Fundamentals of Machine Learning (4341603) - Summer 2024 Solution

Milav Dabgar

June 15, 2024

Question 1(a) [3 marks]

Define Machine Learning using suitable example?

Solution

Machine Learning is a subset of artificial intelligence that enables computers to learn and make decisions from data without being explicitly programmed for every task.

Table 1. Key Components of Machine Learning

Component	Description
Data	Input information used for training
Algorithm	Mathematical model that learns patterns
Training	Process of teaching the algorithm
Prediction	Output based on learned patterns

Example: Email spam detection system learns from thousands of emails labeled as ”spam” or ”not spam” to automatically classify new emails.

Mnemonic

“Data Drives Decisions - Data trains algorithms to make intelligent decisions”

Question 1(b) [4 marks]

Explain the process of machine learning with the help of schematic representation

Solution

The machine learning process involves systematic steps from data collection to model deployment.

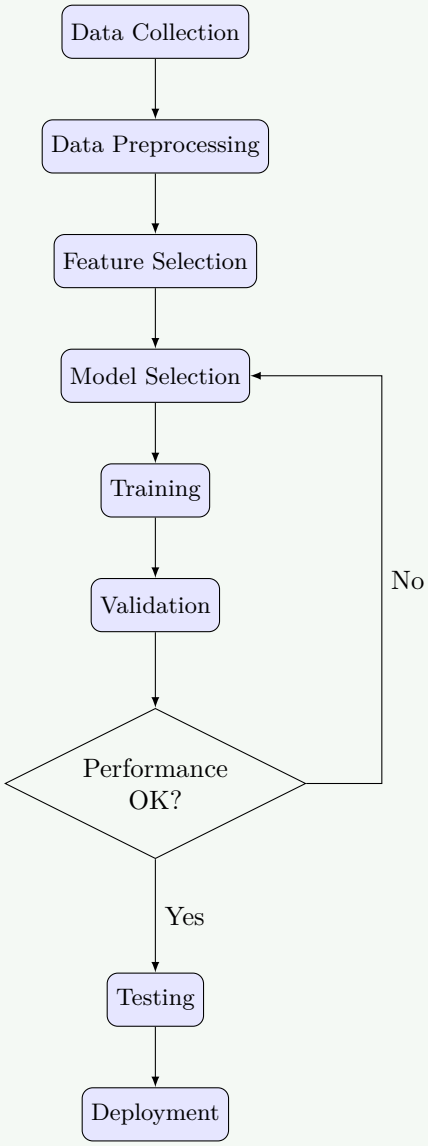


Figure 1. Machine Learning Process

Process Steps:

- **Data Collection**: Gathering relevant dataset
- **Preprocessing**: Cleaning and preparing data
- **Training**: Teaching algorithm using training data
- **Validation**: Testing model performance
- **Deployment**: Using model for real predictions

Mnemonic

"Computers Can Truly Think - Collect, Clean, Train, Test"

Question 1(c) [7 marks]

Explain different types of machine learning with suitable application.

Solution

Machine learning algorithms are categorized based on learning approach and available data.

Table 2. Types of Machine Learning

Type	Learning Method	Data Requirement	Example Application
Supervised	Uses labeled data	Input-output pairs	Email classification
Unsupervised	Finds hidden patterns	Only input data	Customer segmentation
Reinforcement	Learns through rewards	Environment feedback	Game playing AI

Applications:

- **Supervised Learning**: Medical diagnosis, image recognition, fraud detection
- **Unsupervised Learning**: Market research, anomaly detection, recommendation systems
- **Reinforcement Learning**: Autonomous vehicles, robotics, strategic games

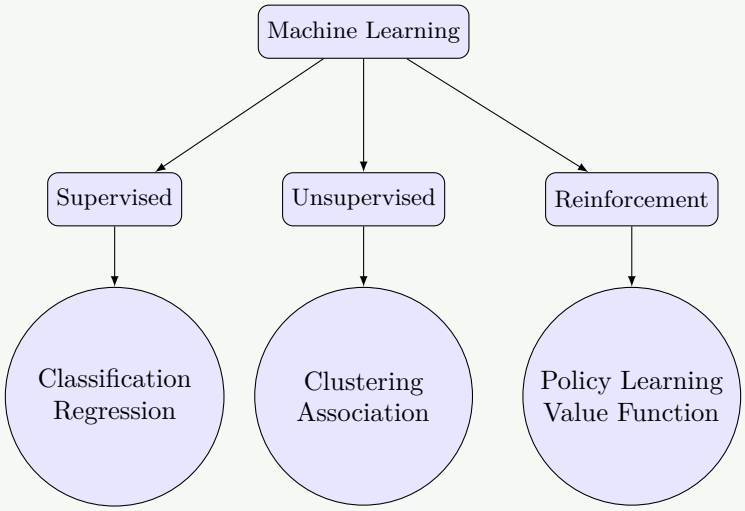


Figure 2. Types of Machine Learning

Mnemonic

"Students Usually Remember - Supervised, Unsupervised, Reinforcement"

Question 1(c) OR [7 marks]

What are various issues with machine learning? List three problems that are not to be solved using machine learning.

Solution

Table 3. Machine Learning Issues

Issue Category	Description	Impact
Data Quality	Incomplete, noisy, biased data	Poor model performance
Overfitting	Model memorizes training data	Poor generalization
Computational	High processing requirements	Resource constraints
Interpretability	Black box models	Lack of transparency

Problems NOT suitable for ML:

1. **Simple rule-based tasks** - Basic calculations, simple if-then logic where rules are explicit.

2. **Ethical decisions** - Moral judgments requiring human values and empathy (e.g., judicial sentencing).
3. **Creative expression** - Original artistic creation requiring genuine human emotion and intent.

Other Issues:

- **Privacy concerns**: Sensitive data handling
- **Bias propagation**: Unfair algorithmic decisions
- **Feature selection**: Choosing relevant input variables

Mnemonic

"Data Drives Quality - Data quality directly affects model quality"

Question 2(a) [3 marks]

Give a summarized view of different types of data in a typical machine learning problem.

Solution

Table 4. Data Types in Machine Learning

Data Type	Description	Example
Numerical	Quantitative values	Age: 25, Height: 170cm
Categorical	Discrete categories	Color: Red, Blue, Green
Ordinal	Ordered categories	Rating: Poor, Good, Excellent
Binary	Two possible values	Gender: Male/Female

Characteristics:

- **Structured**: Organized in tables (databases, spreadsheets)
- **Unstructured**: Images, text, audio files
- **Time-series**: Data points over time

Mnemonic

"Numbers Count Better Than Words - Numerical, Categorical, Binary, Text"

Question 2(b) [4 marks]

Calculate variance for both attributes. Determine which attribute is spread out around mean.

Solution

Given Data:

- Attribute 1: 32, 37, 47, 50, 59
- Attribute 2: 48, 40, 41, 47, 49

Calculations:

Attribute 1:

- Mean $= (32 + 37 + 47 + 50 + 59)/5 = 225/5 = 45$
- Variance $= [(32 - 45)^2 + (37 - 45)^2 + (47 - 45)^2 + (50 - 45)^2 + (59 - 45)^2]/5$
- Variance $= [169 + 64 + 4 + 25 + 196]/5 = 458/5 = 91.6$

Attribute 2:

- Mean $= (48 + 40 + 41 + 47 + 49)/5 = 225/5 = 45$
- Variance $= [(48 - 45)^2 + (40 - 45)^2 + (41 - 45)^2 + (47 - 45)^2 + (49 - 45)^2]/5$
- Variance $= [9 + 25 + 16 + 4 + 16]/5 = 70/5 = 14$

Result: Attribute 1 (variance = 91.6) is more spread out than Attribute 2 (variance = 14).

Mnemonic

"Higher Variance Shows Spread - Greater variance indicates more dispersion"

Question 2(c) [7 marks]

List Factors that lead to data quality issue. How to handle outliers and missing values.

Solution

Table 5. Data Quality Issues

Factor	Cause	Solution
Incompleteness	Missing data collection	Imputation techniques
Inconsistency	Different data formats	Standardization
Inaccuracy	Human/sensor errors	Validation rules
Noise	Random variations	Filtering methods

Handling Outliers:
- **Detection**: Statistical methods (Z-score, IQR)
- **Treatment**: Remove, transform, or cap extreme values
- **Visualization**: Box plots, scatter plots

Handling Missing Values:
- **Deletion**: Remove incomplete records (rows with missing data)
- **Imputation**: Fill with mean, median, or mode
- **Prediction**: Use ML to predict missing values

Code Example:

```
# Handle missing values
df.fillna(df.mean())  # Mean imputation
df.dropna()           # Remove missing rows
```

Mnemonic

"Clean Data Makes Models - Clean data produces better models"

Question 2(a) OR [3 marks]

Give different machine learning activities.

Solution

Table 6. Machine Learning Activities

Activity	Purpose	Example
Data Collection	Gather relevant information	Surveys, sensors, databases
Data Preprocessing	Clean and prepare data	Remove noise, handle missing values
Feature Engineering	Create meaningful variables	Extract features from raw data
Model Training	Teach algorithm patterns	Use training dataset
Model Evaluation	Assess performance	Test accuracy, precision, recall
Model Deployment	Put model into production	Web services, mobile apps

Key Activities:
- **Exploratory Data Analysis**: Understanding data patterns
- **Hyperparameter Tuning**: Optimizing model settings
- **Cross-validation**: Robust performance assessment

Mnemonic

"Data Models Perform Excellently - Data preparation, Model building, Performance evaluation, Execution"

Question 2(b) OR [4 marks]

Calculate mean and median of the following numbers: 12,15,18,20,22,24,28,30

Solution

Given numbers: 12, 15, 18, 20, 22, 24, 28, 30
Mean Calculation: Mean $= (12 + 15 + 18 + 20 + 22 + 24 + 28 + 30)/8 = 169/8 = 21.125$
Median Calculation:
- Numbers are already sorted: 12, 15, 18, 20, 22, 24, 28, 30
- Even count (8 numbers)
- Median = (4th number + 5th number)/2 $= (20 + 22)/2 = 21$

Table 7. Statistical Summary

Measure	Value	Description
Mean	21.125	Average value
Median	21	Middle value
Count	8	Total numbers

Mnemonic

"Middle Makes Median - Middle value gives median"

Question 2(c) OR [7 marks]

Write a short note on dimensionality reduction and feature subset selection in context with data preprocessing.

Solution

Dimensionality Reduction removes irrelevant features and reduces computational complexity while preserving important information.

Table 8. Dimensionality Reduction Techniques

Technique	Method	Use Case
PCA	Principal Component Analysis	Linear reduction
LDA	Linear Discriminant Analysis	Classification tasks
t-SNE	Non-linear embedding	Visualization
Feature Selection	Select important features	Reduce overfitting

Feature Subset Selection Methods:

- **Filter Methods**: Statistical tests, correlation analysis
- **Wrapper Methods**: Forward/backward selection
- **Embedded Methods**: LASSO, Ridge regression

Benefits:

- **Computational Efficiency**: Faster training and prediction
- **Storage Reduction**: Less memory requirements
- **Noise Reduction**: Remove irrelevant features
- **Visualization**: Enable 2D/3D plotting

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
reduced_data = pca.fit_transform(data)
```

Mnemonic

"Reduce Features, Improve Performance - Fewer features often lead to better models"

Question 3(a) [3 marks]

Does bias affect the performance of the ML model? Explain briefly.

Solution

Yes, bias significantly affects ML model performance by creating systematic errors in predictions.

Table 9. Types of Bias

Bias Type	Description	Impact
Selection Bias	Non-representative data	Poor generalization
Confirmation Bias	Favoring expected results	Skewed conclusions
Algorithmic Bias	Model assumptions	Unfair predictions

Effects on Performance:

- **Underfitting**: High bias leads to oversimplified models
- **Poor Accuracy**: Systematic errors reduce overall performance
- **Unfair Decisions**: Biased models discriminate against groups

Mnemonic

"Bias Breaks Better Performance - Bias reduces model effectiveness"

Question 3(b) [4 marks]

Compare cross-validation and bootstrap sampling

Solution

Table 10. Cross-validation vs Bootstrap Sampling

Aspect	Cross-validation	Bootstrap Sampling
Method	Split data into folds	Sample with replacement
Data Usage	Uses all data	Creates multiple samples
Purpose	Model evaluation	Estimate uncertainty
Overlap	No overlap between sets	Allows duplicate samples

Key Differences:

- **Cross-validation**: Divides data into k equal parts. Trains on k-1 parts, tests on 1 part. Repeats k times.
- **Bootstrap Sampling**: Creates random samples with replacement. Generates multiple datasets of same size.

Mnemonic

"Cross Checks, Bootstrap Builds - Cross-validation checks performance, Bootstrap builds confidence"

Question 3(c) [7 marks]

Confusion Matrix Calculation and Metrics

Solution

Given Information:

- True Positive (TP): 83
- False Positive (FP): 7
- False Negative (FN): 5
- True Negative (TN): 5

	Predicted Buy	Predicted No Buy
Actually Buy	83 (TP)	5 (FN)
Actually No Buy	7 (FP)	5 (TN)

Calculations:

a) Error Rate: Error Rate = $(FP + FN)/Total = (7 + 5)/100 = 0.12 = 12\%$

b) Precision: Precision = $TP/(TP + FP) = 83/(83 + 7) = 83/90 = 0.922 = 92.2\%$

c) Recall: Recall = $TP/(TP + FN) = 83/(83 + 5) = 83/88 = 0.943 = 94.3\%$

d) F-measure: F-measure = $2\times(Precision\times Recall)/(Precision+Recall)$ F-measure = $2\times(0.922\times0.943)/(0.922+0.943) = 0.932 = 93.2\%$

Mnemonic

"Perfect Recall Finds Everyone - Precision measures accuracy, Recall finds all positives"

Question 3(a) OR [3 marks]

Define in brief: a) Target function b) Cost function c) Loss Function

Solution

Table 11. Function Definitions

Function	Definition	Purpose
Target Function	Ideal mapping from input to output	What we want to learn
Cost Function	Measures overall model error	Evaluate total performance
Loss Function	Measures error for single prediction	Individual prediction error

Relationship: Cost function is typically the average of loss functions across all training examples.

Mnemonic

"Target Costs Less - Target function is ideal, Cost function measures overall error, Loss function measures individual error"

Question 3(b) OR [4 marks]

Explain balanced fit, underfit and overfit

Solution

Table 12. Model Fitting Types

Fit Type	Training Error	Validation Error	Characteristics
Underfit	High	High	Too simple model
Balanced Fit	Low	Low	Optimal complexity
Overfit	Very Low	High	Too complex model

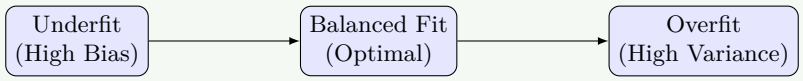


Figure 3. Model Complexity Spectrum

Solutions:
- **Underfit**: Increase model complexity, add features
- **Overfit**: Regularization, cross-validation, more data

Mnemonic

"Balance Brings Best Results - Balanced models perform best on new data"

Question 4(a) [3 marks]

Give classification learning steps.

Solution

Table 13. Classification Learning Steps

Step	Description	Purpose
Data Collection	Gather labeled examples	Provide training material
Preprocessing	Clean and prepare data	Improve data quality
Feature Selection	Choose relevant attributes	Reduce complexity
Model Training	Learn from training data	Build classifier
Evaluation	Test model performance	Assess accuracy
Deployment	Use for new predictions	Practical application

Mnemonic

"Data Preparation Facilitates Model Excellence - Data prep, Feature selection, Model training, Evaluation"

Question 4(b) [4 marks]

Linear Relationship Calculation

Solution

Given Data: Hours (X) vs Exam Score (Y)
Linear Regression Calculation:
Step 1: Calculate means
- $\bar{X} = (2+3+4+5+6)/5 = 4$
- $\bar{Y} = (85+80+75+70+60)/5 = 74$

Step 2: Calculate slope (b)
- Numerator $= \sum(X-\bar{X})(Y-\bar{Y}) = -60$
- Denominator $= \sum(X-\bar{X})^2 = 10$
- $b = -60/10 = -6$

Step 3: Calculate intercept (a)
- $a = \bar{Y} - b \times \bar{X} = 74 - (-6) \times 4 = 74 + 24 = 98$

Linear Equation: $Y = 98 - 6X$
Interpretation: For every additional hour of smartphone use, exam score decreases by 6 points.

Mnemonic

"More Phone, Less Score - Negative correlation between phone use and grades"

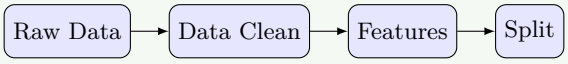
Question 4(c) [7 marks]

Explain classification steps in detail

Solution

Classification is a supervised learning process that assigns input data to predefined categories.
Detailed Classification Steps:
1. Problem Definition
- Define classes and objectives
- Identify input features and target variable

2. Data Collection and Preparation



3. Feature Engineering

- **Feature Selection**: Choose relevant attributes
- **Normalization**: Scale features to similar ranges

4. **Model Selection and Training**

Table 14. Common Classification Algorithms

Algorithm	Best For	Advantages
Decision Tree	Interpretable rules	Easy to understand
SVM	High-dimensional data	Good generalization
Neural Networks	Complex patterns	High accuracy

5. **Model Evaluation**
- **Confusion Matrix**: Detailed performance analysis
- **Metrics**: Accuracy, Precision, Recall, F1-score

6. **Final Evaluation and Deployment**
- Test on unseen data
- Deploy model for production use

Mnemonic

"Proper Data Modeling Evaluates Performance Thoroughly - Problem definition, Data prep, Modeling, Evaluation, Performance testing, Tuning"

Question 4(a) OR [3 marks]

Does the choice of the k value influence the performance of the KNN algorithm? Explain briefly

Solution

Yes, the k value significantly influences KNN algorithm performance.

Table 15. K Value Impact

K Value	Effect	Performance
Small K (k=1)	Sensitive to noise	High variance, low bias
Medium K	Balanced decisions	Optimal performance
Large K	Smooth boundaries	Low variance, high bias

Selection Strategy: Use cross-validation to find optimal k, often starting with $k = \sqrt{n}$.

Mnemonic

"Small K Varies, Large K Smooths - Small k creates variance, large k creates smooth boundaries"

Question 4(b) OR [4 marks]

Define Support Vectors in the SVM model.

Solution

Support Vectors are the critical data points closest to the decision boundary (hyperplane).

Table 16. Support Vector Characteristics

Aspect	Description	Importance
Location	Closest points to hyperplane	Define decision boundary
Distance	Equal distance from boundary	Maximize margin
Role	Support the hyperplane	Determine optimal separation

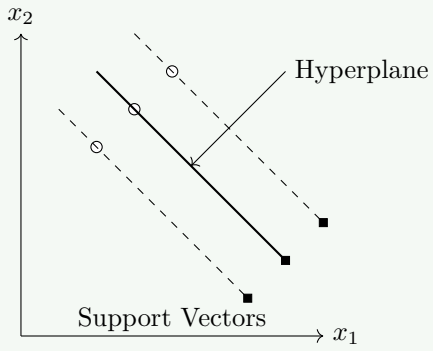


Figure 4. SVM Hyperplane and Support Vectors

Mnemonic

"Support Vectors Support Decisions - These vectors support the decision boundary"

Question 4(c) OR [7 marks]

Explain logistic regression in detail.

Solution

Logistic Regression is a statistical method used for binary classification.
Mathematical Foundation: Sigmoid Function: $\sigma(z) = 1/(1 + e^{-z})$ where $z = \beta_0 + \beta_1 x_1 + ...$

Table 17. Linear vs Logistic Regression

Aspect	Linear Regression	Logistic Regression
Output	Continuous values	Probabilities (0-1)
Function	Linear	Sigmoid (S-curve)
Purpose	Prediction	Classification
Error	Mean Squared Error	Log-likelihood

Key Components:
- **Logistic Function**: S-shaped curve mapping values to [0,1].
- **Decision Rule**: If $P(y = 1|x) > 0.5$, classify as positive.
- **Training**: Uses Maximum Likelihood Estimation.

Applications: Medical diagnosis, Email spam detection, Credit approval.

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train, y_train)
probabilities = model.predict_proba(X_test)
```

Mnemonic

"Sigmoid Squashes Infinite Input - Sigmoid function converts any real number to probability"

Question 5(a) [3 marks]

Write a short note on Matplotlib python library.

Solution

Matplotlib is a comprehensive Python library for creating visualizations.

Table 18. Matplotlib Key Features

Feature	Purpose	Example
Pyplot	MATLAB-like interface	Line plots
Formats	Save in various formats	PNG, PDF, SVG
Subplots	Multiple plots in one figure	Grid layouts

Basic Usage:

```
import matplotlib.pyplot as plt
plt.plot(x, y)
plt.show()
```

Mnemonic

"Matplotlib Makes Pretty Plots - Essential tool for data visualization"

Question 5(b) [4 marks]

K-means clustering for two-dimensional data

Solution

Given Points: Two groups clearly separated. Group 1: (2,3) to (8,3). Group 2: (25,20) to (30,20).
Algorithm Steps: Step 1: Initialize centroids C1 = (4, 3), C2 = (27, 20)
Step 2: Assign points Points (2,3)...(8,3) are closer to C1. Points (25,20)...(30,20) are closer to C2.
Step 3: Update centroids New C1 = Average of Group 1 = (5, 3) New C2 = Average of Group 2 = (27.5, 20)
Final Clusters: Cluster 1: Left group points. Cluster 2: Right group points.

Mnemonic

"Centroids Attract Nearest Neighbors - Points join closest centroid"

Question 5(c) [7 marks]

Give functions and its use of Scikit-learn

Solution

a) Data Preprocessing:
- `StandardScaler()`: Normalize features
- `train_test_split()`: Split dataset

b) Model Selection:
- `GridSearchCV()`: Hyperparameter tuning
- `cross_val_score()`: Cross-validation

c) Model Evaluation:
- `accuracy_score()`: Overall accuracy
- `confusion_matrix()`: Error analysis
- `classification_report()`: Comprehensive metrics

```
from sklearn.metrics import classification_report
print(classification_report(y_true, y_pred))
```

Mnemonic

"Preprocess, Select, Evaluate - Complete ML workflow in Scikit-learn"

Question 5(a) OR [3 marks]

List out the major features of Numpy.

Solution

NumPy is fundamental for scientific computing.
- **N-dimensional Arrays**: Efficient array objects
- **Broadcasting**: Operations on different sized arrays
- **Linear Algebra**: Matrix operations
- **Random Numbers**: Statistical simulations

Mnemonic

"Numbers Need Numpy's Power - Essential for numerical computations"

Question 5(b) OR [4 marks]

K-means clustering for one-dimensional data

Solution

Dataset: 1,2,4,5,7,8,10,11,12,14,15,17
K-means (k=3):
- **Cluster 1**: 1, 2, 4, 5 (Centroid $\approx$ 3)
- **Cluster 2**: 7, 8, 10, 11, 12 (Centroid $\approx$ 9.6)
- **Cluster 3**: 14, 15, 17 (Centroid $\approx$ 15.3)

Mnemonic

"Groups Gather by Distance - Similar points form natural clusters"

Question 5(c) OR [7 marks]

Give function and its use of Pandas library

Solution

a) Preprocessing:
- `read_csv()`: Load data
- `head()`, `tail()`: View data

b) Inspection:
- `info()`: Data types, memory
- `describe()`: Statistical summary
- `isnull()`: Check missing values

c) Cleaning:
- `dropna()`, `fillna()`: Handle missing data
- `groupby()`: Aggregate data

```
df = pd.read_csv('data.csv')
print(df.describe())
```

Mnemonic

"Pandas Processes Data Perfectly - Comprehensive data manipulation tool"