ML

A concise cheat sheet outlining the key concepts, algorithms, and differences between supervised and unsupervised learning methods in machine learning.

Core Concepts

Supervised Learning

Definition: Learning from labeled data to predict outcomes for new, unseen data.

Goal: To map input variables (X) to an output variable (Y) based on a labeled dataset.

Process: Model trained on labeled data → Model predicts outcome on new data → Accuracy measured by comparing predicted vs. actual values.

Common Use Cases: Classification (categorizing data) and Regression (predicting continuous values).

Unsupervised Learning

Definition: Learning from unlabeled data to discover hidden patterns or structures.

Goal: To find inherent structure in data without explicit labels.

Process: Model explores unlabeled data → Identifies patterns, clusters, or reduces dimensionality.

Common Use Cases: Clustering (grouping similar data points) and Dimensionality Reduction (reducing the number of variables).

Supervised Learning Algorithms

Classification Algorithms

Logistic Regression: Predicts the probability of a binary outcome. Good for binary classification problems.
Types:Binary , multinomial and ordinal

Support Vector Machines (SVM): Finds the optimal hyperplane that separates data into classes. Effective in high dimensional spaces.

Decision Trees: Tree-like structure to classify data based on features. Easy to interpret, but prone to overfitting.
Types:ID3,Gini Index and Split Creation

Random Forest: Ensemble of decision trees that improves accuracy and reduces overfitting.

Naive Bayes: Applies Bayes' theorem with strong (naive) independence assumptions between features. Simple and fast, but assumptions may not hold in real-world data.

K-Nearest Neighbors (KNN): Classifies a data point based on the majority class of its k-nearest neighbors.

Confusion Matrix: A table that summarizes the performance of a classification model by showing the counts of true positive, true negative, false positive, and false negative predictions. Used to calculate various metrics like accuracy, precision, recall, and F1-score.

Stochastic Gradient Descent (SGD): An iterative optimization algorithm used to find the minimum of a cost function. It updates model parameters using the gradient of the cost function computed on a small, randomly selected subset of the training data (a mini-batch) at each iteration. SGD is computationally efficient and suitable for large datasets.

Regression Algorithms

Linear Regression: Models the relationship between variables using a linear equation. Simple and interpretable.

Polynomial Regression: Models the relationship using a polynomial equation. Can capture non-linear relationships.

Support Vector Regression (SVR): Uses SVM principles to predict continuous values.

Decision Tree Regression: Uses decision trees to predict continuous values. Prone to overfitting.

Random Forest Regression: Ensemble of decision trees for regression. Improves accuracy and reduces overfitting.