MACHINE LEARNING

Unit I — Fundamentals of Machine Learning

Topics Covered in this Unit

Introduction & Scope of ML, Tom Mitchell's Formal Definition & Real-World Applications
Comparative Analysis: Artificial Intelligence vs. Machine Learning vs. Data Science
Traditional Programming vs. Machine Learning Paradigm (Architecture & Workflow)
Taxonomy of Learning: Supervised, Unsupervised, Semi-Supervised, and Reinforcement Learning
Models of Machine Learning: Geometric, Probabilistic, Logical, Grouping vs. Grading, Parametric Models
Feature Engineering & Data Preprocessing: Scaling, Imputation, One-Hot Encoding & Selection
Curse of Dimensionality: Geometric Sparsity, Distance Breakdown & Computational Bottlenecks
Principal Component Analysis (PCA): Mathematical Derivation, Covariance & Eigen-decomposition
Linear Discriminant Analysis (LDA): Fisher's Criterion, Scatter Matrices & Supervised Projection
Master Comparison Table: PCA vs. LDA & High-Yield University Exam Question Bank

Compiled in a structured, easy-to-understand, textbook style format
Specially curated for B.Tech / B.E. / BCA / MCA / B.Sc Computer Science & Data Science Students

COURSE SYLLABUS & MAPPING

[UNIT I SYLLABUS — PRESCRIBED UNIVERSITY CURRICULUM]

This section contains the official syllabus for Unit I. Verify with your university curriculum mapping.

Unit I Syllabus Breakdown:

Unit I - Fundamentals of Machine Learning: Introduction to machine learning, scope of machine learning, AI vs ML vs Data Science, traditional programming vs ML paradigm, and real-world engineering applications. Types of Learning: Supervised, unsupervised, semi-supervised, and reinforcement learning. Models of Machine Learning: Geometric model, probabilistic models, logical models, grouping and grading models, parametric and non-parametric models. Introduction to Feature Engineering. Feature Transformation: Dimensionality reduction techniques- Principal Component Analysis (PCA); Linear Discriminant Analysis (LDA).

• Course Code: _____ • Semester / Year: _____

TABLE OF CONTENTS

1.1 Introduction & Scope of Machine Learning

Formal Definition: Machine Learning (Tom Mitchell, 1997)

A computer program is said to LEARN from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Classic Example (Spam Filter): Task T = Classifying incoming emails as spam or ham; Experience E = Observing historical labeled emails; Performance P = Percentage of emails correctly classified (Accuracy).

Machine Learning (ML) is a core subfield of Artificial Intelligence (AI) dedicated to the design and development of algorithms that can automatically learn patterns, discover underlying structure, and make highly accurate data-driven predictions from historical empirical data without being explicitly programmed with static deterministic rules.

Why Machine Learning? (The Scope of ML)

- **1. Infeasible Human Programming:** Complex real-world tasks such as facial recognition, continuous speech transcription, and natural language understanding contain millions of subtle variations and noise that cannot be captured by manual hand-crafted 'if-else' rules.
- **2. Dynamic and Adapting Environments:** In domains like financial fraud detection, cybersecurity intrusion defense, and personalized e-commerce recommendations, user behaviors and adversary strategies evolve constantly. ML models continuously adapt to fresh incoming telemetry.
- **3. Massive Big Data Exploitation:** Modern enterprises generate terabytes of high-dimensional multi-modal data. ML algorithms excel at discovering complex non-linear statistical correlations beyond human cognitive capacity.

Comparative Analysis: AI vs. ML vs. Data Science vs. Deep Learning

A fundamental conceptual requirement in university examinations is distinguishing between the overlapping domains of intelligent computing:

Domain	Core Definition & Purpose	Primary Methodology	Relationship & Hierarchy
Artificial Intelligence (AI)	The broad overarching science of engineering intelligent machines capable of mimicking human cognitive functions.	Symbolic logic, knowledge graphs, search algorithms, heuristic reasoning, and ML.	The broad superset containing Machine Learning as a core sub-discipline.

Domain	Core Definition & Purpose	Primary Methodology	Relationship & Hierarchy
Machine Learning (ML)	A subfield of AI focused on statistical algorithms that learn mathematical mapping functions directly from data.	Supervised, Unsupervised, Reinforcement Learning, Optimization, and Probability.	A subset of AI that provides the computational learning engine.
Deep Learning (DL)	A specialized subset of ML based on multi-layered Artificial Neural Networks (ANNs) inspired by the biological brain.	Deep architectures: CNNs, RNNs, Transformers, Backpropagation gradient descent.	A sub-branch of ML specialized for raw unstructured data (images, audio, text).
Data Science (DS)	An interdisciplinary field combining statistics, domain expertise, data engineering, visualization, and ML to extract actionable business insights.	Data wrangling, exploratory data analysis (EDA), statistical inference, SQL, and predictive ML models.	Intersects with AI/ML; applies ML as a tool to solve analytical business problems.

Traditional Programming vs. Machine Learning Paradigm

The fundamental shift in computational philosophy between classical programming and machine learning is illustrated below:

TRADITIONAL PROGRAMMING VS. MACHINE LEARNING PARADIGM

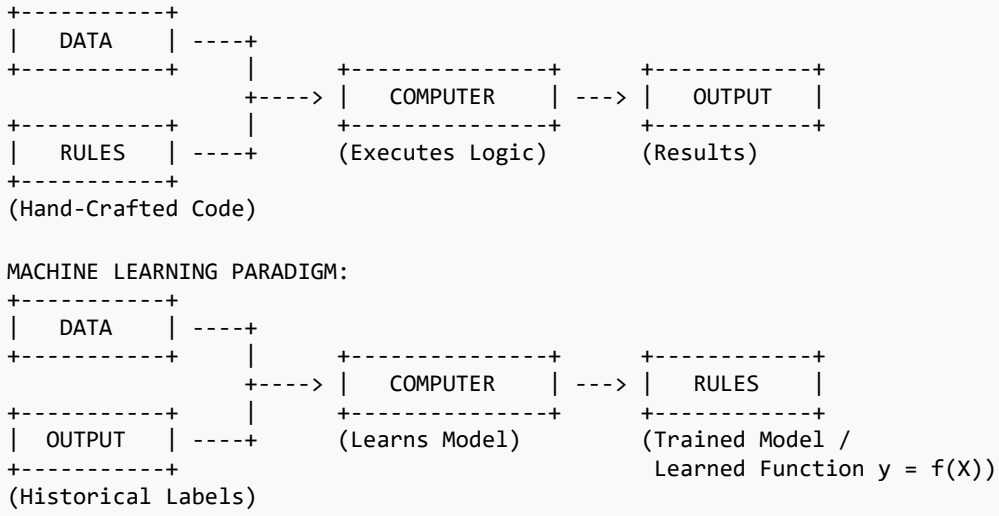

Figure: Classical software requires manual rule-crafting; ML automatically learns the mapping function f(X)

Real-World Engineering Applications of Machine Learning

- **1. Healthcare & Biomedical Engineering:** Automated tumor segmentation in MRI scans, early diabetic retinopathy detection from retinal imagery, protein folding prediction (AlphaFold), and personalized drug discovery.
- **2. Autonomous Transportation & Robotics:** Real-time object detection (YOLO), lane tracking, pedestrian path trajectory forecasting, and sensor fusion across LiDAR, Radar, and camera arrays.
- **3. Finance & Algorithmic Trading:** Real-time credit card fraud detection using anomaly detection, algorithmic high-frequency market making, credit scoring, and automated loan underwriting.
- **4. Natural Language Processing (NLP):** Machine translation (Google Translate), conversational AI and Large Language Models (LLMs), sentiment analysis, and automated document summarization.
- **5. Computer Vision & Security:** Facial recognition biometric access control, automated surveillance anomaly detection, optical character recognition (OCR), and industrial quality defect scanning.

1.2 Types of Machine Learning

Machine Learning paradigms are fundamentally categorized based on the nature of the learning feedback signal, the presence or absence of target supervision, and the operational environment.

1. Supervised Learning (Task-Driven)

In Supervised Learning, the algorithm is provided with a labeled training dataset consisting of input-output pairs: D = {(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)}, where x_i in R^d is the feature vector and y_i is the ground-truth target label. The objective is to learn an optimal mapping function y = f(x) that accurately generalizes to unseen test instances.

- **Two Sub-Branches of Supervised Learning:**
  - • **Regression:** The target variable y is a continuous real-valued numerical quantity (e.g., house price prediction, stock price forecasting, temperature estimation). Algorithms: Linear Regression, Polynomial Regression, Ridge/Lasso, SVR, Decision Trees.
  - • **Classification:** The target variable y is a discrete categorical class label (e.g., Spam vs Ham, Malignant vs Benign tumor). Algorithms: Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, Naive Bayes.

2. Unsupervised Learning (Data-Driven)

In Unsupervised Learning, the training dataset consists purely of unlabelled input feature vectors: D = {x_1, x_2, ..., x_n} without any target outcomes y. The algorithm must autonomously uncover inherent hidden patterns, clusters, density distributions, or low-dimensional manifolds in the data.

- **Core Sub-Branches:**
  - • **Clustering:** Grouping unlabeled instances into homogeneous clusters based on spatial proximity or density (e.g., K-Means, DBSCAN, Hierarchical Agglomerative Clustering, Gaussian Mixture Models).
  - • **Dimensionality Reduction:** Projecting high-dimensional data into a lower-dimensional subspace while preserving maximal variance or geometric structure (e.g., PCA, LDA, t-SNE, UMAP).
  - • **Association Rule Mining:** Discovering co-occurrence relationships among features in transaction databases (e.g., Apriori, FP-Growth).

3. Semi-Supervised Learning

In many practical engineering scenarios, acquiring raw unlabelled data is inexpensive (e.g., crawling millions of medical images), but labeling data requires costly human domain expertise. Semi-Supervised Learning leverages a small pool of labeled data combined with a massive pool of unlabelled data to train superior models.

- **Mechanisms:** Pseudo-Labeling / Self-Training: A model trained on the small labeled set predicts pseudo-labels for high-confidence unlabelled instances, which are then iteratively added to the training set.

4. Reinforcement Learning (Environment-Driven)

Reinforcement Learning (RL) involves an autonomous Agent interacting with a dynamic Environment in discrete time steps. The agent perceives current State s, executes Action a, and receives a scalar Reward r along with the new State s'. The agent learns an optimal Policy pi(a|s) to maximize cumulative long-term discounted rewards through trial and error.

Master Comparison Table: The 4 Paradigms of Machine Learning

Feature	Supervised Learning	Unsupervised Learning	Semi-Supervised Learning	Reinforcement Learning
Training Data	Fully Labeled (Features X + Target y)	Completely Unlabelled (Features X only)	Small Labeled set + Large Unlabelled set	No static dataset; dynamic environment interaction
Primary Objective	Predict target labels for new unseen data	Discover hidden clusters, patterns & manifolds	Boost accuracy using abundant unlabelled data	Learn optimal decision policy via trial & error

Feature	Supervised Learning	Unsupervised Learning	Semi-Supervised Learning	Reinforcement Learning
Feedback Signal	Direct error feedback (Loss between y and y_hat)	No feedback; measures intra-cluster variance	Direct feedback on labeled; pseudo-loss on unlabelled	Delayed scalar reward / penalty signals
Core Task Types	Regression & Classification	Clustering & Dimensionality Reduction	Classification with scarce labels	Control, Game Playing, Autonomous Robotics
Key Algorithms	Linear Regression, SVM, Decision Trees, KNN	K-Means, DBSCAN, PCA, GMM, Hierarchical	Self-Training, Co-Training, Label Propagation	Q-Learning, SARSA, Policy Gradients, Deep Q-Networks

1.3 Models of Machine Learning

Machine Learning models are fundamentally categorized according to their underlying mathematical formulation, geometric interpretation, probabilistic foundation, and structural flexibility.

1. Geometric Models

Geometric models represent instances as points in a continuous metric feature space R^d. Concepts, categories, and decision boundaries are defined geometrically as hyperplanes, decision surfaces, spheres, or polyhedra.

- **Distance & Similarity Metrics:**
  - Euclidean Distance (L2 Norm): d(x, z) = sqrt(sum((x_i - z_i)^2))
  - Manhattan Distance (L1 Norm): d(x, z) = sum(|x_i - z_i|)
  - Cosine Similarity: sim(x, z) = (x . z) / (||x|| * ||z||) (measures angular orientation, invariant to magnitude).
- **Representative Algorithms:** Support Vector Machines (SVM separates classes via maximum margin hyperplanes w^T x + b = 0), K-Nearest Neighbors (KNN), K-Means Clustering.

2. Probabilistic Models

Probabilistic models view learning as estimating underlying statistical probability distributions over data. Predictions are framed as computing conditional probabilities P(Y|X) using Bayes' Theorem:

- **Bayes' Rule Formulation:** P(Y|X) = [P(X|Y) * P(Y)] / P(X), where P(Y) is the Prior Probability, P(X|Y) is the Likelihood, P(X) is the Evidence, and P(Y|X) is the Posterior Probability.
- **Generative vs. Discriminative Models:**

- Generative Models: Model the joint distribution P(X, Y) = P(X|Y)P(Y). Can generate synthetic new data samples (e.g., Naive Bayes, Gaussian Mixture Models, Hidden Markov Models, GANs).
- Discriminative Models: Directly model the conditional posterior boundary P(Y|X) without modeling how data was generated (e.g., Logistic Regression, Linear Discriminant Analysis, Conditional Random Fields).

3. Logical Models

Logical models translate learning into structured declarative rules, logical propositions, and decision trees using Boolean conjunctions (AND), disjunctions (OR), and negations (NOT).

- **Properties:** Highly interpretable (white-box models); human-readable decision rules; invariant to monotonic feature scaling. Examples: Decision Trees (ID3, C4.5, CART), Rule Induction Algorithms (RIPPER).

4. Grouping vs. Grading Models

- **Grouping Models:** Partition the instance space into discrete, mutually exclusive segments or clusters, assigning an identical constant prediction to all instances falling within a group. Examples: Decision Trees (each leaf is a segment), K-Means Clustering.
- **Grading Models:** Assign a continuous score, probability, or ranked grade to instances across the feature space using a global mathematical function. Examples: Linear Regression, Logistic Regression, Support Vector Machines.

5. Parametric vs. Non-Parametric Models

Comparison Parameter	Parametric Models	Non-Parametric Models
Parameter Count	Fixed, finite number of parameters independent of training sample size N	Number of parameters grows dynamically with training dataset size N
Assumptions about Data	Strong functional form assumptions (e.g., linearity, normal distribution)	Minimal/no prior assumptions about underlying data distribution shape
Training Speed & Memory	Fast training; compact memory (only stores weights theta)	Slower training/inference; requires storing training data instances
Overfitting / Flexibility	Simpler; higher bias; lower variance; less prone to overfitting	Highly flexible; low bias; higher risk of overfitting complex noise
Representative Algorithms	Linear Regression, Logistic Regression, Linear SVM, Naive Bayes	K-Nearest Neighbors (KNN), Decision Trees, RBF-Kernel SVM

1.4 Introduction to Feature Engineering

Definition: Feature Engineering

Feature Engineering is the process of using domain knowledge to extract, construct, select, and transform raw variables into refined feature representations that maximize the predictive performance and convergence speed of machine learning algorithms.

Key Steps in Feature Engineering Pipeline

- **1. Handling Missing Data:** Imputation using Mean/Median (for numerical data with outliers), Mode (for categorical features), or predictive imputation using KNN/MICE.
- **2. Categorical Encoding:**
  - One-Hot Encoding: Converts nominal categorical values without intrinsic order (e.g., Red, Blue, Green) into binary columns (prevents artificial numerical ordering bias).
  - Label / Ordinal Encoding: Maps ordered categories to sequential integers (e.g., Low=0, Medium=1, High=2).
- **3. Feature Scaling (Crucial for Distance-Based Algorithms):**
  - Min-Max Normalization: Rescales features strictly into [0, 1] range: x_norm = (x - x_min) / (x_max - x_min). Sensitive to outliers.
  - Standardization (Z-Score Scaling): Rescales features to have mean mu = 0 and standard deviation sigma = 1: z = (x - mu) / sigma. Robust to outliers and ideal for PCA, Gradient Descent, and SVM.

1.5 Dimensionality Reduction: PCA and LDA

The Curse of Dimensionality

As the number of features d increases, the volume of the feature space grows exponentially (V proportional to r^d). Consequently, available training data becomes extremely sparse in high-dimensional space. Distance metrics (Euclidean distance) break down because all pairwise distances become nearly equal (distance concentration effect), computational costs surge, and models suffer catastrophic overfitting.

1. Principal Component Analysis (PCA) — Unsupervised

Principal Component Analysis (PCA) is an unsupervised linear dimensionality reduction technique that finds orthogonal axes (Principal Components) along which the variance of the data is maximized, minimizing reconstruction information loss.

Step-by-Step Mathematical Derivation of PCA

- **Step 1 (Standardization):** Standardize the input dataset X of dimension (n x d) so each feature has mean 0 and variance 1.
- **Step 2 (Covariance Matrix Calculation):** Compute the d x d sample Covariance Matrix Sigma: Sigma = (1 / n) * X^T * X.
- **Step 3 (Eigen-Decomposition):** Solve the characteristic equation to find Eigenvalues lambda and Eigenvectors v: Sigma * v_i = lambda_i * v_i.
- **Step 4 (Sorting & Component Selection):** Sort the eigenvectors in descending order of their corresponding eigenvalues. The eigenvalue lambda_i represents the variance captured by eigenvector v_i. Select the top k eigenvectors to form projection matrix W of size (d x k).
- **Explained Variance Ratio Formula:** Explained Variance = lambda_i / sum(lambda_j for all j).
- **Step 5 (Projection):** Transform the original dataset X into the lower-dimensional subspace Z: Z = X * W (where Z has dimension n x k, k << d).

2. Linear Discriminant Analysis (LDA) — Supervised

Linear Discriminant Analysis (LDA), also known as Fisher's Discriminant Analysis, is a supervised dimensionality reduction technique that seeks a linear projection that maximizes class separability while minimizing intra-class spread.

Mathematical Formulation of LDA (Fisher's Criterion)

- **Between-Class Scatter Matrix (S_B):** Measures the separation between the means of different classes: S_B = sum(N_c * (mu_c - mu) * (mu_c - mu)^T), where mu_c is class c mean and mu is global mean.
- **Within-Class Scatter Matrix (S_W):** Measures the variance/spread of samples within their respective classes: S_W = sum_c sum_{x in c} (x - mu_c) * (x - mu_c)^T.
- **Fisher's Optimization Objective:** Find projection vector w that maximizes Fisher's Criterion: J(w) = (w^T * S_B * w) / (w^T * S_W * w).
- **Solution via Generalized Eigen-Decomposition:** Solving the optimization yields: (S_W^(-1) * S_B) * w = lambda * w.
- **Dimensionality Constraint:** For a dataset with C classes, the maximum number of discriminant components LDA can extract is at most C - 1 (since S_B has rank at most C - 1).

Master Comparison Table: PCA vs. LDA

Comparison Parameter	Principal Component Analysis (PCA)	Linear Discriminant Analysis (LDA)
Learning Paradigm	Unsupervised (Ignores class labels completely)	Supervised (Requires ground-truth class labels)
Optimization Goal	Maximizes total data variance along orthogonal axes	Maximizes between-class scatter while minimizing within-class scatter
Mathematical Formulation	Eigen-decomposition of Covariance Matrix Sigma	Generalized eigen-decomposition of (S_W^(-1) * S_B)
Maximum Reduced Dimensions	Can reduce to any k <= d dimensions	Strictly bounded to k <= C - 1 (where C is number of classes)
Sensitivity to Outliers	Sensitive to extreme outliers (affects variance)	Sensitive to outliers and non-Gaussian class distributions
Ideal Use Cases	Unsupervised data visualization, noise filtering, compression	Supervised classification feature preprocessing (e.g., Face Recognition)

1.6 High-Yield University Exam Question Bank

Part A: Short Answer Questions (2 to 5 Marks)

Q1. State Tom Mitchell's definition of Machine Learning with an example. [2-3 Marks]

- **Answer:** A computer program is said to learn from experience E with respect to task T and performance measure P, if its performance at task T, measured by P, improves with experience E. Example in Spam filtering: T = classifying spam emails, E = past labeled emails, P = classification accuracy percentage.

Q2. Differentiate between Parametric and Non-Parametric ML models. [3 Marks]

- **Answer:** Parametric models assume a fixed functional form with a constant number of parameters independent of data size (e.g., Linear Regression, Logistic Regression; faster, simpler, high bias). Non-parametric models make no rigid assumptions; parameters grow with training data size (e.g., KNN, Decision Trees; flexible, low bias, higher memory).

Q3. What is the Curse of Dimensionality? [3 Marks]

- **Answer:** As feature dimensions increase, space volume grows exponentially, making data points extremely sparse. Pairwise Euclidean distances converge to nearly identical values, degrading distance-based algorithms (KNN, K-Means) and causing severe model overfitting.

Part B: Long Answer Questions (10 to 15 Marks Outline Guides)

- **Q4. Compare AI vs. ML vs. Data Science. Explain the difference between Traditional Programming and Machine Learning paradigms with neat diagrams. [10 Marks]**
- **Q5. Explain the four types of Machine Learning (Supervised, Unsupervised, Semi-Supervised, Reinforcement) with real-world examples, mathematical formulations, and comparison tables. [12 Marks]**
- **Q6. Derive Principal Component Analysis (PCA) step-by-step from covariance matrix calculation to eigen-decomposition and projection. [12 Marks]**
- **Q7. Explain Linear Discriminant Analysis (LDA) and Fisher's Criterion. Compare PCA vs. LDA in detail. [10 Marks]**

1.7 Unit Summary & Key Takeaways

- Machine learning algorithms automatically learn mapping functions y = f(X) from empirical experience (data) without explicit static programming.
- ML types include Supervised (Regression/Classification), Unsupervised (Clustering/PCA), Semi-Supervised, and Reinforcement Learning (Agent/Reward).
- ML models are categorized into Geometric (distance hyperplanes), Probabilistic (Bayes rule), Logical (Decision trees), Grouping vs Grading, and Parametric vs Non-parametric.
- Feature engineering involves imputation, encoding, and scaling (Min-Max vs Z-Score standardization).
- PCA is an unsupervised technique maximizing variance via covariance eigen-decomposition.
- LDA is a supervised technique maximizing Fisher's class separability criterion J(w) = (w^T S_B w) / (w^T S_W w) bounded to at most C-1 dimensions.