Data Science Pocket Dictionary

IMPORTANT DATA SCIENCE TERMS

ramchandrapadwal

A

Accuracy

A metric that measures the correctness of a model's predictions, calculated as the number of correct predictions divided by the total number of predictions.

Activation Function

A mathematical function that introduces non-linearity in neural networks, allowing them to learn complex patterns.

A/B Testing

A statistical method used to compare two versions of a product or webpage to determine which performs better.

Anomaly Detection

The process of identifying patterns or data points that deviate significantly from the expected behavior in a dataset.

API (Application Programming Interface)

A set of protocols and tools that allows different software applications to communicate and interact with each other.

Artificial Intelligence (AI)

The simulation of human intelligence in machines that can perform tasks typically requiring human intelligence, such as learning, reasoning, and problem-solving.

Association Rule Mining

A technique in data mining that discovers interesting relationships or patterns within datasets.

Active Learning

A learning approach where the model actively selects which data points to be labeled by an oracle to improve performance.

Autoencoder

A type of neural network used for unsupervised learning that aims to reconstruct its input, forcing the model to learn meaningful representations.

B

Backpropagation

A training algorithm used in neural networks to update the model's weights by minimizing the error between predicted and actual outputs.

Bagging

A technique in ensemble learning where multiple models are trained independently and their predictions are combined to improve performance.

Batch Gradient Descent

An optimization algorithm used in machine learning to update model parameters using the average gradient of the loss function across a batch of training examples.

Bayesian Statistics

A statistical approach that uses Bayes' theorem to update the probability of a hypothesis based on new evidence.

Bias

The difference between the true value and the average of predicted values in a model.

Big Data

Extremely large and complex datasets that require specialized tools and techniques for processing and analysis.

Binary Classification

A type of classification problem where the goal is to categorize data into two distinct classes.

Bootstrap Sampling

A statistical technique that involves resampling the dataset to estimate the variability of a model's performance.

Box Plot

A graphical representation of the distribution of a dataset, displaying median, quartiles, and outliers.

Bag-of-Words:

A text representation technique that converts a document into a sparse vector of word frequencies.

Bayesian Optimization:

A hyperparameter optimization technique that uses Bayesian inference to find the best model configuration.

BERT (Bidirectional Encoder Representations from Transformers)

A pre-trained natural language processing model based on transformer architecture.

C

Categorical Variable

A variable that represents categories or groups, often encoded as strings or numerical codes.

Clustering

A unsupervised learning technique that groups similar data points together based on their features.

Confusion Matrix

A table used to evaluate the performance of a classification model, showing true positives, true negatives, false positives, and false negatives.

Convolutional Neural Network (CNN)

A type of deep learning architecture specifically designed for image recognition tasks.

Correlation

A statistical measure indicating the strength and direction of a linear relationship between two variables.

Cross-Entropy

A loss function commonly used in classification tasks, particularly in logistic regression and neural networks.

Cross-Validation

A resampling technique used to assess the performance of a model by splitting the data into subsets for training and testing.

Causal Inference

The process of identifying causal relationships between variables from observational data.

Collaborative Filtering

A recommendation system technique that uses user-item interactions to predict user preferences.

Cronbach's Alpha:

A measure of internal consistency used to assess the reliability of a scale or questionnaire.

CROSSTABS

A function used in data analysis tools to create contingency tables and analyze associations between categorical variables.

D

Data Cleaning

The process of identifying and correcting errors or inconsistencies in a dataset.

Data Engineering

The process of collecting, storing, and processing data to make it accessible for analysis.

Data Mining

The process of discovering patterns, relationships, or insights from large datasets using various techniques.

Data Science

An interdisciplinary field that uses scientific methods, algorithms, processes, and systems to extract knowledge and insights from structured and unstructured data.

Data Visualization

The graphical representation of data to provide insights and aid in understanding patterns or trends.

Decision Tree

A tree-like model used for classification and regression tasks, where each internal node represents a feature, and each leaf node represents a class or value.

Deep Learning

A subset of machine learning that involves neural networks with multiple layers, allowing them to learn complex representations from data.

Dimensionality Reduction

The process of reducing the number of features in a dataset while preserving important information.

Dropout

A regularization technique used in neural networks to prevent overfitting by randomly deactivating neurons during training.

Data Augmentation

The process of generating additional training data by applying transformations to existing data points.

Decision Boundary

The boundary separating different classes or categories in a classification model.

DevOps

A set of practices that combine software development and IT operations to shorten the systems development life cycle.

E

Ensemble Learning

A technique that combines multiple models to improve overall predictive performance.

Ethics in Data Science

The consideration of ethical principles and potential biases when collecting, analyzing, and using data.

Exploratory Data Analysis (EDA)

The process of visually and statistically exploring datasets to uncover patterns, trends, and relationships.

Eigenvalue and Eigenvector

In linear algebra, eigenvalues and eigenvectors represent the properties of a transformation or a matrix.

F

Feature Engineering

The process of creating new features or transforming existing ones to improve a model's performance.

F1 Score

A metric that combines precision and recall to evaluate the performance of a binary classification model.

Forward Propagation

The process in neural networks where input data is fed through the network's layers to generate predictions.

Frequentist Statistics

A statistical approach that focuses on estimating fixed parameters from observed data.

F-measure

A metric that combines precision and recall using a weighted harmonic mean.

Feature Importance

A measure indicating the relative importance of features in a machine learning model.

Fine-Tuning

The process of further training a pre-trained model on a specific task or dataset.

Forward Selection

A feature selection technique that starts with no features and iteratively adds the most relevant ones.

G

Gaussian Distribution (Normal Distribution)

A symmetrical probability distribution commonly found in nature and often used in statistical modeling.

Gradient Descent

An optimization algorithm used to update model parameters iteratively, searching for the minimum of a loss function.

Grid Search

A hyperparameter tuning technique that exhaustively searches through a predefined set of hyperparameter combinations to find the best model.

GAN (Generative Adversarial Network)

A type of deep learning model that uses a generator and a discriminator to create new data samples.

Gradient Boosting

An ensemble learning technique that combines weak learners (e.g., decision trees) to create a strong model.

H

Hadoop

An open-source framework for distributed storage and processing of big data.

Hyperparameter

A parameter set before model training that influences the learning process, such as learning rate and number of hidden layers.

Hypothesis Testing

A statistical method used to make inferences about a population based on a sample of data.

Hashing Trick

A technique used to convert large categorical datasets into a fixed-size vector representation.

Heteroscedasticity:

In statistics, a condition where the variability of a variable's distribution changes across different levels of another variable.

I

Imputation

The process of filling missing values in a dataset using various techniques.

Independent Variable

A variable in a study that is manipulated or controlled to understand its effect on the dependent variable.

Information Gain

A measure used in decision trees to evaluate the effectiveness of a feature in splitting the data.

Interpolation

The process of estimating values within a range based on existing data points.

Inference

The process of drawing conclusions from data, often involving hypothesis testing.

K

k-Nearest Neighbors (k-NN)

A supervised learning algorithm used for classification and regression tasks, where predictions are based on the similarity to k-nearest data points.

K-Means Clustering

An unsupervised learning algorithm used to partition data into k clusters based on similarity.

Kernel

A function used in support vector machines (SVM) to map data into a higher-dimensional space for better separation.

K-Fold Cross-Validation

A method of cross-validation that partitions the data into k subsets and iteratively uses each subset as a validation set while the rest serve as training data.

Keras

An open-source deep learning library that provides a high-level neural networks API.

Kolmogorov-Smirnov Test

A statistical test used to compare the distribution of a sample with a theoretical distribution.

L

L1 Regularization (Lasso)

A regularization technique that adds the absolute values of the model's coefficients to the loss function.

L2 Regularization (Ridge)

A regularization technique that adds the square of the model's coefficients to the loss function.

Learning Rate

A hyperparameter in gradient-based optimization algorithms that controls the step size during parameter updates.

Logistic Regression

A statistical method used for binary classification, where the output is transformed using the logistic function.

Loss Function

A function that measures the difference between predicted and actual values, used to optimize model parameters.

Lag Plot

A graphical tool used to detect patterns or relationships in time series data.

Latent Dirichlet Allocation (LDA)

A probabilistic model used for topic modeling.

Leaky ReLU

An activation function used in neural networks that addresses the "dying ReLU" problem by allowing small negative values.

LSTMs (Long Short-Term Memory)

A type of recurrent neural network architecture that can model long-term dependencies in sequences.

M

Machine Learning

The study of algorithms and statistical models that enable computers to perform tasks without explicit programming.

Mean Absolute Error (MAE)

A metric that measures the average absolute difference between predicted and actual values.

Mean Squared Error (MSE)

A metric that measures the average squared difference between predicted and actual values.

Model Selection

The process of choosing the best model among different candidates based on performance metrics.

Manifold Learning:

A dimensionality reduction technique that preserves the intrinsic structure of high-dimensional data.

Markov Chain

A sequence of events where the probability of each event depends only on the previous event.

Mean Absolute Percentage Error (MAPE)

A metric that measures the percentage difference between predicted and actual values.

Memory-based Collaborative Filtering

A recommendation system technique that uses user-item interactions directly for predictions.

Multicollinearity

A condition in linear regression where two or more predictor variables are highly correlated, causing issues with coefficient interpretation.

Multiclass Classification:

A classification task where data is categorized into more than two classes.

Multilayer Perceptron (MLP)

A type of feedforward neural network with multiple hidden layers.

Multi-Task Learning

A machine learning technique where a model is trained on multiple related tasks simultaneously.

N

Natural Language Processing (NLP)

A field of AI focused on enabling computers to understand and process human language.

Neural Network

A type of machine learning model inspired by the human brain's structure, consisting of interconnected layers of artificial neurons.

Normalization

The process of scaling data to have a mean of zero and a standard deviation of one.

O

One-Hot Encoding

A technique used to convert categorical variables into binary vectors to be used as input in machine learning models.

Outlier

A data point that significantly deviates from the general pattern of the dataset.

Overfitting

A phenomenon where a model performs well on the training data but poorly on unseen data due to memorizing noise instead of learning general patterns.

Object Detection

A computer vision task where the goal is to identify and locate objects within images or videos.

One-vs-Rest (OvR):

A strategy for multiclass classification where each class is treated as a binary problem against all other classes.

Over-sampling

A technique used to balance imbalanced datasets by generating synthetic examples of the minority class.

P

PCA (Principal Component Analysis)

A dimensionality reduction technique that transforms data into a lower-dimensional space while preserving the most important information.

Pearson Correlation Coefficient

A statistical measure indicating the linear relationship between two continuous variables.

Precision

The number of true positive predictions divided by the total number of positive predictions in a classification model.

Principal Component

The transformed feature resulting from PCA that captures the most significant variation in the data.

Probability Distribution

A function that describes the likelihood of different outcomes occurring in a random experiment.

Python

A popular programming language commonly used in data science and machine learning.

Pandas

An open-source Python library used for data manipulation and analysis.

Perceptron

The simplest form of a neural network with one layer, used for binary classification.

Pipeline

A series of data processing steps and models combined into a single workflow.

Poisson Distribution

A discrete probability distribution often used for modeling count data.

Precision-Recall Curve

A graphical tool to evaluate the trade-off between precision and recall for different classification thresholds.

Proportional Hazard Model (Cox Regression)

A statistical model used for survival analysis in time-to-event data.

Q

QuickSort

A fast sorting algorithm commonly used in data processing.

R

Random Forest

An ensemble learning technique that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.

Recall

The number of true positive predictions divided by the total number of actual positive instances in a classification model.

Regularization

Techniques used to prevent overfitting by adding penalties to the loss function for large model parameters.

Regression

A type of supervised learning task that predicts continuous numerical values.

Reinforcement Learning

A type of machine learning where an agent learns to make decisions by interacting with an environment and receiving rewards.

Resampling

Techniques such as bootstrapping and cross-validation used to repeatedly draw samples from a dataset for statistical analysis.

Root Mean Squared Error (RMSE)

A metric that measures the square root of the average squared difference between predicted and actual values.

R-Squared (R2)

A metric that measures the proportion of variance in the dependent variable explained by the model.

Random Search

A hyperparameter tuning technique that randomly samples hyperparameters from predefined ranges.

Recommender System

A type of information filtering system that suggests items or content based on user preferences.

Root Cause Analysis

The process of identifying the fundamental reason for a problem or issue in data analysis or system performance.

S

Scikit-learn

An open-source machine learning library for Python.

Semi-Supervised Learning

A learning paradigm where a model is trained on both labeled and unlabeled data.

Sentiment Analysis:

A natural language processing task that determines the sentiment or emotion expressed in a piece of text.

Singular Value Decomposition (SVD)

A technique used in dimensionality reduction and matrix factorization.

Support Vector Machine (SVM)

A supervised learning algorithm used for classification and regression tasks.

T

TensorFlow

An open-source deep learning library developed by Google.

Time Series Analysis

The study of data collected over time to identify patterns and make predictions.

Training Set

A subset of data used to train a machine learning model.

U

Unsupervised Learning

A type of machine learning where the model learns from unlabeled data to find patterns and relationships.

V

Variance

The measure of how spread out the data points are in a dataset.

Vectorization

The process of converting non-numeric data into a numerical format suitable for machine learning models.

X

XGBoost

An optimized gradient boosting library widely used for machine learning tasks.

Y

Yield

In the context of data processing, the percentage of successfully processed data compared to the total amount of data.

Z

Z-Score

A measure that standardizes data points by subtracting the mean and dividing by the standard deviation.

Zero-Inflated Model

A statistical model used when a significant number of data points have zero values and require special handling.