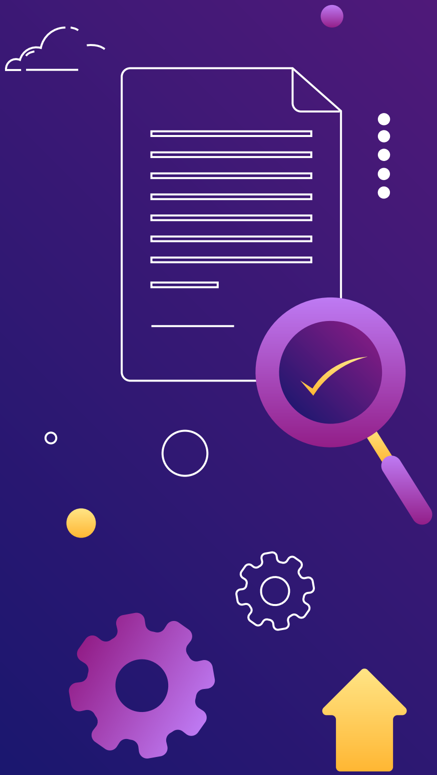
MDL

MINIMAL DESCRIPTION LENGTH PRINCIPLE

Attribute selection method in Decision Tree

By Group 4



OUR TEAM

Lim KimHoung
Sok Sopheak
Dorn Dawin
Nang Sreynich
Long Channleap
Koh Tito

But Cheableng
Phoung Bunthoen
Phai Ratha
Pheng Sothea
Kheang tongheang

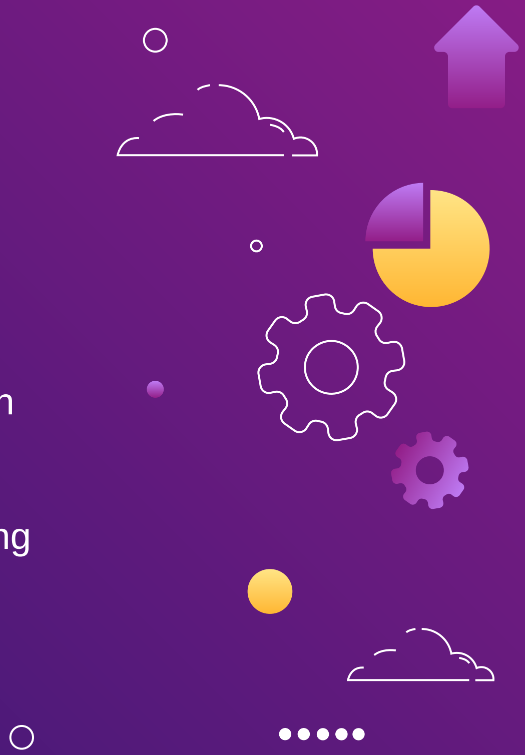


TABLE OF CONTENT

01 ABOUT DECISION TREE

Introduction to decision tree

02 INTRODUCTION TO MDL IN BUILDING DECISION TREE

What is MDL? Talking about concept MDL in Decision Tree

03 COMPRESSION AND ENTROPY

How does it work?

01

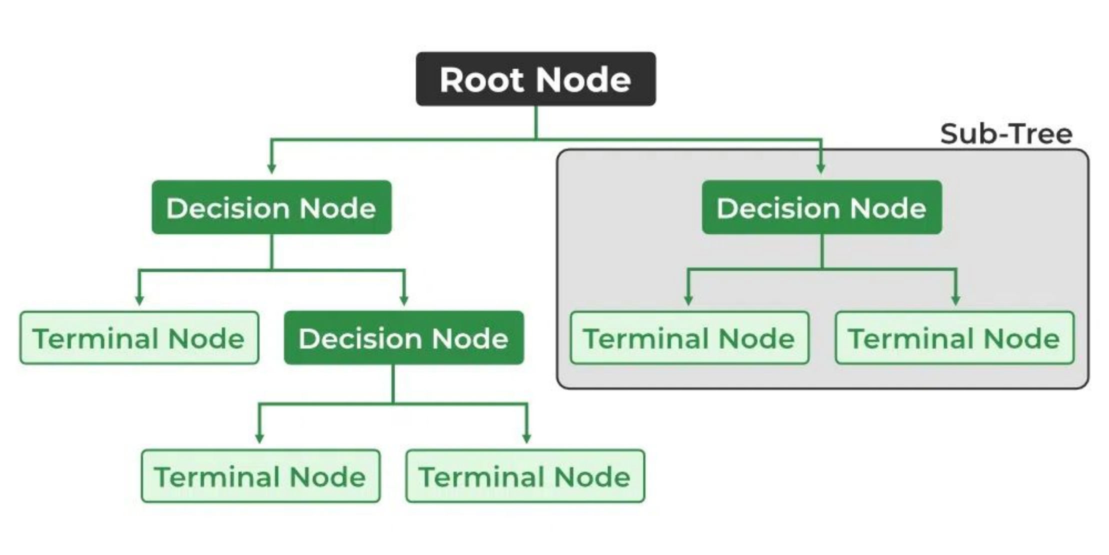
OVERVIEW ON DECISION TREE

Here we introduce you a little bit about decision tree

Decision Tree is a flowchart-like tree structure.

It consist of **root node, decision node, terminal node.**

It can be use in both regression and classification problem.



02

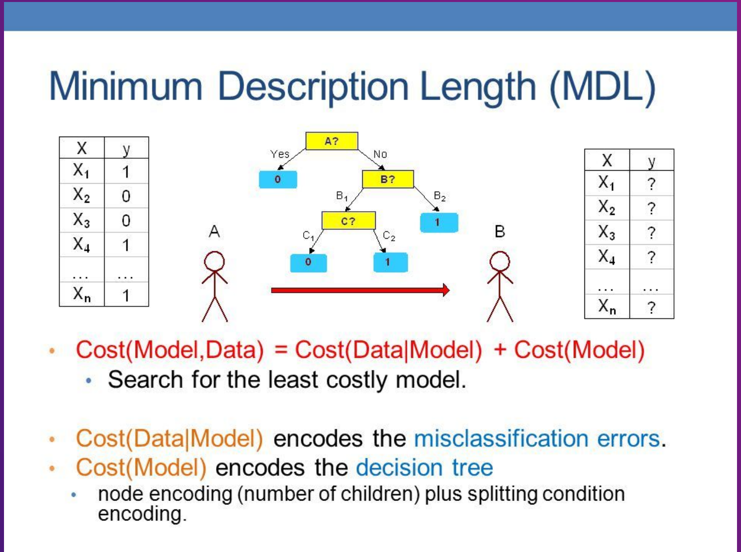
ABOUT

MDL IN DECISION TREE

Here we introduce you MDL principle

WHAT IS MDL PRINCIPLE?

Minimum Description Length(MDL) principle that can be used in Decision Tree algorithms to guide the process of selecting the best features and creating an optimal tree structure.



Minimum Description Length (MDL)

X	y
X_1	1
X_2	0
X_3	0
X_4	1
…	…
X_n	1

X	y
X_1	?
X_2	?
X_3	?
X_4	?
…	…
X_n	?

- Cost(Model,Data) = Cost(Data|Model) + Cost(Model)
  - Search for the least costly model.

- Cost(Data|Model) encodes the misclassification errors.
- Cost(Model) encodes the decision tree
  - node encoding (number of children) plus splitting condition encoding.

HERE IS SOME GENERAL CONCEPT OF MDL IN ML

- **Model Selection:** The MDL principle is often used to select the best model from a set of several models. By choosing the model that provides the best representation of the data, the MDL principle helps prevent overfitting and ensures that the selected model.
- **Clustering:** The MDL principle can be used for clustering, which involves grouping similar objects together. By applying the MDL principle to clustering, we can find to identify the most important features that distinguish one cluster from another.
- **Time series analysis:** The MDL principle can be used for time series analysis, which involves analyzing data that generates the data and make accurate predictions about future values. For example, suppose you have a dataset of stock prices over time.

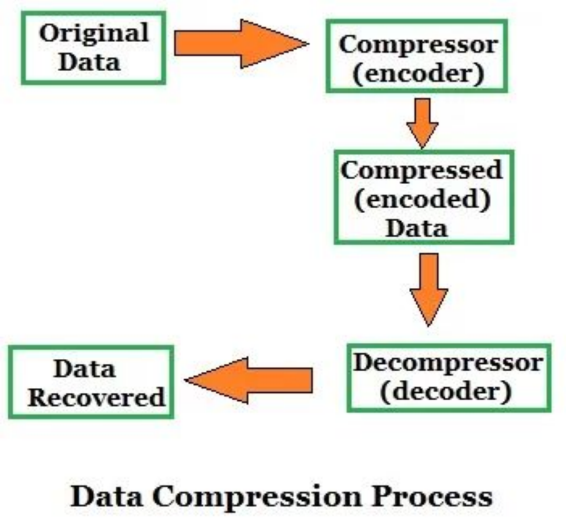
03

COMPRESSION AND ENTROPY

The way on how to choose the variables

DATA COMPRESSION

Data compression is the process of encoding information using fewer bits than what the original representation uses.



ENTROPY

Entropy is a measure of uncertainty. It quantifies the uncertainty in a random variable as the information required to specify its value.

For example, if the variable represents the sex of an individual, then the number of possible values is two:
female and male. If the variable represents the salary of individuals expressed in whole dollar amounts, then the values can be in the range $0-$10B, or billions of unique values. Clearly it takes more information to specify an exact salary than to specify an individual’s sex.

VALUES OF A RANDOM VARIABLE

Statistical Distribution

Information (the number of bits) depends on the statistical distribution of the values of the variable as well as the number of values of the variable. If we are judicious in the choice of Yes/No questions, then the amount of information for salary specification cannot be as much as it first appears

Significant Predictors

Suppose that for some random variable there is a predictor that when its values are known reduces the uncertainty of the random variable.

THANK

YOU