Cleaning and Transfrom data: clean data first to ensure accuracy, consistency, and reliability by removing errors, handling missing values, and standardizing formats, which helps improve analysis and model performance. – Type and name conversion. – Filtering of missing or inconsistent data. – Unifying semantic data representations. – Matching entries from different sources. – Rescaling and optional dimensionality reduction.
Bar Chart A bar chart is used to compare categorical or ordinal data. It displays frequencies across distinct categories, making it easy to identify the highest or lowest values.
Histogram A histogram is used to visualize the distribution of continuous numerical data. It groups data into intervals (bins) to show patterns such as spread, skewness, and outliers.
Scatter Plot A scatter plot shows the relationship between two continuous variables. It helps identify correlations, trends, or clusters by plotting individual data points on a two-dimensional axis.
Pie Chart A pie chart is used to represent proportions of a whole for categorical data. It visually compares parts to the whole using slices of a circle.
Pivot Table A pivot table summarizes large datasets by grouping and aggregating data. It is useful for organizing, sorting, and calculating key metrics efficiently.
Python is great for prototyping – Interpreted: direct execution without compilation. – Dynamically-typed: don't have to declare a static type. – Readable: easy-to-understand syntax. – Deployable: easy to incorporate in applications.
Feature Series DataFrame 1. Structure One-dimensional Two-dimensional 2. Data Type Holds a single data type Can hold multiple data types across columns 3. Indexing Single index Row and column indexing 4. Columns Has only one column (implicit) Has multiple named columns 5. Usage Used for a single column or array of data Used for structured data with rows and columns
Box plots summarise data based on 5 numbers: – Lower inner fence – Q1–1.5*IQR. – First quartile (Q1) – equivalent to 25th percentile. – Median (Q2) – equivalent to 50th percentile. – Third quartile (Q3) – equivalent to 75th percentile. - Upper inner fence – Q3+1.5*IQR. – Values outside fences are outliers. – Sometimes include outer fences at Q1-3*IQR and Q3+3*IQR.
Pearson Correlation Pearson correlation measures the linear relationship between two continuous variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear correlation. It assumes that the variables are normally distributed and related linearly. Pearson is sensitive to outliers and is best used when both variables are interval or ratio scale.
Spearman Correlation Spearman correlation measures the monotonic relationship between two variables using the ranked values rather than raw data. It also ranges from -1 to 1, where the values indicate the strength and direction of the association. It is non-parametric and does not assume a normal distribution, making it useful when the relationship is not linear or the data contains outliers.
Kendall Correlation Kendall correlation (also known as Kendall's Tau) measures the strength of dependence between two variables based on the concordance and discordance of data pairs. It is also non-parametric and works with ordinal data. Kendall is generally more robust for smaller datasets and is more accurate when data have many tied ranks or when the relationship is weak and non-linear.
Relational Database A relational database is a type of database that stores data in tables (relations) consisting of rows and columns. It uses structured query language (SQL) to manage and manipulate data with relationships maintained through keys.
Relational Database Schema A relational database schema defines the structure of the database, including tables, fields, data types, and relationships. It serves as a design plan or blueprint for how data is organized and connected.
Relational Database Instance A relational database instance is the current state or snapshot of the data in the database. It includes the actual data stored in the tables at a specific moment in time, based on the schema.
tables qualify as a relation: – Every relation must have a unique name. – Attributes (columns) in tables must have unique names. • The order of the columns is irrelevant. – All tuples in a relation have the same structure • Constructed from the same set of attributes. – Every attribute value is atomic (not multivalued, not composite). – Every row is unique. • Can't have two rows with exactly the same values for all their fields. – The order of the rows is immaterial...
DDL (Data Definition Language) DDL commands are used to define or modify the structure of database objects such as tables, schemas, and indexes. Common DDL commands include CREATE, ALTER, DROP, and TRUNCATE....
DML (Data Manipulation Language) DML commands are used to insert, update, delete, or retrieve data from existing tables. Common DML commands include INSERT, UPDATE, DELETE, and SELECT.... Domain constraints restrict the type, format, or range of values that can be stored in a column to maintain data integrity. Common domain constraints include:
NULL: Allows a column to have missing or unknown values.
NOT NULL: Ensures that a column must always have a value (no empty entries allowed).
CHECK: Enforces a condition on the values in a column (e.g., age ¿ 0).
DEFAULT: Assigns a default value to a column if no value is provided during data entry.
A primary key is a unique identifier for each record in a table. It must be unique and NOT NULL, ensuring that no two rows have the same value in the primary key column(s).
A foreign key is a column in one table that references the primary key of another table. It is used to maintain referential integrity between related tables.
A candidate key is any column or set of columns that can uniquely identify a row in a table. A table can have multiple candidate keys, but only one is chosen as the primary key.
Data modeling is the process of designing the structure of a database by defining entities, attributes, relationships, and constraints. It helps in organizing data logically to ensure accuracy, consistency, and efficient access.
Feature OLTP (Online Transaction Processing) OLAP (Online Analytical Processing) Purpose Manages day-to-day transactions Supports data analysis and decision making Data Type Handles current, real-time data Uses historical, consolidated data Operations Frequent insert, update, delete actions Mainly read and aggregate queries Normalization Highly normalized for efficiency Often denormalized for faster querying
A data warehouse is a centralized repository that stores large volumes of integrated, historical data from multiple sources to support business intelligence and analytical reporting. It is optimized for querying and analysis, not for transaction processing.
Characteristics Subject-Oriented: Organized around key subjects like sales, finance, or customer data.
Integrated: Combines data from different sources into a consistent format.
Time-Variant: Stores historical data to track changes over time.
Non-Volatile: Data is stable and not frequently updated or deleted, only added for analysis. Feature Star Schema Snowflake Schema Structure Uses denormalized dimension tables directly linked to the fact table. Uses normalized dimension tables split into multiple related tables. Complexity Simpler design with fewer tables and joins. More complex design with more tables and joins. Storage Requires more storage due to data redundancy. More storage-efficient by eliminating redundancy. Query Performance Faster queries because of fewer joins. Slower queries due to multiple joins. Maintenance Easier to maintain because of its straightforward structure. More difficult to maintain due to normalized tables. The LIKE operator used to search for patterns in text using wildcards Feature Observational Study Experimental Study Control No control over variables; only observation Variables are actively manipulated Purpose Identify correlations and patterns Establish cause-and-effect relationships Intervention No intervention Involves intervention or treatment Data Collection Passive, records natural behavior or outcomes Active, measures outcomes after manipulation Causality Cannot prove causation Can demonstrate causation under controlled conditions Hypothesis is an assumption or statement that we want to test.
Research Question asks whether the independent variable has any effect.... Null Hypothesis (H) is the assumption that there is no effect or difference.... Alternative Hypothesis (H) is the opposite of the null hypothesis, suggesting there is an effect. P-value is the probability of observing the result (or more extreme) if the null hypothesis is true. If p-value ¡ alpha, there is strong evidence against H → reject the null hypothesis. If p-value ¿ alpha, there is weak evidence against H → accept the null hypothesis. If p-value = alpha, the evidence is marginal → no clear decision about H. Alpha Value () is the significance level (commonly 5Parametric Tests These tests rely on assumptions like normal distribution, equal variances, and interval/ratio level data. They are powerful and efficient if the data meets these conditions.
One-Sided t-test: Used when we want to test if the mean of a sample is either significantly greater than or less than a specific value, not both.
Unpaired t-test (Independent Samples t-test): Used to compare the means of two unrelated groups, like test scores of two different classes.
ANOVA (Analysis of Variance): Used to compare the means of three or more independent groups to see if at least one group differs significantly from the others.... Non-Parametric Tests These tests are used when data doesn't meet parametric assumptions (like normality). They use medians or ranks instead of means and work well with ordinal data or non-normal distributions. Mann-Whitney U Test: A non-parametric alternative to the unpaired t-test. It compares two independent groups by ranking all values and checking if their distributions differ. Kruskal-Wallis Test: A non-parametric alternative to ANOVA. It compares three or more independent groups based on ranks to check if at least one group is different. Confusion Matrix:TP-FN-FP-TN Accuracy = (TP + TN) / (TP + TN + FP + FN) → Proportion of correct predictions among all predictions.
Precision (Homogeneity) = TP / (TP + FP) → Of all predicted positives, how many were actually positive?
Recall (Completeness) = TP / (TP + FN) → Of all actual positives, how many were correctly predicted?
F1-Score (V-measure) = 2 × (Precision × Recall) / (Precision + Recall) → Harmonic mean of precision and recall; balances both. Evaluating Classifier Accuracy Holdout Method splits the dataset into two parts: one for training and one for testing, typically using a ratio like 70:30. It's simple but can give biased results if the split isn't representative. Cross-Validation divides the data into k equal parts (folds), trains on k–1 folds, tests on the remaining one, and repeats this process k times—providing a more reliable estimate of model performance. Machine Learning Machine Learning is a branch of artificial intelligence that enables systems to learn from data and improve performance without being explicitly programmed. It uses algorithms to identify patterns, make decisions, and predict outcomes based on input data.
Common Problems in Machine Learning Overfitting – Model performs well on training data but poorly on new/unseen data.
Underfitting – Model is too simple and fails to capture underlying patterns in the data.
Imbalanced Data – One class dominates the dataset, leading to biased predictions.
High Dimensionality – Too many features can lead to increased complexity and reduced performance.
Supervised vs Unsupervised Learning Feature Supervised Learning Unsupervised Learning Labeled Data Uses labeled data (input-output pairs) Uses unlabeled data Goal Predict output/target based on input Discover hidden patterns or groupings Examples Classification, Regression Clustering, Dimensionality Reduction Evaluation Accuracy, precision, recall, F1-score Silhouette score, clustering quality
Association Rule Mining is a data mining technique used to discover interesting relationships, patterns, or associations among items in large datasets, especially in transactional databases. Items: Individual elements or products in a transaction dataset.
Itemset: A collection of one or more items (e.g., milk, bread).
k-itemset: An itemset that contains exactly k items.
Support Count: The number of transactions that contain a particular itemset.
Support: Proportion of transactions that include the itemset; Support = (Support Count) / (Total Transactions).
Frequent Itemset: An itemset whose support is greater than or equal to a predefined minimum support threshold.
Association Rule: An implication of the form X Y, meaning if X occurs, Y is likely to occur.
Confidence: The probability that Y occurs given X has occurred; Confidence = Support(X Y) / Support(X).
Apriori is an algorithm used to find frequent itemsets and generate association rules from transaction data.
Key Principles Apriori Property: All subsets of a frequent itemset must also be frequent.... Anti-Monotonicity: If an itemset is infrequent, its supersets are also infrequent.... Candidate Generation: Combine frequent k-itemsets to form (k+1)-itemsets.
Pruning: Remove candidates with any infrequent subset to save time.
FP-Growth (Frequent Pattern Growth) is a fast and efficient algorithm that mines frequent itemsets without candidate generation by using a compact data structure called an FP-tree. It reduces database scans and improves performance over Apriori, especially on large datasets.Clustering is an unsupervised learning technique that groups similar data points together based on distance or similarity without using labeled data. It helps in discovering hidden patterns or natural groupings in datasets. Minkowski Distance: D=(—x1y1—^{p} + $|x2y2|^p + +|xnyn|^p)1/p$
Manhattan Distance (p=1): D=—x1y1—+—x2y2—++—xnyn—
Euclidean Distance (p=2): D=$(x1y1)^2 + (x2y2)^2 + +(xnyn)^2$
Hierarchical Clustering: Builds a tree-like structure (dendrogram) by either merging (agglomerative) or splitting (divisive) clusters based on distance.
Partitioning Clustering: Divides data into a fixed number of clusters by optimizing a criterion, like k-means minimizing within-cluster variance.K-means is a popular partitioning method that assigns points to the nearest cluster centroid and updates centroids iteratively. Its time complexity is O(n×k×i×d), where n = number of points, k = clusters, i = iterations, and d = dimensions.Agglomerative (Bottom-Up): Starts with each data point as its own cluster and iteratively merges the closest clusters until all points form a single cluster.
Divisive (Top-Down): Starts with all data points in one cluster and recursively splits clusters into smaller ones until each point is separate.Cluster validity measures assess the quality of clustering results and are classified into three types:
Internal Measures: Evaluate clusters based on the data, focusing on compactness and separation. An example is the Silhouette Coefficient (S), which for each point uses: a = average distance to points in the same cluster b = minimum average distance to points in the nearest different cluster The Silhouette Coefficient S is: S=1a/b, if a¡b; b/a1, if a¿b
External Measures: Compare clustering results to known labels using metrics like Precision, Recall, and F1 Score.
Relative Measures: Compare different clustering results on the same dataset to select the best model without ground truth.PCA is a dimensionality reduction technique that transforms correlated variables into a smaller set of uncorrelated variables called principal components. It simplifies data by capturing the most important variance, making analysis, visualization, and modeling easier and faster.
To do this, PCA first computes the covariance matrix, which measures how variables vary together—positive covariance means variables increase or decrease together, while negative covariance means they move inversely. This matrix reveals relationships between variables and helps identify directions of maximum variance.
From the covariance matrix, PCA calculates eigenvectors and eigenvalues:
Eigenvectors represent the directions (principal components) along which the data varies most.
Eigenvalues measure the amount of variance captured by each eigenvector.
PCA selects the principal components with the highest eigenvalues to reduce data dimensionality while preserving essential information.
Uses of PCA include:
Reducing feature space to simplify models and prevent overfitting.
Speeding up machine learning algorithms by lowering data complexity.
Visualizing high-dimensional data in 2D or 3D plots.
Removing noise and redundancy by focusing on key variance directions.
Enabling data compression for efficient storage.Simple Linear Regression Models the linear relationship between x and y with an error term.... y=+x+
Error (Residual) Difference between observed and predicted values. ei=yiyˆi
Sum of Squared Errors (SSE) Total squared difference between observed and predicted values; lower is better. SSE=$(yiy\hat{i})^2$
Ordinary Least Squares (OLS) Estimates Calculates slope and intercept by minimizing SSE. $\hat{}=(xix)(yiy)/(xix)^2, \hat{}=y\hat{}x$
Total Sum of Squares (SST) Total variance in the observed data; generally higher indicates more variability to explain. SST=(yiy)$^2...RegressionSumofSquares(SSR)Varianceexplainedbythemodel;higherisbetter.SSR = (\hat{y}iy)^2...CoefficientofDetermination(R^2)Proportionofvarianceexplainedbythemodel;rangesfrom0to1(higherisbetter).IfR^2isnegative,itmeansthemodelfitsworsethanahorizontalline(meanofy).R^2 = SSR/SST = 1SSE/SST$
Multiple Linear Regression Models the relationship between multiple independent variables and a dependent variable: y=+1x1+2x2++pxp+
Cost Function (Mean Squared Error) Measures how well the model predicts by calculating average squared error: J()=1/2ni=1n(h(x(i))y(i))$^2 whereh(x(i))isthepredictedvalue,y(i)istheactualvalue,andnisthenumberofsamples.$
Polynomial Transformation Extends linear regression by including powers of features to capture nonlinear relationships: y=+1x+2x^2 + 3x^3 + +dx^d+
Overfitting Occurs when a model learns the training data—including noise and outliers—too well, causing it to perform very accurately on training data but poorly on unseen data. This happens because the model is too complex and captures random fluctuations instead of the true pattern.
Underfitting Happens when a model is too simple to capture the underlying structure of the data, resulting in poor performance on both training and new data. The model fails to learn the important patterns, leading to high bias.
Correct Fit A model that generalizes well by capturing the true underlying patterns without being influenced by noise. It balances bias and variance, performing well on both training and unseen data.Regularization Regularization adds a penalty to the cost function to reduce model complexity and prevent overfitting by shrinking coefficients. L1 Regularization (Lasso): Adds the sum of absolute values of coefficients to the cost function, promoting sparsity by driving some coefficients to zero.
L2 Regularization (Ridge): Adds the sum of squared coefficients to the cost function, encouraging smaller coefficients but rarely making them exactly zero.L2 Regularization (Ridge):
J()=1/2ni=1n(h(x(i))y(i))$^2 + j = 1pj^2$
L1 Regularization (Lasso):
J()=1/2ni=1n(h(x(i))y(i))$^2 + j = 1p|j|$
controls the penalty strength.
Gradient Descent Gradient Descent updates model parameters iteratively to minimize the cost function: j:=jJ/j
Convex function: Has one global minimum; its graph is a curve that never crosses itself.
Non-convex function: Has multiple minima; the curve may cross itself or have local minima.
Gradient Descent always converges to the global minimum for convex functions if the learning rate is small enough.
Learning Rate () A hyperparameter controlling the step size in each gradient descent iteration.
If is too small, gradient descent can be very slow.
If is too large, gradient descent might overshoot the minimum.
One good practice is gradually shrinking the learning rate during training to balance speed and stability. Batch Gradient Descent: Uses all training samples to compute the gradient and update parameters once per iteration.... :=(1/n)i=1nJ(;x(i),y(i))
Stochastic Gradient Descent (SGD): Updates parameters using the gradient from a single training sample at each iteration. :=J(;x(i),y(i))
Aspect Regression Classification Output Predicts continuous numeric values Predicts discrete class labels Example Task Predict house price, temperature Identify spam email, recognize digits Algorithm Type Models relationships between variables Assigns input to predefined categories Evaluation Metrics like MSE, RMSE, R² Metrics like accuracy, precision, recallLogistic Regression Sigmoid Function: The sigmoid function maps any real-valued input to a probability between 0 and 1: (z)=1/(1+e^z)$wherez =^T x.$
Hypothesis: The predicted probability that y=1 given input x is: h(x)=$(^T x)CostFunction(LogLoss/Cross - EntropyLoss) : Thecostfunctionmeasurestheerrorbetweenpredictedprobabilitiesandactuallabels : J() = (1/m)i = 1m[y(i)log(h(x(i))) + (1y(i))log(1h(x(i)))]...Fory = 1, costislog(h(x))Fory = 0, costislog(1h(x))Thiscostfunctionisconvex, whichensuresefficientoptimization....GradientDescentforLogisticRegression : Gradientdescentiterativelyupdatesparameterstominimizethecost : j := j(1/m)i = 1m(h(x(i))y(i))xj(i)whereisthelearningrateandjindexestheparameters.DecisionTree : Adecisiontreeisasupervisedlearningmodelthat$
Root Node: The topmost node of the tree representing the entire dataset where the first decision split occurs.
Branches: The connections between nodes that represent the outcomes of tests on features, leading to further splits or leaf nodes.
Leaf Node: The end nodes that provide the final decision or predicted class after all splits. Entropy Entropy measures the impurity or disorder in a dataset. It is defined as: H(S)=i=1c pi log2(pi) where pi is the proportion of class i in set S, and c is the number of classes. Lower entropy means purer nodes.
Information Gain (IG) Information Gain measures the reduction in entropy after a dataset is split on an attribute: IG(S,A)=H(S)vValues(A)—Sv—/—S— H(Sv) where Sv is the subset of S for which attribute A has value v.
Higher Information Gain is better because it means the attribute provides a more effective split, reducing uncertainty. $\max_depth : Limitsthemaximumdepthofthetreetocontroloverfitting.$
$\min_samples_split : Minimumnumberofsamplesrequiredtosplitaninternalnode.$
$\min_samples_leaf : Minimumnumberofsamplesrequiredtobeataleafnode.$
$\max_features : Numberoffeaturestoconsiderwhenlookingforthebestsplit.$
criterion: The function to measure the quality of a split (e.g., "gini" for Gini impurity or "entropy" for information gain).valuation Steps for Decision Trees Train-Test Split: Divide data into training and testing sets to evaluate model performance on unseen data.
Confusion Matrix: Analyze true positives, false positives, true negatives, and false negatives to assess classification accuracy.
Cross-Validation: Use k-fold cross-validation to validate model stability and generalization. Feature Engineering
This involves transforming raw data into meaningful input features that improve model learning. It includes creating new features, selecting the most relevant ones, and preprocessing data (e.g., scaling, encoding categorical variables).
Ensembles
Ensemble methods combine predictions from multiple models to reduce errors and increase robustness. For decision trees, popular ensembles include Random Forests (averaging many trees built on random subsets of data and features) and Gradient Boosting (building trees sequentially to correct previous errors).Aspect Structured Data Unstructured Data Format Organized in rows and columns (tables) No predefined format or organization Data Type Numeric, text, dates in fixed fields Text, images, videos, audio, emails, etc. Storage Stored in relational databases Stored in files, documents, multimedia Processing Easy to search, query, and analyze Requires advanced techniques like NLP or image recognition Examples Spreadsheets, SQL databases Emails, social media posts, videosFeature Vector from Text: It converts a text document into a list of numbers, where each number shows how many times a particular word appears. This helps machines understand and process the text numerically.
Tokenization: This breaks down text into smaller units called tokens, usually words or sentences, which makes it easier to analyze and process language data.
Normalization: It scales different data values to a common range or standard form so that no single feature dominates, improving the performance and speed of learning algorithms.
Indicator Features: These create binary vectors from text, marking each word with a 1 if it appears or 0 if it doesn't, counting each word only once, which helps convert categorical data into a numeric format for models.Term Frequency: Measures how often a word appears in a document, showing its importance within that text.
Damping: A technique used to reduce the effect of very frequent words by scaling down their counts to prevent them from dominating the feature vector.TF: log(occurrences of term in document) IDF: log(total number of documents / number of documents containing the term)
Naïve Bayes Classifier: A probabilistic machine learning model based on Bayes' Theorem, assuming features are independent given the class. It is commonly used for text classification (like spam detection).... Bayes' Theorem Formula: P(C—X)=P(X—C)P(C)/P(X) Where: P(C—X): Posterior probability of class C given input X P(X—C): Likelihood of input X given class C P(C): Prior probability of class C P(X): Probability of input X (can be ignored for classification since it's constant for all classes)... Naïve Assumption: It assumes features in X=(x1,x2,...,xn) are independent, so: P(X—C)=P(x1—C)P(x2—C)...P(xn—C)
Laplace Smoothing (Add-One Smoothing): Laplace Smoothing is used in Naïve Bayes to handle the problem of zero probabilities for unseen features by adding a small value (usually 1) to all counts.
Formula (for categorical features): P(xi—C)=(count(xi,C)+1)/(count(C)+k) Where: count(xi,C): Number of times feature xi appears in class C count(C): Total count of all features in class C k: Total number of unique features (vocabulary size)