**WIE3007 2023/2024 S1**

CHEONG HUI TING U2005292

Tutorial 8

1. Describe the historical evolution of decision trees. Why did they gain popularity in the realm of data analysis and machine learning? Mention at least three major milestones in their development.

* The concept of decision trees began in 1960s.
* 1970s, decision tree algorithms utilizing entropy as a metric
* 1980s, gained better handling of missing data and continuous attributes.
* 1990s, decision tree algorithms become a fundamental tool in data mining, and CART algorithm, which extended decision trees to handle both classification and regression tasks
* 2000s. ensemble methods combined multiple decision trees to be more accurate. Random Forests, introduced by Leo Breiman, and Gradient Boosting Machines, pioneered by Jerome Friedman, provided substantial improvements in predictive performance.

It gets popular because of their interpretability and predictive power is high of decision tree, making them preferable in industries.

1. Imagine you've been given a dataset of customer transactions for an e-commerce website. Detail the steps you would take to:

* Import this dataset into SAS Enterprise Miner.

1. Launch SAS enterprise miner, right click on “Data Source” and select “Import Data.”
2. Choose the appropriate data source type (e.g., CSV, Excel) and browse to the location of your dataset.
3. Specify the settings for importing your dataset, such as data format, field delimiters, etc.
4. Confirm the import and review the imported data in the Data Sources panel.

* Identify and handle any missing values.

-Using “replace missing value” node to impute or handle missing value. Common methods include mean imputation, median imputation, or using predictive modelling techniques.   
-Other than that, “Filter Data” node can be use to exclude observation with missing value.

* Specify the role of each variable, especially identifying the target variable (e.g., customer churn

-Use the "Variable Roles" node to specify the role of each variable in your analysis. Roles include Input, Output, Target, and Input Rejected.

-Identify the variable that represents the target variable. This is typically the variable you want to predict or model, such as customer churn.

1. Using the SAS Enterprise Miner, you've added a Decision Tree node to your workspace.

Describe the primary functionalities this node offers.

If you wanted to change the splitting criteria of the tree to "Gini impurity", where within the node settings would you do this?

Primary functionalities:

-build decision trees based on the input data

-provide visual representation of the decision tree

-assess model’s performance on training and validation datasets

Gini Impurity setting:

1. open setting and tab named “Tree Specifications”

2. find the splitting criteria option and dropdown menu to select ‘Gini Impurity’

option

3. apply setting and run the decision tree node

1. You've built a decision tree using the default configurations in SAS Enterprise Miner, but you notice signs of overfitting.

What are some symptoms of overfitting in a decision tree?

Suggest three advanced configurations or tuning techniques you would apply to address this overfitting.

Symptoms:

* High accuracy on training data but low accuracy on validation of data
* The tree structure is too complex with many branched and leaves
* Poor generalization to unseen data

Techniques:

* Pruning involves removing branches (subtrees) from the tree that do not significantly contribute to improving predictive accuracy on a validation dataset.
* Limit the maximum depth of the tree to prevent it from growing too deep
* Cross-validation helps detect overfitting by evaluating the model on multiple training/validation splits

1. Explain the primary difference between Bagging and Boosting as ensemble methods for decision trees.

Using the Random Forest algorithm as an example, describe how Bagging is implemented.

Highlight a scenario where Boosting might be more advantageous than Bagging and explain why.

|  |  |  |
| --- | --- | --- |
| Differences | Bagging | Boosting |
| Sampling | Use bootstrap sampling | Re-weights or re-samples the dataset |
| Model building | Built independently | Built sequentially |
| Final prediction | Average or majority vote | Assign weight to model prediction based on accuracy |
| Error reduction | Reducing variance | Reducing bias and variance |
| Random Forest | Based on decision trees, it introduces additional randomness by selecting a subset of features for each split | - |

When to use

Bagging:

When the model has high variance, deep decision tree

When the dataset is large

Boosting:

When model has high bias, shallow decision tree

When high predictive accuracy is required

1. Describe the improvements made to the decision tree in 2004 as mentioned in the presentation. What were the key enhancements, and how did they impact the performance of the decision tree?
2. Slide 5 (The presentation) mentions the use of SQL Logs (Select statements) as the input for the improved decision tree. How can a decision tree be applied to analyze SQL Logs? What kind of outputs can be expected?

Apply decision tree to analyze SQL Logs:

* Feature engineering to extract relevant features to be used as decision tree input variables
* Organize the SQL log data into a structured dataset with each row representing a unique SQL query and the selected features as columns

Expected Outputs:

* Organize the SQL log data into a structured dataset with each row representing a unique SQL query and the selected features as columns
* Identify anomalies or unusual patterns in SQL queries

1. Discuss the role of decision trees in data mining systems (The presentation) . How do decision trees contribute to the effectiveness of these systems?
2. Interpretability: This interpretability makes decision trees valuable for extracting actionable insights and communicating findings
3. In classification, the tree predicts the class or category of a given instance, while in prediction, it estimates a numerical value for a continuous target variable, making it suitable for data mining applications like customer segmentation, fraud detection and predictive modelling.
4. Handle missing value: The algorithm can make decisions based on available data, and missing values do not necessarily hinder the building of the tree

Note: The presentation is derived from the following video: https://www.youtube.com/watch?v=OGIiGg8qKFg&ab\_channel=UniversitiMalaya