SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING-IoT Including CS & BCT

COURSE NAME :19SB701 PATTERN RECOGNITION TECHNIQUES IN CYBER CRIME

IV YEAR / VII SEMESTER

Unit III- **NONPARAMETRIC TECHNIQUE AND NON-METRIC METHODS**

Topic :Non Metric Methods- Introduction- Decision Trees- CART

Non-metric Methods

- We have focused on real-valued feature vectors or discrete valued numbers with a natural measure of distance between vectors (metric)

- Some classification problems describe a pattern by a list of attributes: a fruit may be described by 4-tuple (red, shiny, sweet, small)

- How to learn categories using non-metric data where distance between attributes can not be measured?

- Decision tree, a.k.a. hierarchical classifier, multistage classification, rule-based methods

Data Type and Scale

• **Data type:** degree of quantization in the data – binary feature: two values (Yes-No response) – discrete feature: small number of values (image gray values) – continuous feature: real value in a fixed range • Data scale: relative significance of numbers – qualitative scales

• **Nominal (categorical):** numerical values are simply used as names; e.g., (yes, no) response can be coded as (0,1) or (1,0) or (50,100)

• **Ordinal:** numbers have meaning only in relation to one another (e.g., one value is larger than the other); e.g., scales (1, 2, 3), and (10, 20, 30) are equivalent – quantitative scales

• **Interval:** separation between values has meaning; equal differences on this scale represent equal differences in temperature, but temperature of 30 degrees is not twice as warm as one of 15 degrees.

• **Ratio:** an absolute zero exists along with a unit of measurement; ratio between two numbers has meaning (height)

General Class of Metrics

- Minkowski metric

$$L_k(\mathbf{a},\mathbf{b}) = \left(\sum_{i=1}^{d} | a_i - b_i |^k \right)^{1/k}$$

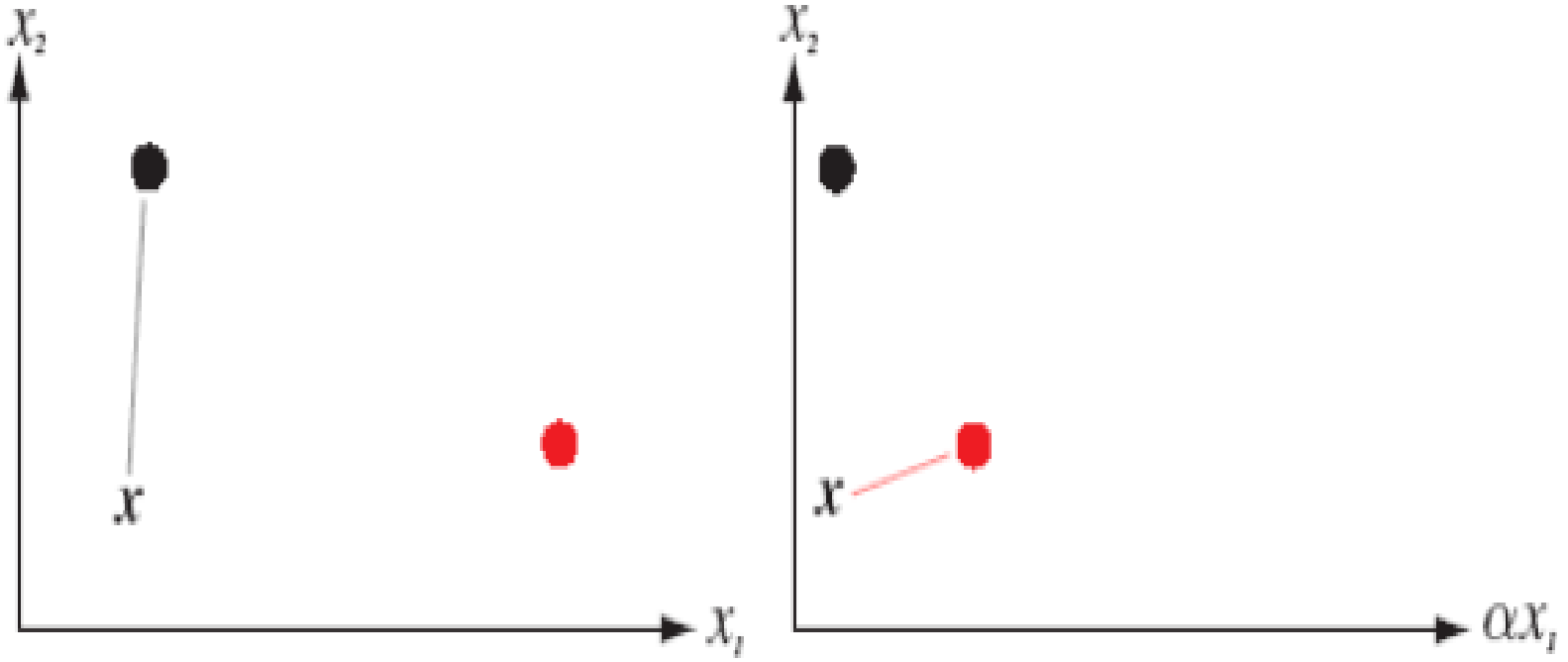
- Manhattan distance

$$L_1(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{d} | a_i - b_i |$$

Scaling the Data

- Although one can always compute the Euclidean distance between two vectors, the results may or may not be meaningful

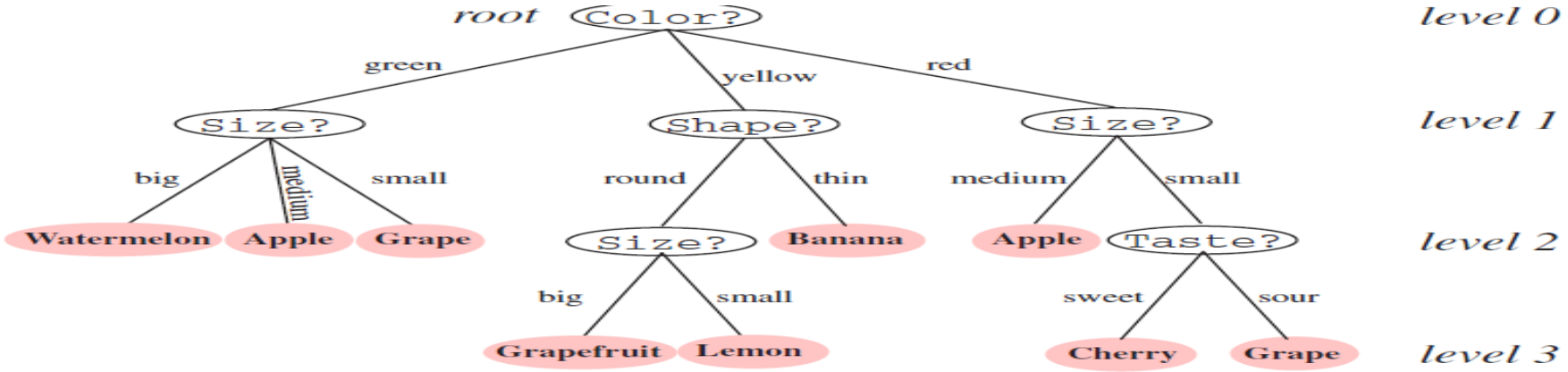
- If the space is transformed by multiplying each coordinate by an arbitrary constant, the Euclidean distance in the transformed space is different from original distance relationship; such scale changes can have a major impact on NN classifiers


X_2
x
X_1
X_2
x
αX_1

Decision Tree

The Decision Tree algorithm is a hierarchical tree-based algorithm that is used to classify or predict outcomes based on a set of rules. It works by splitting the data into subsets based on the values of the input features.

Seven-class, 4-feature classification problem Apple = (green AND medium) OR (red AND medium) = (Medium AND NOT yellow)



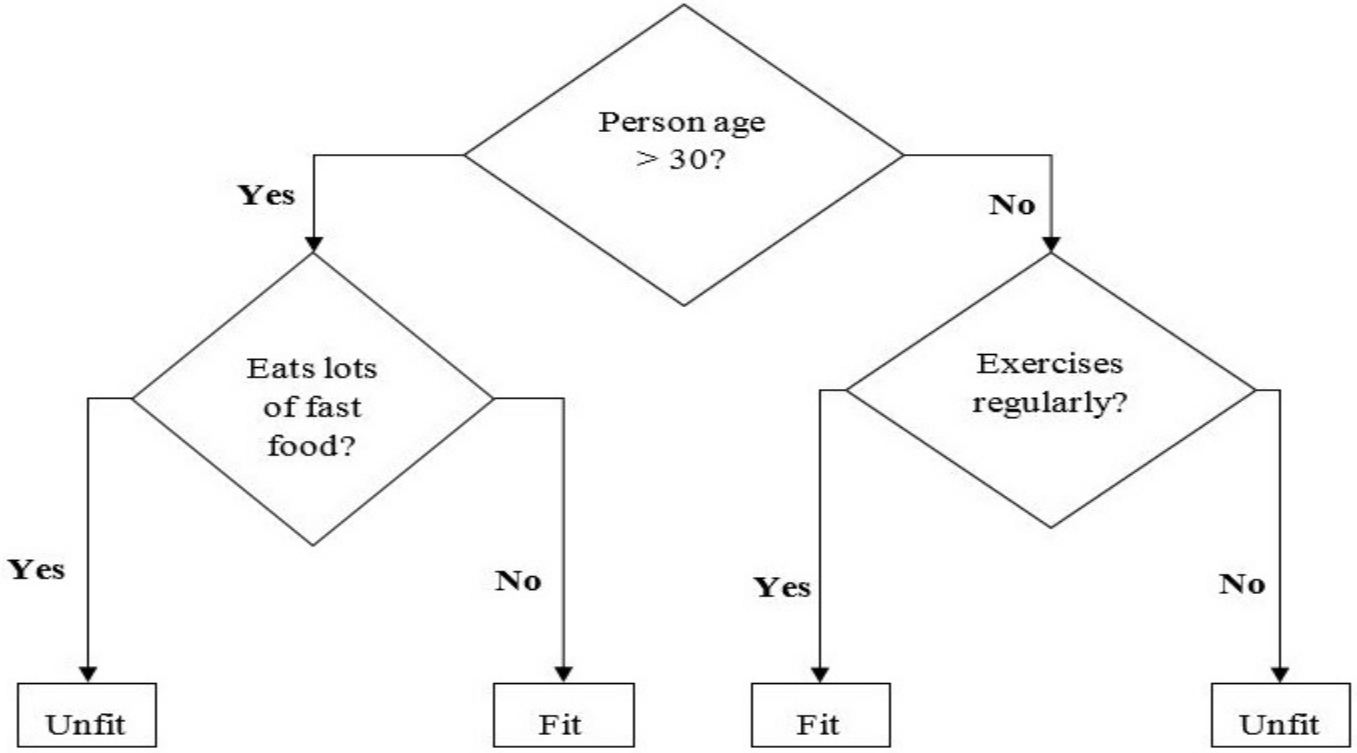
Types of Decision Tree Algorithm

There are two main types of Decision Tree algorithm –

Classification Tree – A classification tree is used to classify data into different classes or categories. It works by splitting the data into subsets based on the values of the input features and assigning each subset to a different class.

Regression Tree – A regression tree is used to predict numerical values or continuous variables. It works by splitting the data into subsets based on the values of the input features and assigning each subset a numerical value.

The example of a binary tree for predicting whether a person is fit or unfit providing various information like age, eating habits and exercise habits, is given below –



CART

CART(Classification And Regression Tree) for Decision Tree

CART is a predictive algorithm used in Machine learning and it explains how the target variable's values can be predicted based on other matters.

It is a decision tree where each fork is split into a predictor variable and each node has a prediction for the target variable at the end.

The term CART serves as a generic term for the following categories of decision trees:

Classification Trees: The tree is used to determine which "class" the target variable is most likely to fall into when it is continuous.

Regression trees: These are used to predict a continuous variable's value.

CART Algorithm

Classification and Regression Trees (CART) is a decision tree algorithm that is used for both classification and regression tasks.

It is a supervised learning algorithm that learns from labelled data to predict unseen data.

Tree structure: CART builds a tree-like structure consisting of nodes and branches.

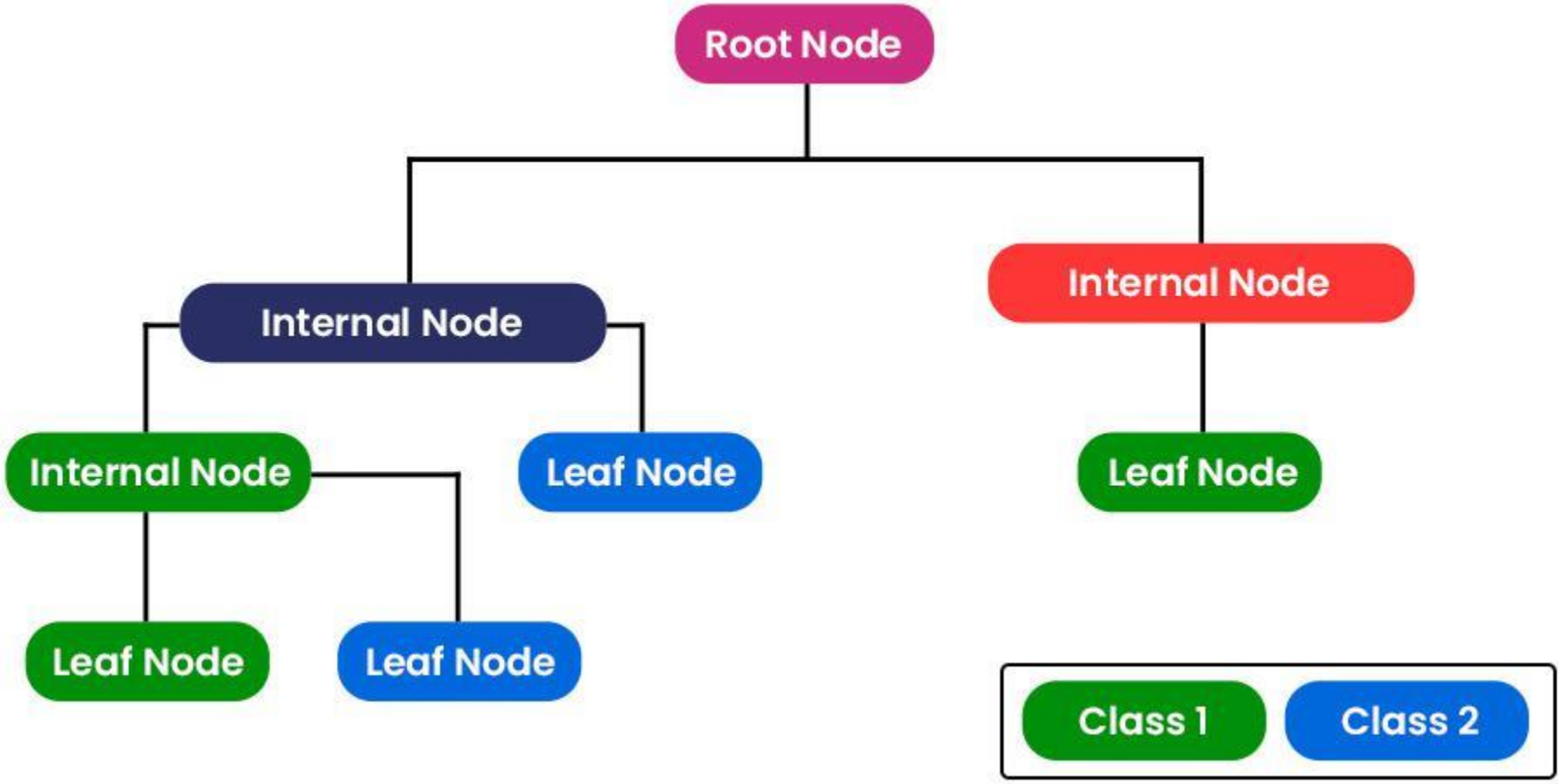
Splitting criteria: CART uses a greedy approach to split the data at each node.

Pruning: To prevent overfitting of the data, pruning is a technique used to remove the nodes that contribute little to the model accuracy.

How does CART algorithm works?

The CART algorithm works via the following process:

The best-split point of each input is obtained.
Based on the best-split points of each input in Step 1, the new “best” split point is identified.
Split the chosen input according to the “best” split point.
Continue splitting until a stopping rule is satisfied or no further desirable splitting is available.


Root Node
Internal Node
Internal Node
Internal Node
Leaf Node
Leaf Node
Leaf Node
Leaf Node
Class 1
Class 2

CART for Classification

A classification tree is an algorithm where the target variable is categorical.

The algorithm is then used to identify the “Class” within which the target variable is most likely to fall.

Classification trees are used when the dataset needs to be split into classes that belong to the response variable(like yes or no)

CART for Regression

A Regression tree is an algorithm where the target variable is continuous and the tree is used to predict its value.

Regression trees are used when the response variable is continuous.

For example, if the response variable is the temperature of the day.

Any Query????

Thank you……