Lesson 13

Decision Tree

Tree-based Learning Algorithms

- Simpler and efficient supervised learning methods
- Provide accurate, persistent and ease of analysis to predictive models
- Solving classification and regression problems

Tree-based Learning Algorithms

- Suitable for representing non-linear relationships as well
- Examples of tree-based learning algorithms: Decision trees, Random Forest and Gradient Boosting

Decision Tree Terms Used

1. Root Node: Represents the entire dataset
2. Splitting: A process of dividing a node into two or more sub-nodes
3. Decision Node: When a sub-node splits into further sub-nodes
4. Leaf/Terminal Node: Nodes that do not split further

Decision Tree Terms Used

5. Pruning: Process of removing sub-nodes of a decision node (opposite of splitting)
6. Branch/Sub-Tree: A sub-section of the entire tree
7. Parent Node: A node divided into sub-nodes;
8. Child Node: A node derived from a parent node.

Decision Tree

- A supervised learning algorithm
- Gives a desired response value
- Mostly used in classification problems
- Works for both categorical and continuous input and output variables

Three Variable Example

1. A dataset of 100 students with three variables
2. Gender (Boy/Girl),
3. Branch (CS/EC) and
4. GPA scores of previous year: ≥ 9.0 and ≤ 9.0

Categorical-variable Decision Tree

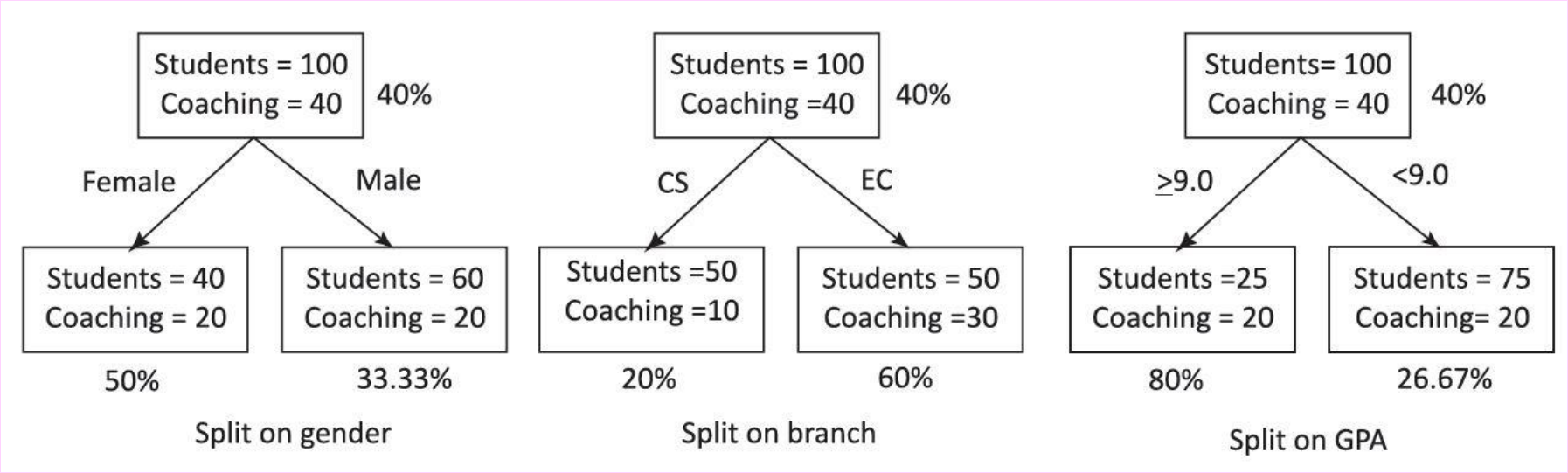
- Decision tree, which has categorical response variable
- . For example, Response variable is "Will more students enroll in coaching class or not" and the answer is NO (Figure 6.17) "Is GPA split is more significant compare to Gender?" YES

Continuous-variable Decision Tree

2: Decision tree, which has continuous response variable

For example, 50% female students enroll in coaching classes, 33% of male students in coaching classes out of admitted students in the coaching classes (Figure 6.17)

An Example of Decision Tree; Students in Coaching Class and Non-Coaching Classes



Steps in Three Variable Case

- Segregate the datasets based on all values of three variables
- Identify the variable, which creates the best homogeneous sets of datasets (which are heterogeneous to each other)

Steps

- First, split the dataset when classifying a response variable
- That is based on most significant splitter/ differentiator in input variables into two or more subsets

Applications

1. Identify the best combination of products and marketing strategies that target specific sets of consumers in a marketing area.
2. Customer behavior analysis, customer retention strategy planning
3. Fraud detection in industries
4. Medicines for diagnosis of diseases

Decision Tree Usages' Advantages

1. Decision tree output in the form of graphical representation is very easy to understand.
2. Useful in predicting significant response variable.
3. Not influenced by outliers and missing values to a fair degree compared to other classifiers

Decision Tree Usages' Advantages

4. Handles both numerical and categorical variables.
5. A non-parametric method, thus, decision trees have no assumptions about space distribution and classifier structure

Disadvantages

1. Over fitting, [Setting constraints on model parameters and pruning solve this problem]
2. When the numerical variables are continuous, the decision tree loses information while categorizing the variables in different categories

Decision Tree Metrics to find best split variables

- Gini index
- Chi-square
- Entropy and Information Gain.

GINI Index ($p^2 + q^2$)

- Sum of square of probability for success (p) and failure (q)
- Higher Gini index means greater significance in decision (Example 6.20)

Entropy and Information Gain

- Entropy in thermodynamics is a measure of disorder or randomness of a system (randomness in dataset)
- A pure node requires less information to describe it, and an impure node requires more information.
- Information Gain = 1 − Entropy

Entropy

- = 0 when the subset or the dataset is completely homogeneous
- = 1 if the sample is an equally divided (50% – 50%)

Entropy Computations

- $H = \sum_{i=1}^{n} p_i \log_2 p_i$
- p = probability of success at a node

Steps to calculate entropy for a split

1. Calculate entropy of parent node.
2. Calculate entropy of each individual node of split and calculate the weighted average of all sub-nodes available in split. (Example 6.22)

A Decision Tree Algorithm CART

- Classification and Regression Tree
- Uses Gini index to split the node
- Results in a binary decision tree (each node will have only two child nodes)

Summary

We learnt:

- Decision Tree
- Responses in Categorical and Continuous variables
- Splits on Multiple Variables
- Decision Tree metrics: GINI index, Chi-square, Entropy and Information Gain
-

End of Lesson 13 on

Decision Tree