitting conditions
- It can handle missing data effectively.
- It can detect outliers by considering instances that end up in separate leaf nodes or exhibit unusual decision paths

Random Forest



Equation

The random forest algorithm does not have a specific equation. Instead, it is a combination of decision trees.

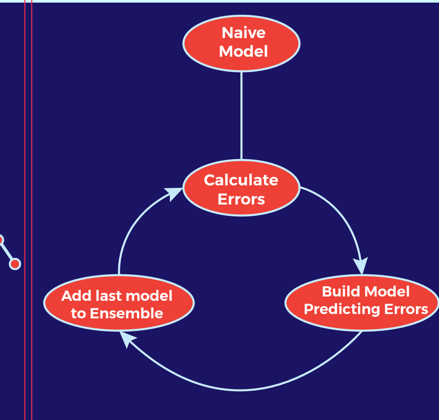
Concepts

- Ensemble Learning
- Decision Trees
- Bagging
- Out of Bag Error
- Random Feature Selection

Scenarios

- It can be used for classification and regression tasks.
- It can handle high-dimensional data effectively.
- Random Forests can be used for outlier detection by analyzing the anomaly score of instances
- It can handle imbalanced datasets by adjusting class weights or using sampling techniques such as oversampling or undersampling.
- It can capture non-linear relationships between the predictors and the target variable.
- It can serve as a benchmark model for comparison against other complex algorithms.

XGBoost



Equation

While the exact mathematical equation for XGBoost is complex and involves numerous terms and parameters, the underlying principle revolves around iteratively improving the ensemble of decision trees by minimizing a combined loss and regularization function.

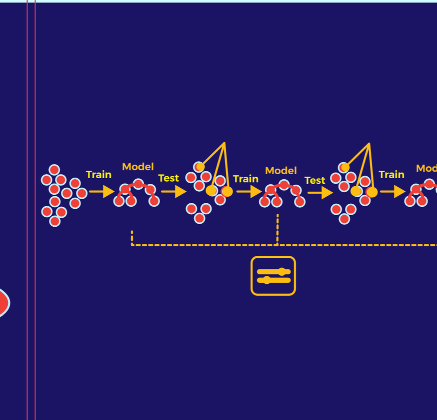
Concepts

- Gradient Boosting
- Decision Trees
- different loss functions (e.g., logistic loss, squared loss) and regularization terms (e.g., L1 regularization, L2 regularization)
- techniques like pre-sorting, column blockings, and approximate algorithms
- cross-validation

Scenarios

- It can be used to solve classification and regression problems.
- It can be used for building ranking models and recommendation systems.
- XGBoost can be used for detecting anomalies in various domains.
- It can be used for time series forecasting.

Gradient Boosting Regression



Equation

The equation for gradient boosting regression can be summarized as:

Prediction = Initial Prediction + Learning Rate * Sum of Weak Learner Predictions

where the weak learner predictions are iteratively added to the initial prediction, multiplied by the learning rate, to improve the overall prediction accuracy.

Concepts

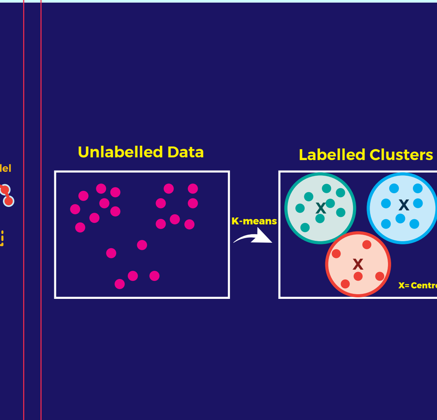
- Gradient boosting algorithm
- Weak learners (e.g., decision trees)
- Residuals and error minimization,
- Learning rate and shrinkage
- Gradient descent optimization
- Ensembling of weak learners
- Overfitting prevention techniques (e.g., regularization)
- Hyperparameter tuning
- Loss functions
- Feature importance analysis

Scenarios

- It is commonly used for predictive modeling tasks such as housing price prediction, demand forecasting, and customer churn prediction
- It can be applied in financial analysis for tasks such as stock price prediction, credit risk assessment
- It can be used to build recommendation systems.

Unsupervised Learning

K-Means Clustering



Equation

The equation for K-Means algorithm can be summarized as:

$$\arg\min(C)\ \Sigma\ ||x - \mu\_i||^2$$

where

- C represents the assignment of data points to clusters.
- x denotes the data points.
- μ_i denotes the centroids of the clusters.
- ||x - μ_i||^2 represents the squared Euclidean distance between a data point and its assigned cluster centroid. The objective of the K-Means algorithm is to minimize the sum of squared distances between data points and their corresponding cluster centroids by iteratively updating the cluster assignments and recalculating the centroids.

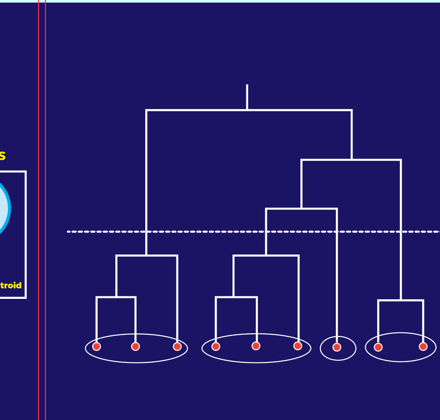
Concepts

- Centroids
- Euclidean distance
- Cluster assignment
- Cluster centroid update
- Sum of squared distances
- Initialization methods
- Convergence criteria

Scenarios

- Customer Segmentation
- Document Clustering
- Anomaly Detection
- Market Segmentation
- Fraud Detection
- Image Segmentation

Hierarchical Clustering



Equation

There are two types of approaches to hierarchical clustering:

1. Agglomerative Hierarchical Clustering: Start with each data point as an individual cluster, then iteratively merge the two closest clusters until a single cluster remains.
2. Divisive Hierarchical Clustering: Start with all data points in a single cluster, then iteratively split the cluster into two based on the dissimilarity between data points, creating a hierarchy of clusters.

In both cases, the choice of proximity or dissimilarity metric, such as Euclidean distance or correlation coefficient, determines the merging or splitting of clusters. The result is a dendrogram that represents the hierarchical structure of the data points and their relationships.

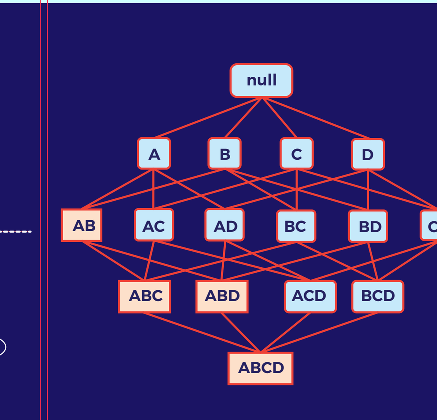
Concepts

- Proximity or dissimilarity measures (e.g., Euclidean distance, correlation coefficient)
- Agglomerative and divisive clustering
- Dendrogram
- Linkage methods (e.g., single linkage, complete linkage, average linkage)
- Hierarchical structure
- Merge and split criteria
- Distance matrix
- Cluster distance or dissimilarity calculation
- Cut-off or threshold determination for cluster formation.

Scenarios

- Customer Segmentation
- Crime Pattern Analysis
- Clustering Documents
- Gene Expression Analysis
- Social Network Analysis

Apriori Algorithm



Equation

The Apriori algorithm does not have a specific equation, as it operates based on a set of steps rather than a mathematical formula.

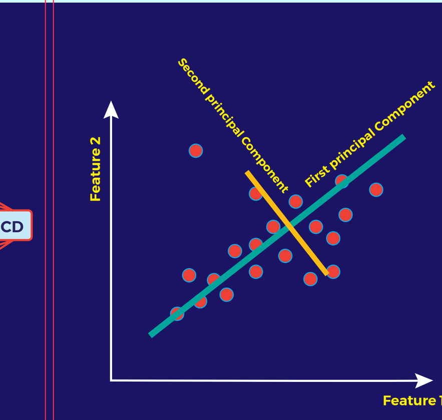
Concepts

- Association rule mining
- Frequent itemsets
- Support and confidence measures
- Minimum support threshold
- Candidate itemset generation
- Pruning strategies
- Apriori property
- Apriori principle
- Iterative search and counting mechanisms.

Scenarios

- Market basket analysis
- Customer behavior analysis
- Recommender systems
- Web usage mining
- Cross-selling and upselling
- Text mining and document analysis.

Principal Component Analysis



Equation

The equation for Principal Component Analysis (PCA) involves computing eigenvectors and eigenvalues of the covariance matrix. The transformation is represented as Y = X * P, where Y is the reduced-dimensional representation of the data matrix X obtained by multiplying it with the projection matrix P.

Concepts

- Covariance Matrix
- Eigenvectors and eigenvalues
- Dimensionality reduction
- Variance explained
- Principal components
- Orthogonality
- Projection matrix
- Explained variance ratio
- Data standardization

Scenarios

- Dimensionality reduction
- Feature extraction
- Data visualization
- Noise reduction
- Image and video processing
- Signal processing Collaborative filtering.

ProjectPro