Machine Learning Glossary

Evan Misshula

June 8, 2025

1 Glossary

1.1 Section 1: ML Concepts and Isolation Forests

Anomaly Score
A value between 0 and 1 indicating how easily a data point can be isolated in an isolation forest; higher values suggest anomalies.

Classification
A machine learning task that assigns labels to input data from a set of discrete categories.

Data Cleaning
The process of correcting or removing inaccurate records from a dataset, including handling missing values, outliers, and duplicates.

Data Collection
The act of gathering and measuring information from various sources to answer a research question or model a problem.

Dependent Variable
The target or outcome variable that a model is trained to predict, also known as label, response, or outcome variable.

Duplicate
A record that appears more than once in a dataset and represents the same real-world entity.

Expected Path Length
The average number of steps required to isolate a data point across all trees in an isolation forest.

Feature Engineering
: The process of selecting, modifying, or creating new features from raw data to improve model performance.

Isolation Forest
: An anomaly detection method that isolates observations by randomly selecting features and split values.

Isolation Tree
: A binary tree used in isolation forests where each node splits data based on randomly chosen feature and split value.

Missing Value
: A data entry where no value is recorded, often due to errors in data collection or system limitations.

Model Deployment
: The process of integrating a trained machine learning model into a production environment where it can make predictions.

Outlier
: A data point that differs significantly from other observations and may indicate an anomaly or data error.

Path Length
: The number of edges from the root of an isolation tree to the node where a data point is isolated.

Regression
: A type of machine learning task that involves predicting continuous numeric values.

Tree Depth
: The length of the longest path from the root to a leaf in a tree structure.

Unsupervised Learning
: A category of machine learning where the model learns patterns from data without labeled responses.

Workflow
: The sequence of steps in a machine learning process including data collection, cleaning, modeling, and deployment.

1.2 Section 2: Exploratory Data Analysis and Feature Engineering

Analysis
The phase in ML workflow where insights are extracted from data, including EDA and feature engineering.

Boxplot
A graphical representation of data distribution highlighting the median, quartiles, and outliers.

Data Leakage
Occurs when information from outside the training dataset is used to create the model, leading to overly optimistic results.

EDA (Exploratory Data Analysis)
A process of analyzing datasets to summarize their main characteristics, often using visual methods.

Feature Extraction
Creating new features from raw data, often by dimensionality reduction techniques like PCA or t-SNE.

Feature Selection
Choosing a subset of relevant features for use in model construction, e.g., using RFE.

Histogram
A plot showing the frequency distribution of a dataset, useful for identifying skewness and modality.

Joint Distribution
A probability distribution over multiple random variables describing their simultaneous behavior.

KL Divergence
A non-symmetric measure of how one probability distribution diverges from a second, expected distribution.

Marginal Distribution
The distribution of a single variable within a joint distribution, integrating out others.

Mutual Information
: A measure of the mutual dependence between two variables, capturing both linear and nonlinear associations.

Normalization/Scaling
: Techniques such as MinMax or StandardScaler that adjust features to a common scale.

Pair Plot
: A matrix of scatter plots to visualize pairwise relationships between variables.

Q-Q Plot
: A quantile-quantile plot used to assess if a dataset follows a particular distribution.

Recursive Feature Elimination (RFE)
: An iterative method to select features by recursively removing the least important features.

Skewness
: A measure of asymmetry in the distribution of data.

StandardScaler
: A normalization technique that rescales features to have zero mean and unit variance.

t-SNE
: A nonlinear dimensionality reduction technique that visualizes high-dimensional data by preserving local structure.

UMAP
: A dimensionality reduction technique that preserves both local and global structure in data.

1.3 Section 3: Model Training and Evaluation

Bias-Variance Tradeoff
: The tension between model complexity and prediction error: simpler models have high bias and low variance, while complex models have low bias and high variance.

Cross-Validation
: A technique for estimating the generalization error of a model by training and evaluating it on different data subsets.

Empirical Risk
: The average loss computed on the training data; also called training error.

Evaluation Metrics
: Quantitative measures used to assess the performance of a machine learning model, e.g., accuracy, precision, recall, MSE.

Expected Risk
: The true generalization error computed over the full data distribution.

F1 Score
: The harmonic mean of precision and recall, used as a balanced performance metric in classification tasks.

Generalization Gap
: The difference between expected risk and empirical risk; large gaps indicate overfitting.

Loss Function
: A function that quantifies the error between predicted and true values; minimized during training.

Mean Squared Error (MSE)
: A regression metric that averages the squared differences between predicted and true values.

Model Selection
: The process of choosing the best-performing model based on validation performance, not test data.

Model Training
: The process of fitting a model's parameters to minimize a specified loss function on training data.

Overfitting
: A situation where a model performs well on training data but poorly on unseen data due to high variance.

Precision
: The proportion of true positives among all predicted positives in a classification task.

Recall
: The proportion of true positives detected among all actual positives in a classification task.

Rš Score
: A metric for regression that indicates the proportion of variance explained by the model.

Training Error
: The average loss of a model on training data; also referred to as empirical risk.

Validation Set
: A subset of the dataset used to tune model hyperparameters and evaluate performance during model selection.

Workflow
: The structured pipeline of steps in machine learning, including data preparation, modeling, and evaluation.