Data Science, Statistics and Machine Learning Dictionary

Data dictionaries, manuals, and all explanatory materials.

A

Algorithm - a set of repeatable instructions for solving a particular task.

Anaconda - an open-source distribution for Data Science, which simplifies packages and deployment. Mostly used for Python and R programming language users.

Apache Spark - an open-source computing framework and set of libraries for real-time, large-scale data processing.

Artificial Intelligence - is the ability of a computer or a computer-controlled robot to perform tasks that are usually done by humans as they require human intelligence.

B

Backpropagation - sometimes abbreviated as "backprop". It is the messenger who tells the Neural Network whether it made a mistake when it made a prediction, by iteratively adjusting the weights.

Bagging - also known as bootstrap is a technique where predictions are made using a combination of predictions from multiple models created on a subset of data.

Bayes' Theorem - a mathematical formula used to determine conditional probability.

Bayesian Networks - a type of probabilistic graphical model, which aims to develop a model that maintains known conditional dependence between random variables and conditional independence in all other cases.

Bias - is a systematic error due to incorrect assumptions in a Machine Learning process.

Big Data - is data that can be very impractical due to its great variety, increasing volumes, and velocity.

Binary - in the context of a variable, these only have two unique values such as “Yes” and “No”

Binomial Distribution - is a method of calculating probabilities for experiments having a fixed number of trials.

Boosting - is a sequential process where a model corrects errors learning from previous models.

C

Categorical variables - are variables that have discrete qualitative values, such as race or religion.

Chi-square test - a statistical method used to test and compare observed results with expected results.

Classification - is predicting a label, by identifying which category an object belongs to based on different parameters.

Clustering - an unsupervised algorithm process of dividing data points into particular groups.

Computer Vision - a form of AI that allows computers to visualize, process, and identify images/videos in the same way that a human vision does.

Confidence interval - a statistical method used to estimate what percent of a population fits in a particular category based on the results from a sample population.

Confusion matrix - a table that is used to describe the performance of a classification model.

Continuous variables - are variables that have an infinite number of values, such as speed and distance.

Convex function - if the line segment between any two points on the graph lies either above or on the graph.

Correlation - is the ratio of covariance/relationship between two or more variables.

Cost function - a statistical approach used to define and measure the error of the model.

Covariance - a measure of the relationship between two random variables.

Cross-Entropy - a measure of the difference between two probability distributions for a set of events.

Cross-Validation - a statistical technique used that evaluates and compares machine learning algorithms by dividing data into two segments.

D

Data Engineers - data professionals that are responsible for setting up and maintaining the organization's data infrastructure.

Data Mining - the process of extracting useful information from both structured and unstructured data.

Data Science - preparing data for the analysis process of cleansing, manipulating, algorithmic development, and more to the data to perform advanced data analysis.

Dashboard - an information management tool used to track, analyze and display

performance. The most common tools for building dashboards include Excel and Tableau.

Database - is a structured collection of data, which is organized in an accessible way. The common database language is SQL.

Data Augmentation - a technique used to increase the amount of data by adding slight adjustments to existing data.

Decision Trees - a non-parametric supervised learning method used for classification and regression, which aims to build a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Deep Learning - a Machine Learning method that teaches computers to do what comes naturally to humans. It trains an algorithm to predict outputs, given a set of inputs.

Dependent Variable - the variable that is measured and is affected by the independent variable.

Dimensionality Reduction - the process of reducing the number of input variables in training data.

E

Early stopping - a technique used to avoid overfitting when training a machine learning model using an iterative method.

Exploratory data analysis - is a critical process in the initial investigations of data to provide more insights through visualization or statistical analysis.

ETL - is a popular acronym that stands for Extract, Transform, and Load. An ETL system extracts data from source systems, enforcing its quality as well as presenting the data.

Evaluation metric - metrics that are used to measure the quality of a machine learning model, such as AUC.

F

False Negative - predictions that are true but are incorrectly predicted as false

False Positive - predictions that are false but are incorrectly predicted as true.

Feature Reduction - is the process of reducing the number of features to improve the efficiency of a computation-intensive task without losing information.

Feature Selection - is the process of reducing the number of input variables by selecting relevant features to use in the model.

F-score - is a measure of the model's accuracy on a dataset.

G

GPU - stands for graphic processing unit, a specialized processor which can process pieces of data for machine learning, video editing, and gaming applications.

Gradient Boosting - is the process of relying on using previous models to improve the next model, and minimizing the overall prediction error.

Gradient Descent - is an optimization algorithm that helps to find a local minimum/maximum of a given function.

H

Hadoop - an open-source framework used to efficiently store and process large datasets.

Hierarchical Clustering - an algorithm that groups similar data points into groups called clusters.

Histogram - a graphical representation organizing a group of data points in continuous variables

Holdout Sample - a random sample taken from a data set that is not used in the model fitting process.

Hyperparameter Tuning - the process of finding the optimal hyperparameters for machine learning algorithms.

I

Independent Variable - the variable that can manipulate or have a direct effect on the dependent variable.

Iteration - the process of repeating a statement/block of code a specific number of times, producing an output one after another

J

Jupyter Notebook - a web-based interactive computing platform that is used for creating and sharing computational documents.

K

K-means - an unsupervised learning algorithm used to group data points to the nearest centroid through distance.

Keras - an open-source software library developed by Google for implementing neural networks.

K-nearest neighbors (KNN) - a supervised machine learning algorithm used for both regression and classification tasks. It is used to make predictions on the test data set by calculating the distance between the current training data points.

Kubernetes - an open-source platform for automating application deployment, scaling, and management.

L

Labeled dataset - data that has a “label”, “class” or “tag” associated with it.

Lasso Regression - the process of shrinking or regularizing to avoid overfitting to minimize prediction error.

Linear Regression - is used to make predictions on continuous dependent variables with the use of independent variables.

Logistic Regression - is used to predict the categorical dependent variable with the use of independent variables, to classify outputs, which can only be between 0 and 1.

Log Loss - measures the performance of a classification model, where the output is a probability with values between 0 and 1.

Long Short Term Memory Networks - a type of Recurrent Neural Network which can learn and memorize long-term dependencies. An LSTM aims to remember past information for long periods.

M

Machine Learning - a process where models use historical data as an input to predict new output values, a method used to identify and learn patterns in data analysis.

Machine Learning Operations (MLOps) - a core function for Machine Learning engineering which focused on the process of taking machine learning models to production, and then maintaining and monitoring them.

Management information system (MIS) - a computer system consisting of hardware and software that serves as the backbone of an organization's operations

Maximum Likelihood Estimation - a probabilistic framework to get more robust parameter estimates.

Mean - is the average value of all the numbers

Mean Absolute Error - also known as L1 regularization computes the mean of squares of errors between labeled data and predicted data.

Mean Square Error Loss - also known as L2 regularization tells you how close a regression line is to a set of data points.

Median - is the middle value in a list ordered from smallest to largest.

Mode - the value that appears most frequently in a data set.

Model selection - the process of selecting a statistical model from a set of known models.

Monte Carlo method - a mathematical technique used to estimate the possible outcomes of an uncertain event.

Multi-Class Classification - classification problems that have more than one class in the target variable.

Multilayer Perceptrons - is a feedforward artificial neural network, where a set of inputs are fed into the Neural Network to generate a set of outputs.

Multivariate analysis - the process of comparing and analyzing the dependency of multiple variables over one another.

N

Naive Bayes - a process by using a classifier that assumes the independence between attributes of data points, based on the Bayes Theorem.

NaN - this stands for ‘not a number’ and refers to a numeric data type value that is undefined or unrepresented. This will be considered missing or represented incorrectly in a dataset.

Natural Language Processing - is the ability of a computer to be able to detect and understand human language, through speech and text just the way we humans can.

Neural Network - is a network made up of neurons that contain three different layers: an input layer, one or more hidden layers, and an output layer

NoSQL - stands for Not only SQL and is a database that provides storage and retrieval of data.

Nominal Variable - a type of the variable used to name, label, or categorize particular attributes that are being measured.

Normal distribution - a probability distribution function that represents the distribution of random variables in a bell-shaped graph.

Normalization - a scaling technique used to shift and rescale data into the range [0, 1].

NumPy - a library used for Python that has mathematical functions in processing multidimensional array objects, linear algebra, and matrix calculation functions.

O

One Hot Encoding - a process where categorical variables are converted into a form for machine and deep learning algorithms to improve predictions and accuracy of a model.

Ordinal Variable - are variables that have discrete values with some form of order involved.
Outlier - an observation that is far away from the overall pattern in a sample.

Overfitting - is when a statistical model fits exactly against its training data. It is a modeling error when a function is too closely fits a limited set of data points.

P

Pandas - an open-source Python library for data manipulation and analysis

Parameters - is a set of measurable factors that define a system and the part of the model that is learned from past training data.

Precision - is the measure of total actual positive cases and the quality of a positive prediction made by the model.

Predictive modeling - the process of using a mathematical approach to predict future events or outcomes by analyzing patterns in a given set of input data.

Predictor variable - is the variable used to make a prediction for dependent variables.

Pre-trained model - a model created by someone else that can solve a similar problem, rather than building a model from scratch.

Principal Component Analysis - a technique used to reduce the dimensionality of datasets by increasing model interpretability without minimizing information loss.

Probability distribution - a statistical function that describes all possible values and their occurrence.

P-value - is the probability that the results from your sample data occurred by chance, therefore a low p-value is good.

R

R - an open-source programming language as well as a software environment for statistical computing, machine learning, and data visualization.

Random Forest - is made up of many decision trees and an ensemble learning method for classification, regression, and other tasks that contain multiple Decision Trees.

Regression - a technique used for investigating the relationship between independent variables or features and a dependent variable or outcome.

Regularization - a technique used to solve overfitting in statistical models.

Reinforcement Learning - the aim is to train a model to return an optimum solution using a sequence of solutions and/or decisions that have been created for a specific problem.

Ruby - an open-source programming language that is mostly used for building web applications.

S

Scikit-learn - a library for the Python users which contains tools for machine learning, and statistical modeling such as classification, regression, clustering, and dimensionality reduction

SQL - stands for Structured Query Language, and is used to manage databases by performing tasks such as updating the data, retrieving data, etc.

Standard Deviation - tells you the variation of the data around the mean

Standard Error - tells you the variation of the different means calculated

Stochastic Gradient Descent - the aim is to minimize the Cost Function by incrementally changing the weights of the network.

Supervised Learning - the type of learning when an algorithm learns on a labeled dataset and analyses the training data

Support Vector Machine - a supervised learning model that creates a line or a hyperplane which separates the data into classes

T

T-Distribution - a probability distribution that describes the standardized distances of sample means to the population mean, similar to normal distribution.

T-Value - The variance between and within groups, where a big T-Value means different groups, and a small T-Value mean similar groups.

TensorFlow - is an open-source library for deep learning applications that make it easy to build models through large-scale neural networks with many layers using data flow graphs.

Tokenization - the process of splitting a text string into units called tokens and is a part of Natural Language Processing.

Transfer Learning - a machine learning method where the application of knowledge obtained from a model used in one task, can be reused as a foundation point for another task.

**True Positive - You predicted positive and its actually positive

True Negative - You predicted negative and its actually negative

T-test - a test used to compare two populations by finding the difference in their population means.

Type I error - the decision to reject the null hypothesis as it could be incorrect.

Type II error - the decision to retain the null hypothesis as it could be incorrect.

U

Underfitting - a modeling error that can neither model training data nor generalizes new data and does not perform well on the training set.

Unsupervised Learning - where a model learns on unlabeled data, inferring more about hidden structures to produce accurate and reliable outputs.

V

Variance - is used to measure the spread of a given set of numbers.

**Vectors* - *are used to represent numeric characteristics known as features in a mathematical and easily analyzable form.

REFERENCES

- Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python Book by Jason Brownlee
- https://www.qualtrics.com/experience-management/employee/performance-appraisal/ (https://www.qualtrics.com/experience-management/employee/performance-appraisal/)
- https://www.businessnewsdaily.com/15967-how-to-evaluate-employee-performance-.html (https://www.businessnewsdaily.com/15967-how-to-evaluate-employee-performance-.html)
- https://www.businessnewsdaily.com/5760-write-good-performance-review.html (https://www.businessnewsdaily.com/5760-write-good-performance-review.html)
- https://www.perkbox.com/uk/resources/blog/employee-performance (https://www.perkbox.com/uk/resources/blog/employee-performance)
- https://medium.com/low-code-for-advanced-data-science/modeling-employee-performance-analysis-15eec3821d84 (https://medium.com/low-code-for-advanced-data-science/modeling-employee-performance-analysis-15eec3821d84)

- https://medium.com/@millicentnyuguto211/employee-performance-analysis-57731b12d9eb (https://medium.com/@millicentnyuguto211/employee-performance-analysis-57731b12d9eb)

- Cross Validated, Machine Learning Questions and Answers. https://stats.stackexchange.com/ (https://stats.stackexchange.com/)

- Stack Overflow, Programming Questions and Answers. https://stackoverflow.com/ (https://stackoverflow.com/)

- Scikit-Learn API Reference. https://scikit-learn.org/stable/modules/classes.html (https://scikit-learn.org/stable/modules/classes.html)

- Matplotlib API Reference. http://matplotlib.org/api/index.html (http://matplotlib.org/api/index.html)

- https://www.kdnuggets.com/2022/05/data-science-statistics-machine-learning-dictionary.html (https://www.kdnuggets.com/2022/05/data-science-statistics-machine-learning-dictionary.html)
- Finding Structure With Randomness: Probabilistic Algorithms For Constructing Approximate Matrix Decompositions, 2009. https://arxiv.org/abs/0909.4061 (https://arxiv.org/abs/0909.4061)

- Dimensionality reduction, Wikipedia. https://en.wikipedia.org/wiki/Dimensionality_reduction (https://en.wikipedia.org/wiki/Dimensionality_reduction)

- Curse of dimensionality, Wikipedia. https://en.wikipedia.org/wiki/Curse_of_dimensionality (https://en.wikipedia.org/wiki/Curse_of_dimensionality)

- Singular value decomposition, Wikipedia. https://en.wikipedia.org/wiki/Singular_value_decomposi (https://en.wikipedia.org/wiki/Singular_value_decomposi)
- sklearn.linear model.LogisticRegression API. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

- Non-linear transformation, scikit-learn Guide. https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-transformer (https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-transformer)

- sklearn.preprocessing.PowerTransformer API. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer) .html