1. Introduction to Classification

What is classification in Machine Learning?

Classification in machine learning is a supervised learning process where a model learns from labeled data to assign new, unseen data points to predefined categories or classes

How is it different from regression?

Regression = predicts continuous numerical values, such as predicting prices, temperature, income, the output is a real-valued number rather than a category

Classification = predicts categorical labels or classes, such as boy or girl, or donkey versus monkey, It sorts data into predefined groups based on features learned from labeled data

Give one real-life example of classification and one of regression?

Classification example:

A credit card company detecting if transaction is fraudulent or not Output: Yes / No

Regression example:

Predicting a car fuel efficiency based on engine size, weight, and speed Output: 32.7 MPG

2A. Classification Algorithms

Describe the following algorithms:

Logistic Regression

Logistic Regression it predicts the probability that an input belongs to a specific class It is used for binary classification where the output can be one of two possible categories such as Yes/No, True/False or 0/1 It uses sigmoid function to convert inputs into a probability value between 0 and 1

Decision Trees

Decision trees are supervised learning method used for classification and regression, they splitting the dataset into subsets based on feature values, model of decisions,

Decision trees are easy to interpret and visualize and handle both categorical and numerical data

Random Forest

Random forest is builds multiple decision tree training and merges their results to improve classification accuracy and control overfitting, Random forests are powerful and robust, often providing higher accuracy than single decision tree

For each algorithm, explain:

How it works basic idea?

One real-world use case?

Main advantages and limitations?

Logistic Regression

How it works: models the probability of a binary outcome class using a logistic sigmoid function outputs a probability between 0 and 1 which is converted to a class label based on a threshold, commonly 0.5, sigmoid function makes it suitable for classification by mapping any real number to a probability score

Real-world use case: Marketing Predict whether user will click on an online ads Click / No Click

Advantages

1. Simple and easy to implement
2. Works well binary classification problems
3. Provides probability outputs
4. you can understand how each feature affects the outcome)

Limitations

1. Not good for non-linear or complex patterns
2. Sensitive to outliers
3. Doesn't handle high-dimensional or irrelevant features well without preprocessing

Decision Trees:

How it works: It segments the data into homogenous groups for classification

Real-world use case: E-commerce Recommending whether a customer is likely to buy a product or not based on browsing and purchase history

Advantages

1. Easy to visualize and interpret
2. Works for both classification & regression
3. Captures non-linear relationships.
4. No need for feature scaling normalization/standardization

Limitations

1. Can easily overfit
2. Small changes in data can change the tree

Random Forest:

How it works: Random forest builds an ensemble of decision trees using random subsets of data and features, this reduces overfitting and improves accuracy compared to a single decision tree

Real-world use case: Fraud detection in banking, where multiple trees analyze transaction patterns to classify transactions as legitimate or fraudulent

Advantages

1. Reduces overfitting by combining many decision trees
2. Works well with both classification & regression
3. Handles missing data and large datasets effectively
4. More stable and accurate than a single decision tree

Limitations

1. More complex, harder to interpret compared to a single tree
2. Slower to train and predict since it builds many trees
3. Can require tuning for best performance

2B. Extended Task – Research Another Algorithm

What problem is this algorithm good at solving?

Neural networks is good at solving complex problems that involve recognizing patterns and modeling nonlinear relationships in data

1. Image and video recognition: used face recognition, object detection
2. Natural language processing: used language translation, & chatbots
3. Autonomous vehicles: used self-driving
4. Medical imaging: They assist in diagnosing diseases MRIs, X-rays, & CT scans with high accuracy
5. Financial services: detect fraud, predict market trends, credit score
6. Speech recognition: it converts speech to text
7. Robotics and industrial automation: guided robots and automate tasks

How does it work (the core idea)?

A neural network is a system made of layers of neurons, where data enters the input layer and passes through hidden layers in which each neuron applies weights, biases, and an activation function to learn complex patterns, The output layer produces the result, while during training, the network uses a loss function and backpropagation to adjust weights and biases, reducing errors and enabling it to learn intricate relationships and patterns in the data, such as for image or language understanding.

One real-world application where it is used?

used in self-driving cars, Neural networks process and analyze the data from vehicle sensors and cameras to detect and recognize objects such as pedestrians, other vehicles, and road signs, this enables the car to make real-time decisions

What are its strengths and weaknesses compared to the algorithms in Section 2?

compared	Neural Networks	Logistic Regression	Decision Trees	Random Forest
Strengths	1: Can model complex non-linear patterns 2: handles unstructured data high accuracy on large datasets	1.Simple and easy to implement 2.Works well binary 3.Provides probability outputs	1.Easy to visualize & interpret 2.Works for both classification and regression 3.Captures non-linear relationships. 4.No need for feature scaling normalization & standardization	1.Reduces overfitting 2.Works for both classification and regression 3.Handles missing data and large dataset effectively
Weaknesses	1: Needs large data and computation 2: complex design 3: less interpretable	1.Not good for non-linear or complex patterns 2.Sensitive to outliers 3.Doesn't handle high-dimensional	1.Can easily overfit 2.Small changes in data can change the tree	1.More complex, harder to interpret 2.Slower predict since it builds many trees 3.Can require tuning for best performance
Use Case Suitability	Best for complex problems with large data (image recognition, NLP)	Best for simple, linearly separable problems	Good for decision-making with mixed data and interpretability	Ideal for high accuracy on complex data often top performer

3. Classification Metrics

Define and explain what each of these metrics measures:

Accuracy, Precision, Recall, F1-Score, Confusion Matrix

Add a comparison table showing their differences (when to use, what they focus on, weaknesses)

Accuracy: measures the proportion of all correctly predicted instances both positive and negative but can be misleading if the dataset is imbalanced in terms of class

Precision: is the ratio of true positive predictions to the total predicted positives true + false positives It measures the correctness of positive predictions Precision is crucial when false positives are costly

Recall (or sensitivity) is the ratio of true positives to total actual positives true positives + false negatives It measures the ability of the model to find all the positive instances

F1-Score It balances both and is a better measure when the class distribution is imbalanced, or when you want to balance between false positives and false negatives

Confusion Matrix is a table used to describe the performance of a classification model, showing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)

METRIC	WHAT IT MEASURES	WHEN TO USE	FOCUS	WEAKNESSES
ACCURACY	Overall correctness of model predictions	Balanced class distribution	Correct predictions over total	Misleading with imbalanced datasets
PRECISION	Correctness of positive predictions	When false positives are costly	Avoiding false positives	Can ignore false negatives
RECALL	Ability to identify all positive instances	When missing positives is costly	Avoiding false negatives	Can ignore false positives
F1-SCORE	Balance between precision and recall	Imbalanced classes or when both errors matter	Trade-off between precision and recall	Not intuitive, single number abstraction
CONFUSION MATRIX	Detailed breakdown of predictions	Always useful for all classification tasks	Visual and detailed error analysis	Needs interpretation, no single score

4: Imbalanced Data Problem

Explain the meaning of unbalanced data in classification tasks.

Why does the accuracy of an unbalanced dataset suffer?

Which measures of reliability are appropriate in such cases, and why?

When classes are not evenly distributed

One class the majority is significantly more numerous than the other the minority

Example: dataset of 1000 samples 950 belong to class A, and only 50 belong to class B

1. Accuracy = Correct Predictions / Total Predictions
    a. example: Predicting all as majority class can still give high accuracy
2. Precision = Among predicted positives, how many are correct?
    a. Use when false positives are costly
    b. Example: Diagnosing cancer — wrong positive causes stress & tests
3. Recall (Sensitivity) = Among actual positives, how many are detected?
    a. Use when missing positives is costly
    b. Example: Fraud detection — missing fraud is very bad
4. F1-Score = Harmonic mean of Precision & Recall.
    a. Good when need balance between precision & recall

5: Real-World Case Study

Research a real-world project or study that used classification (e.g., spam detection, fraud detection, medical diagnosis, customer churn prediction).

Summarize:

1. The goal of the project
2. The data they used
3. The classification model applied
4. The key results or insights

1 Spam Detection_Email

1. Goal: Detect spam emails automatically
2. Data: dataset thousands of real company emails
3. Model: SVM, Decision Trees
4. Key Result: showed that even simple models can handle spam filtering effectively if data is preprocessed properly

2 Fraud Detection_Credit Card Transactions

1. Goal: Identify fraudulent credit card transactions in real-time
2. Data: Large & highly imbalanced dataset most genuine, very few fraud
3. Model: Random Forest
4. Key Result: Accuracy was misleading; Precision, Recall better evaluation

3) Medical Diagnosis_Cudurka Sonkorta

1. Goal: Detect diabetic eye disease early to prevent blindness
2. Data: Over 100,000 retina images labeled by expert doctors
3. Model: Neural Networks
4. Key Result: The artificial intelligence Achieved performance close to human doctors

References:

1: https://www.ibm.com/think/topics/classification-machine-learning

2: https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/

3: https://www.ibm.com/think/topics/logistic-regression

4: https://jayantb1019.github.io/Analysis/telecom-churn-case-study-1.html