Fundamentals of Machine Learning (4341603) - Winter 2024 Solution

Milav Dabgar

November 28, 2024

Question 1(a) [3 marks]

Describe human learning in brief.

Solution

Human learning is the process by which humans acquire knowledge, skills, and behaviors through experience, practice, and instruction.

Table 1. Human Learning Process

Aspect	Description
Observation	Gathering information from environment
Experience	Learning through trial and error
Practice	Repetition to improve skills
Memory	Storing and retrieving information

- **Learning Types**: Visual, auditory, kinesthetic learning styles.
- **Feedback Loop**: Humans learn from mistakes and successes.
- **Adaptation**: Ability to apply knowledge to new situations.

Mnemonic

"Observe, Experience, Practice, Memory, Adapt (OEPMA)"

Question 1(b) [4 marks]

Differentiate: Supervised Learning v/s Unsupervised Learning

Solution

Table 2. Supervised vs Unsupervised Learning

Parameter	Supervised Learning	Unsupervised Learning
Training Data	Labeled data (input-output pairs)	Unlabeled data (only inputs)
Goal	Predict output for new inputs	Find hidden patterns
Examples	Classification, Regression	Clustering, Association
Feedback	Direct feedback available	No direct feedback

- **Supervised**: Teacher guides learning with correct answers.
- **Unsupervised**: Self-discovery of patterns without guidance.

Mnemonic

"SL-Labels, UL-Unknown"

Question 1(c) [7 marks]

List out machine learning activities. Explain each in detail.

Solution

Table 3. Machine Learning Activities

Activity	Purpose	Description
Data Collection	Gather raw data	Collecting relevant data from various sources
Data Preprocessing	Clean and prepare data	Handling missing values, normalization
Feature Selection	Choose important features	Selecting relevant attributes for learning
Model Training	Build learning model	Training algorithm on prepared dataset
Model Evaluation	Assess performance	Testing model accuracy and effectiveness
Model Deployment	Put model to use	Implementing model in real-world applications

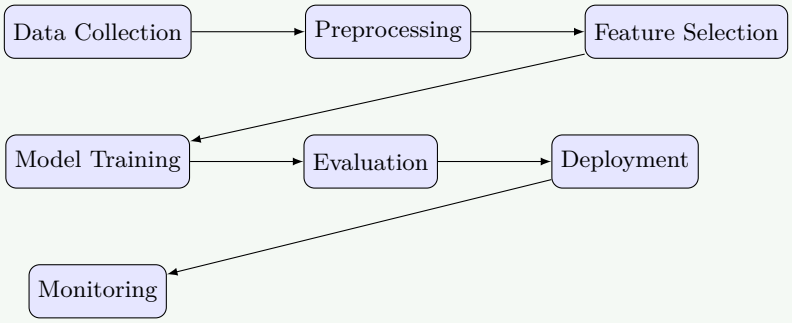

Figure 1. Machine Learning Activity Flow

- **Iterative Process**: Activities repeat for model improvement.
- **Quality Control**: Each step ensures better model performance.

Mnemonic

"Collect, Preprocess, Feature, Train, Evaluate, Deploy, Monitor (CPFTEDM)"

Question 1(c) OR [7 marks]

Find mean, median, and mode for the following data: 1, 1, 1, 2, 4, 5, 5, 6, 6, 7, 7, 7, 7, 8, 9, 10, 11

Solution

Data: 1, 1, 1, 2, 4, 5, 5, 6, 6, 7, 7, 7, 7, 8, 9, 10, 11 (Sorted, N=17)

Table 4. Data Analysis

Statistic	Formula	Calculation	Result
Mean	Sum/Count	100 / 17	5.88
Median	Middle value	9th position	6
Mode	Most frequent	Value 7 (4 times)	7

Step-by-step Calculation:

- **Count (N)**: 17 values
- **Sum**: $1+1+1+2+4+5+5+6+6+7+7+7+7+8+9+10+11=100$
- **Mean**: $100/17=5.88$
- **Median**: Odd number of values, so $(N+1)/2=9$th value. The 9th value in the sorted list is 6.
- **Mode**: The number 7 appears most frequently (4 times).

Mnemonic

"Mean=Average, Median=Middle, Mode=Most frequent (MMM)"

Question 2(a) [3 marks]

Write down steps to use hold out method for model training.

Solution

Table 5. Hold Out Method Steps

Step	Action	Purpose
1	Split dataset (70-80% train, 20-30% test)	Separate data for training and evaluation
2	Train model on training set	Build learning algorithm
3	Test model on testing set	Evaluate model performance

- **Random Split**: Ensure representative distribution in both sets.
- **No Overlap**: Testing data never used in training.

Mnemonic

"Split, Train, Test (STT)"

Question 2(b) [4 marks]

Explain structure of confusion matrix.

Solution

Confusion Matrix Structure

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 6. Confusion Matrix Layout

Components Explanation:

- **TP**: Correctly predicted positive cases.
- **TN**: Correctly predicted negative cases.

- **FP**: Incorrectly predicted as positive (Type I error).
- **FN**: Incorrectly predicted as negative (Type II error).

Performance Metrics:

- **Accuracy** = $(TP + TN)/(TP + TN + FP + FN)$
- **Precision** = $TP/(TP + FP)$

Mnemonic

"TPFN-FPTN for matrix positions"

Question 2(c) [7 marks]

Define data pre-processing. Explain various methods used in data pre-processing.

Solution

Data pre-processing is the technique of preparing raw data by cleaning, transforming, and organizing it for machine learning algorithms.

Table 7. Data Pre-processing Methods

Method	Purpose	Techniques
Data Cleaning	Remove noise/inconsistencies	Handle missing values, remove duplicates
Data Transformation	Convert data format	Normalization, standardization
Data Reduction	Reduce dataset size	Feature selection, dimensionality reduction
Data Integration	Combine multiple sources	Merge datasets, resolve conflicts

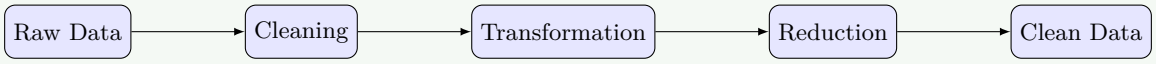

Figure 2. Preprocessing Steps

Key Techniques:

- **Missing Values**: Use mean, median, or mode for imputation.
- **Outliers**: Detect and handle extreme values.
- **Feature Scaling**: Normalize data to same scale.

Mnemonic

"Clean, Transform, Reduce, Integrate (CTRI)"

Question 2(a) OR [3 marks]

Explain histogram with suitable example.

Solution

A **Histogram** is a graphical representation showing the frequency distribution of numerical data by dividing it into bins.

Table 8. Histogram Components

Component	Description
X-axis	Data ranges (bins)
Y-axis	Frequency of occurrence
Bars	Height represents frequency

Example: Student marks distribution.
- Bins: 0-20, 21-40, 41-60, 61-80, 81-100.
- Heights show number of students in each range.

Mnemonic

"Bins, Axes, Range (BAR)"

Question 2(b) OR [4 marks]

Relate the appropriate data type of following examples: i) Gender of a person ii) Rank of students iii) Price of a home iv) Color of a flower

Solution

Table 9. Data Types Classification

Example	Data Type	Characteristics
Gender of person	Nominal Categorical	No natural order (Male/Female)
Rank of students	Ordinal Categorical	Has meaningful order (1st, 2nd)
Price of home	Continuous Numerical	Can take any value within range
Color of flower	Nominal Categorical	No natural order (Red, Blue)

- **Categorical**: distinct categories.
- **Numerical**: mathematical operations possible.
- **Ordinal**: categories with sequence.

Mnemonic

"Nominal, Ordinal, Continuous (NOCO)"

Question 2(c) OR [7 marks]

Describe K-fold cross validation in details.

Solution

K-fold cross validation is a model evaluation technique that divides dataset into K equal parts for robust performance assessment.

Table 10. K-fold Process

Step	Action	Purpose
1	Divide data into K equal folds	Create K subsets
2	Use K-1 folds for training	Train model
3	Use 1 fold for testing	Evaluate performance
4	Repeat K times	Each fold acts as test set once
5	Average all results	Final performance metric

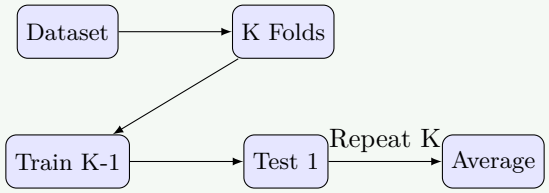

Figure 3. K-Fold Cross Validation

Advantages:
- **Robust Evaluation**: Every data point used for both training and testing.
- **Reduced Overfitting**: Multiple validation rounds.

Mnemonic

"Divide, Use, Repeat, Average, Test (DURAT)"

Question 3(a) [3 marks]

List out applications of regression.

Solution

Table 11. Regression Applications

Domain	Application	Purpose
Finance	Stock price prediction	Forecast market trends
Healthcare	Drug dosage calculation	Determine optimal treatment
Marketing	Sales forecasting	Predict revenue
Real Estate	Property valuation	Estimate house prices

Mnemonic

"Finance, Healthcare, Marketing, Real estate (FHMR)"

Question 3(b) [4 marks]

Write a short note on single linear regression.

Solution

Single linear regression models the relationship between one independent variable (X) and one dependent variable (Y) using a straight line.

Table 12. Linear Regression Components

Component	Formula	Description
Equation	$Y = a + bX$	Linear relationship
Slope (b)	$\Delta Y/\Delta X$	Rate of change
Intercept (a)	Y when X=0	Starting point

- **Goal**: Find best-fit line minimizing errors.
- **Method**: Least squares optimization.

Mnemonic

"Y equals a plus b times X (YABX)"

Question 3(c) [7 marks]

Write and discuss K-NN algorithm.

Solution

K-Nearest Neighbors (K-NN) is a lazy learning algorithm that classifies data points based on the majority class of their K nearest neighbors.

Table 13. K-NN Algorithm Steps

Step	Action	Description
1	Choose K value	Select number of neighbors
2	Calculate distances	Find distance to all training points
3	Sort distances	Arrange in ascending order
4	Select K nearest	Choose K closest points
5	Majority voting	Assign most common class

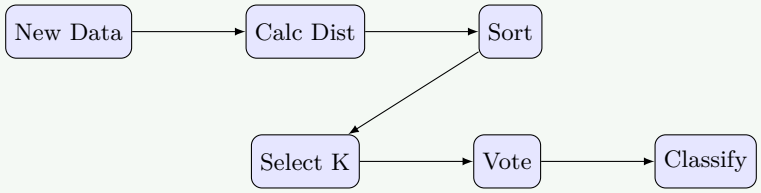

Figure 4. K-NN Process

Advantages:
- Simple to understand and implement.
- No training phase (Lazy).

Disadvantages:
- Computationally expensive on large data.
- Sensitive to K value.

Mnemonic

"Choose, Calculate, Sort, Majority vote (CCSM)"

Question 3(a) OR [3 marks]

Write any three examples of supervised learning in the field of healthcare

Solution

Table 14. Healthcare Supervised Learning

Application	Input	Output
Disease Diagnosis	Symptoms, tests	Disease type
Drug Prediction	Patient genetics	Drug response
Image Analysis	X-rays, MRI	Tumor detection

Mnemonic

"Diagnosis, Drug response, Medical imaging (DDM)"

Question 3(b) OR [4 marks]

Differentiate: Classification v/s Regression.

Solution

Table 15. Classification vs Regression

Aspect	Classification	Regression
Output Type	Discrete categories	Continuous values
Goal	Predict class labels	Predict numerical values
Examples	Spam/Not Spam	House price
Evaluation	Accuracy, Precision	MSE, R-squared

Mnemonic

"CLASS-Categories, REG-Real numbers"

Question 3(c) OR [7 marks]

Explain classification learning steps in details.

Solution

Classification learning involves training a model to assign input data to predefined categories or classes.

Table 16. Classification Steps

Step	Process	Description
1	Data Collection	Gather labeled training examples
2	Preprocessing	Clean and prepare data
3	Feature Selection	Choose relevant attributes
4	Model Selection	Choose algorithm
5	Training	Learn from labeled data
6	Evaluation	Test performance

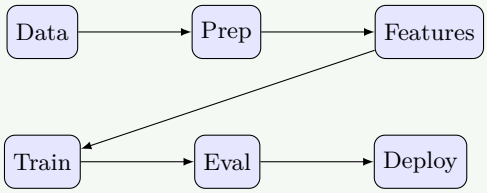

Figure 5. Classification Pipeline

Key Concepts:
- **Supervised Learning**: Requires labeled data.
- **Performance**: Measured by accuracy, precision, recall.

Mnemonic

"Data, Clean, Features, Model, Train, Evaluate, Deploy (DCFMTED)"

Question 4(a) [3 marks]

Differentiate: Clustering v/s Classification.

Solution

Table 17. Clustering vs Classification

Aspect	Clustering	Classification
Learning Type	Unsupervised	Supervised
Training Data	Unlabeled	Labeled
Goal	Find patterns	Predict classes
Output	Groups	Predictions

Mnemonic

"CL-Unknown groups, CLASS-Known categories"

Question 4(b) [4 marks]

List out advantages and disadvantages of apriori algorithm.

Solution

Table 18. Apriori Pros and Cons

Advantages	Disadvantages
Easy to understand	Computationally expensive
Finds all frequent itemsets	Multiple database scans
Generates association rules	Large memory requirements
Simple logic	Poor scalability

Mnemonic

"Easy to use but slow performance (EASY-SLOW)"

Question 4(c) [7 marks]

Write and explain applications of unsupervised learning.

Solution

Unsupervised learning discovers hidden patterns in data without labeled examples.

Table 19. Applications

Domain	Application	Technique
Marketing	Customer segmentation	Clustering
Retail	Market basket analysis	Association rules
Security	Fraud detection	Anomaly detection
Compression	Dimensionality reduction	PCA

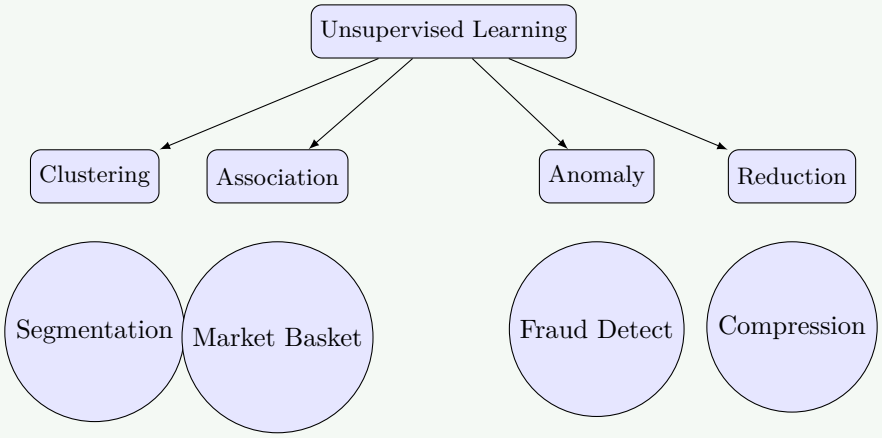

Figure 6. Unsupervised Applications

Mnemonic

"Marketing, Retail, Anomaly, Dimensionality (MRAD)"

Question 4(a) OR [3 marks]

List out applications of apriori algorithm.

Solution

Table 20. Apriori Applications

Domain	Application	Purpose
Retail	Market basket analysis	Find items bought together
Web Mining	Website usage patterns	Discover page visit sequences
Bioinformatics	Gene pattern analysis	Identify gene associations

Mnemonic

"Retail, Web, Bioinformatics (RWB)"

Question 4(b) OR [4 marks]

Define: Support and Confidence.

Solution

Table 21. Association Rule Metrics

Metric	Formula	Description
Support	$Count(A)/Total$	How often itemset appears
Confidence	$Support(A \cup B)/Support(A)$	How often rule is true

Example:

- **Support**: If Bread & Milk appear in 30% of transactions.
- **Confidence**: If 60% of Bread buyers also buy Milk.

Mnemonic

"SUP-How often, CONF-How reliable"

Question 4(c) OR [7 marks]

Write and explain K-means clustering approach in detail.

Solution

K-means clustering partitions data into K clusters by minimizing within-cluster sum of squares.

Table 22. K-means Steps

Step	Action	Description
1	Choose K	Select number of clusters
2	Initialize	Place K centroids randomly
3	Assign	Points to nearest centroid
4	Update	Recalculate centroids
5	Repeat	Until convergence

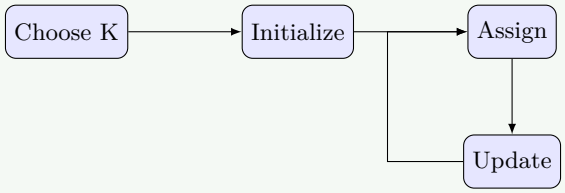

Figure 7. K-means Process

Choosing K: Use Elbow Method to find optimal K value.

Mnemonic

"Choose K, Initialize, Assign, Update, Repeat (CIAUR)"

Question 5(a) [3 marks]

Give the difference between predictive model and descriptive model.

Solution

Table 23. Predictive vs Descriptive Models

Aspect	Predictive	Descriptive
Purpose	Forecast future	Explain present
Output	Predictions	Insights
Examples	Forecasting	Segmentation

Mnemonic

"PRED-Future, DESC-Present"

Question 5(b) [4 marks]

List out application of scikit-learn.

Solution

Table 24. Scikit-learn Applications

Category	Algorithm	Application
Classification	SVM, Random Forest	Spam filtering
Regression	Linear Regression	Price prediction
Clustering	K-means	Customer grouping
Preprocessing	Scalers	Data cleaning

Mnemonic

"Classification, Regression, Clustering, Preprocessing (CRCP)"

Question 5(c) [7 marks]

Explain features and applications of Numpy.

Solution

NumPy is the fundamental library for scientific computing in Python.

Table 25. NumPy Features

Feature	Benefit
N-dim Arrays	Efficient data storage
Broadcasting	Flexible computations
Math Functions	Complete math toolkit
Performance	Fast C implementation

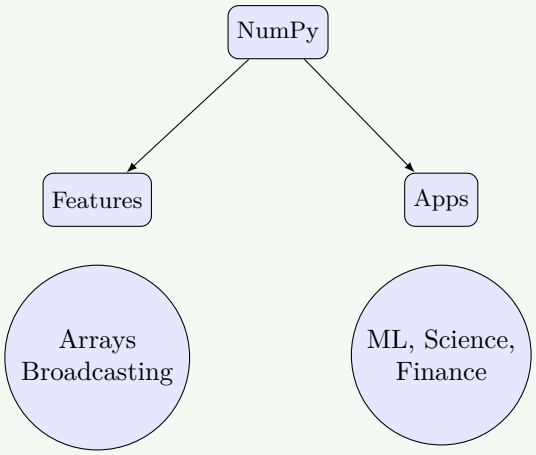

Figure 8. NumPy Functions

Mnemonic

"N-dimensional, Fast, Arrays, Math, Scientific (NFAMS)"

Question 5(a) OR [3 marks]

Write a short note on bagging

Solution

Bagging (Bootstrap Aggregating) improves model performance by training multiple models on different data subsets.

Table 26. Bagging Process

Step	Purpose
Bootstrap	Create diverse training sets
Train	Build independent models
Aggregate	Average predictions to reduce variance

Mnemonic

"Bootstrap, Train, Aggregate (BTA)"

Question 5(b) OR [4 marks]

List out features of Pandas.

Solution

Table 27. Pandas Features

Feature	Description	Benefit
DataFrame	Structured container	Easy manipulation
File I/O	Read/Write	Format support
Cleaning	Missing values	Data prep
Grouping	Aggregation	Analysis

Mnemonic

"DataFrame, File I/O, Indexing, Grouping (DFIG)"

Question 5(c) OR [7 marks]

Explain features and applications of Matplotlib.

Solution

Matplotlib is a comprehensive 2D plotting library for creating publication-quality figures.

Table 28. Matplotlib Capabilities

Feature	Description
Plot Types	Line, bar, scatter, histogram
Customization	Full control over style
Interactive	Zoom, pan support
Output	PNG, PDF, SVG support

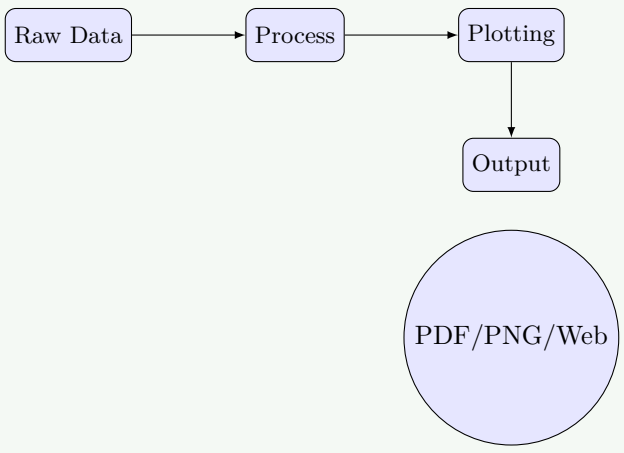

Figure 9. Visualization Pipeline

Mnemonic

"Multiple plots, Visualization, Interactive, Customizable, Scientific (MVICS)"