Machine Learning Algorithms Every Data Scientist Should Know

Plain Language with Math Explanation

Madhusudhan Pandey, PhD

May 21, 2025

What is Machine Learning?

- Teaching computers to learn from examples.
- Given inputs and outputs, the model learns to predict new outputs.
- Like teaching a kid by showing questions and answers.

Supervised Learning

- Inputs (features) and correct outputs (labels) are given.
- The model learns to predict outputs for new inputs.
- Two main types:
  - Classification (output is category)
  - Regression (output is number)

Classification: Naïve Bayes

Example: Sorting emails into "spam" or "not spam."

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}$$

- $P(C|X)$: Probability of class C given data X.
- $P(X|C)$: Probability of data X if class is C.
- $P(C)$: How common class C is.
- $P(X)$: How common data X is overall.

Classification: Logistic Regression

Example: Predict loan approval (yes/no).

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}}$$

- $P(y = 1|x)$: Probability output is 1 (e.g., loan approved).
- e: Euler's number (2.718).
- β_i: Coefficients learned for each input feature x_i.

Classification: K-Nearest Neighbor (KNN)

Example: Recommend movies based on friends with similar tastes.

$$\text{Find } k \text{ nearest neighbors by distance } \|x_i - x_j\|_2$$

- k: Number of closest neighbors to consider.
- $\|x_i - x_j\|_2$: Euclidean distance between data points.

Classification: Random Forest

Example: Diagnose illness by combining many decision trees.

$$\hat{y} = \text{majority vote from decision trees}$$

- $\hat{y}$: Final predicted class.
- Each tree votes; majority vote wins.

Classification: Support Vector Machine (SVM)

Example: Recognize faces by finding the best boundary.

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t. } y_i(\mathbf{w} \cdot x_i + b) \geq 1$$

- $\mathbf{w}$: Vector defining boundary direction.
- b: Bias term (boundary shift).
- y_i: True class (+1 or -1).
- x_i: Feature vector.
- $\mathbf{w} \cdot x_i$: Dot product of weights and features.

Classification: Decision Tree

Example: Predict customer churn by asking yes/no questions.

Split data using measures like entropy or Gini index that measure group purity.

Regression: Simple Linear Regression

Example: Predict house price from size.

$$y = \beta_0 + \beta_1 x + \epsilon$$

- y: Predicted output (price).
- x: Input feature (size).
- β_0: Intercept (starting value).
- β_1: Effect of x on y.
- ϵ: Error term.

Regression: Multivariate Regression

Example: Predict salary from experience, education, location.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{y}$: Output vector.
- $\mathbf{X}$: Input feature matrix.
- $\boldsymbol{\beta}$: Coefficients vector.
- $\boldsymbol{\epsilon}$: Errors vector.

Example: Select important features for sales prediction.

$$\min_{\beta} \left\{ \sum (y_i - X_i\beta)^2 + \lambda \sum |\beta_j| \right\}$$

- y_i: Actual output.
- $X_i\beta$: Predicted output.
- λ: Penalty controlling feature selection.
- $|\beta_j|$: Absolute value of coefficients.

Unsupervised Learning

- Only inputs given; no output labels.
- Goal: Find hidden patterns or groupings.

Clustering: K-Means

Example: Group customers by spending habits.

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

- k: Number of clusters.
- C_i: Cluster i.
- x: Data points.
- μ_i: Cluster center.

Other Clustering Methods

DBSCAN: Groups dense regions, no fixed number of clusters needed.
PCA (Principal Component Analysis): Compress data by finding new axes that capture most variation.

$$\max_{W} \|XW\|^2 \quad \text{s.t. } W^T W = I$$

- X: Data matrix.
- W: New directions (axes).
- $W^T W = I$: New axes are perpendicular.

Independent Component Analysis (ICA)

Example: Separate voices from a noisy recording.

$$X = AS \quad \Rightarrow \quad \text{Find } A^{-1} \text{ to get } S$$

- X: Mixed signals.
- A: Mixing matrix.
- S: Original independent signals.

Example: If a customer buys bread and butter, likely buys milk too.

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- $P(B|A)$: Probability of buying B given A.
- $P(A \cap B)$: Probability of buying both A and B.
- $P(A)$: Probability of buying A.

Anomaly Detection: Z-Score

Example: Spot unusual bank transactions.

$$Z = \frac{x - \mu}{\sigma}$$

- x: Observed value.
- μ: Mean.
- σ: Standard deviation.
- Z: Number of standard deviations away from mean.

Anomaly Detection: Isolation Forest

Isolates anomalies by partitioning data; anomalies are isolated quickly.
(No explicit formula)

Semi-Supervised Learning

- Some data labeled, most unlabeled.
- Use labeled data to help label unlabeled.
- Examples:
  - Self-Training
  - Co-Training

Reinforcement Learning

- Learn by trial and error with rewards.
- Like teaching a child to walk.

Policy Optimization

$$\max_{\theta} \mathbb{E}\left[\sum_{t} \gamma^{t} r_{t}\right]$$

- θ: Policy parameters.
- r_t: Reward at time t.
- γ: Discount factor (future rewards importance).

Q-Learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

- $Q(s,a)$: Value of action a in state s.
- α: Learning rate.
- r: Reward.
- s': Next state.
- γ: Discount factor.