“Don't let the digital supply chain scare you.”

Assignment 4

Induce decision trees using three attribute selection measures. For each tree, plot test and training error. Plot at least 20 points for training error and 40 for test error. Clearly mention the protocol used for plotting the error curves.

27.03.2020

Nitesh Yadav(35), Rohit Rana(40), Rohit Shakya(41)
M.Sc. Computer Science
Department of Computer Science,
University of Delhi

Overview

Predict the outcome of the game with the given positions of 2 kings and a rook.

About Data

- **Instances:** 28056 **Attributes:** 7 and **No Missing Values**

Attribute Details :

Name	Type	Description
white_king_file	string	Column location on the chess board of the white king
white_king_rank	string	Row location on the chess board of the white king
white_rook_file	string	Column location on the chess board of the white rook
white_rook_rank	string	Row location on the chess board of the white rook
black_king_file	string	Column location on the chess board of the black king
black_king_rank	string	Row location on the chess board of the black king
result	string	Predictor Class. optimal depth-of-win for White in 0 to 16 moves, otherwise drawn Values: {draw, zero, one, two, ..., sixteen}

Sample Dataset

Used factorize function for handling Categorical Attributes.

HeatMap - Shows correlation between attributes.

This clearly shows that results are dependent on white king rank more than any other feature.

Methods

Gini Index: It is calculated by subtracting the sum of squared probabilities of each class from one. It favors larger partitions and is easy to implement whereas information gain favors smaller partitions with distinct values.

Entropy: It is a measure of disorder or uncertainty and the goal of machine learning models and Data Scientists in general is to reduce uncertainty.

Train and Test error for different size of test tests and at different depth of Decision trees

Incremented test size every time by 2%.

Gini Index

Entropy

Results

Gini

- **Mean Train_error 0.098414**
- **Mean Test_error 0.340853**

Entropy

- **Mean Train_error 0.099461**
- **Mean Test_error 0.338760**

From here we can't conclude that which one is better both are almost equally effective.

Nuts and Bolts

- **At max_depth = 14 and test size of 20% gives the best results in both the scenarios (Gini Index and Entropy).**
- **As the max_depth of the tree increases the accuracy of our model gets better.**
- **When test size is less than (50%-60%) then the relationship between training and testing error becomes linear.**
- **When the training error increases exponentially, The model is Under-Fitting.**
- **When test error starts increasing exponentially. The model is overfitting.**