

Machine Learning - Comprehensive Guide

Basics of Machine Learning

Machine Learning (ML) is a subset of Artificial Intelligence that focuses on developing systems that learn from data and improve performance over time without being explicitly programmed.

Types include:

- Supervised Learning (e.g. regression, classification),
- Unsupervised Learning (e.g. clustering, dimensionality reduction),
- Semi-Supervised Learning, and
- Reinforcement Learning (learning through feedback from environment).

Common Algorithms

- Linear Regression: Predicts continuous outputs using a linear relationship.
- Logistic Regression: Binary classification using the sigmoid function.
- Decision Tree: Tree structure for decision-making, interpretable but can overfit.
- Random Forest: Ensemble of decision trees to reduce overfitting.
- SVM: Finds hyperplanes to separate classes effectively.
- k-NN: Classifies based on proximity to neighbors.
- XGBoost: Fast and accurate gradient boosting algorithm.
- K-Means: Clustering technique for grouping similar items.
- PCA: Reduces data dimensions while preserving variance.

Key ML Concepts

- Overfitting: Model fits training data too well, performs poorly on new data.
- Underfitting: Model is too simple, fails to capture patterns.
- Bias-Variance Tradeoff: Balancing model complexity and accuracy.
- Cross-Validation: Ensures model generalization.
- Evaluation Metrics: Accuracy, Precision, Recall, F1-Score for classification; MSE, RMSE, MAE for

Machine Learning - Comprehensive Guide

regression.

Data Preprocessing

- Handle missing values (imputation, deletion).
- Encode categorical variables (One-hot, Label).
- Feature scaling (Standardization, Normalization).
- Split data into training/testing sets (e.g., 80/20 split).

Model Deployment

- Train and export model using Pickle/Joblib.
- Create API (e.g., Flask/FastAPI) to serve model.
- Host model on cloud (AWS, Azure, etc.).
- Monitor performance and retrain as needed.

Useful Libraries

- Scikit-learn: ML algorithms and tools.
- Pandas: Data manipulation.
- NumPy: Numerical computation.
- Matplotlib/Seaborn: Visualization.
- TensorFlow/PyTorch: Deep learning frameworks.
- XGBoost/LightGBM: Gradient boosting methods.

ML Workflow

1. Define the problem.
2. Collect and clean the data.

Machine Learning - Comprehensive Guide

3. Perform EDA.
4. Engineer features.
5. Choose and train a model.
6. Tune hyperparameters.
7. Evaluate and validate.
8. Deploy model.
9. Monitor and maintain model.

Glossary

- Feature: Input variable.
- Label: Target output variable.
- Epoch: One full training pass over the dataset.
- Learning Rate: Controls step size in optimization.
- Gradient Descent: Optimization to minimize loss.