Machine Learning

Classification and Regression Trees (CART)

Satishkumar L. Varma

Department of Information Technology
SVKM's Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai.
ORCID | Scopus | Google Scholar | Google Site | Website

Outline

- Learning with Regression and Trees
  - Learning with Regression
    - Simple Linear Regression
    - Multiple Linear Regression
    - Logistic Regression
  - Learning with Trees
    - Decision Trees
    - Constructing Decision Trees using Gini Index
    - Classification and Regression Trees (CART)

Classification and Regression Trees (CART)

- An algorithm can be transparent only if its decisions can be read and understood by people clearly.
- Even though DL is better than ML, it is an opaque algorithm and we do not know the reason of decision.
- Decision tree algorithms still keep their popularity because they can produce transparent decisions.
- ID3 uses information gain
- C4.5 uses gain ratio for splitting.
- CART is an alternative decision tree building algorithm.
    - It can handle both classification and regression tasks.
    - It uses a new metric named gini index to create decision points for classification tasks.
    - CART are a type of decision tree algorithm used in ML for predictive modeling.
    - It is used for both classification (predicting categorical outcomes) and regression (predicting continuous outcomes) tasks.
    - It is a decision tree algorithm that splits a dataset into subsets based on the most significant variable.
    - The goal is to create the purest subsets possible,
        - where "pure" means that the subset contains only instances of a single class (for classification) or has minimal variance (for regression).

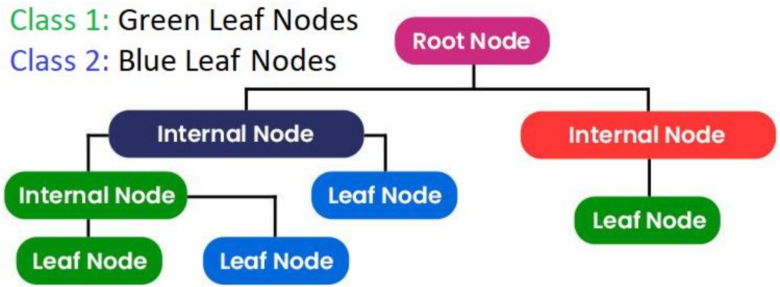
Classification and Regression Trees (CART)

- Types of CART
- Classification Trees:
  - Used when the target variable is categorical.
  - For example, predicting whether an email is spam or not.
- Regression Trees:
  - These are used to predict a continuous variable's value.
  - Used when the target variable is continuous.
  - For example, predicting house prices based on features like size and location.
- In the decision tree, nodes are split into sub-nodes based on a threshold value of an attribute.
- The root node is taken as the training set and is split into two by considering the best attribute and threshold value.
- Further, the subsets are also split using the same logic.
- This continues till the last pure sub-set is found in the tree or the maximum number of leaves possible in that growing tree.

Classification and Regression Trees (CART)

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

- CART algorithm uses Gini Impurity (Gini index) to split the dataset into a decision tree.
- It does that by searching for the best homogeneity for the sub nodes, with the help of the Gini index criterion.



Constructing Decision Trees using Gini Index

$$Gini = 1 - \sum_{i=1}^{n}(p_i)^2$$

- Gini Index (Gini Impurity)
- The Gini index is a metric for the classification tasks in CART.
- It stores the sum of squared probabilities of each class.
- It computes the degree of probability of a specific variable.
- It works on categorical variables, provides outcomes either "successful" or "failure" and
  - hence conducts binary splitting only.
- The degree of the Gini index varies from 0 to 1,
- Where 0 depicts that all the elements are allied to a certain class, or only one class exists there.
- Gini index close to 1 means a high level of impurity,
  - where each class contains a very small fraction of elements, and
  - A value of 1–1/n occurs when the elements are uniformly distributed into n classes and
  - each class has an equal probability of 1/n.
  - For example, with two classes, the Gini impurity is 1 – 1/2 = 0.5.
- where p_i is the probability of an object being classified to a i^{th} class.
- Gini impurity is the probability of misclassification,
  - assuming independent selection of the element and its class based on the class probabilities.

Classification and Regression Trees (CART)

- 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)
- For classification in decision tree learning algorithm that creates a tree-like structure to predict class labels.
- The tree consists of nodes, which represent different decision points, and branches, which represent the possible result of those decisions.
- Predicted class labels are present at each leaf node of the tree.

Classification and Regression Trees (CART)

- CART for Classification
- How Does CART for Classification Work?
- CART for classification works by recursively splitting the training data into smaller and smaller subsets based on certain criteria.
- The goal is to split the data in a way that minimizes the impurity within each subset. Impurity is a measure of how mixed up the data is in a particular subset.
- For classification tasks, CART uses Gini impurity
- Gini Impurity
    - Gini impurity measures the probability of misclassifying a random instance from a subset labeled according to the majority class.
    - Lower Gini impurity means more purity of the subset.
- Splitting Criteria
    - The CART algorithm evaluates all potential splits at every node and chooses the one that best decreases the Gini impurity of the resultant subsets.
    - This process continues until a stopping criterion is reached, like a maximum tree depth or a minimum number of instances in a leaf node.

Classification and Regression Trees (CART)

- 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.
- CART for regression is a decision tree learning method that creates a tree-like structure to predict continuous target variables.
- The tree consists of nodes that represent different decision points and branches that represent the possible outcomes of those decisions.
- Predicted values for the target variable are stored in each leaf node of the tree.

Classification and Regression Trees (CART)

- CART for Regression
- How Does CART works for Regression?
- Regression CART works by splitting the training data recursively into smaller subsets based on specific criteria.
- The objective is to split the data in a way that minimizes the residual reduction in each subset.
- Residual Reduction
  - Residual reduction is a measure of how much the average squared difference between the predicted values and the actual values for the target variable is reduced by splitting the subset.
  - The lower the residual reduction, the better the model fits the data.
- Splitting Criteria
  - CART evaluates every possible split at each node and selects the one that results in the greatest reduction of residual error in the resulting subsets.
  - This process is repeated until a stopping criterion is met, such as reaching the maximum tree depth or having too few instances in a leaf node.

Classification and Regression Trees (CART)

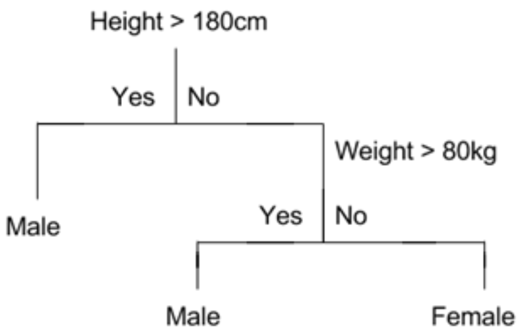
- CART-BASED ALGORITHMS:
- CART (Classification and Regression Trees)
  - The original algorithm that uses binary splits to build decision trees.
- C4.5 and C5.0:
  - Extensions of CART that allow for multiway splits and handle categorical variables more effectively.
- Random Forests:
  - Ensemble methods that use multiple decision trees (often CART) to improve predictive performance and reduce overfitting.
- Gradient Boosting Machines (GBM):
  - Boosting algorithms that also use decision trees (often CART) as base learners, sequentially improving model performance.

Classification and Regression Trees (CART)

- **Advantages of CART**
  - Results are simplistic.
  - Classification and regression trees are Nonparametric and Nonlinear.
  - Classification and regression trees implicitly perform feature selection.
  - Outliers have no meaningful effect on CART.
  - It requires minimal supervision and produces easy-to-understand models.
- **Limitations of CART**
  - Overfitting.
  - High Variance.
  - low bias.
  - the tree structure may be unstable.
- **Applications of the CART algorithm**
  - For quick Data insights.
  - In Blood Donors Classification.
  - For environmental and ecological data.
  - In the financial sectors.

Classification and Regression Trees (CART)

- CART Model Representation
- The representation for the CART model is a binary tree.
- Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).
- The leaf nodes of the tree contain an output variable (y) which is used to make a prediction.
- Example:
  - Dataset with two inputs (x)
    - height (centimeters) and
    - weight (kilograms)



Classification and Regression Trees (CART)

- ID3 uses information gain whereas C4.5 uses gain ratio for splitting.
- CART is an alternative decision tree building algorithm.
- CART can handle both classification and regression tasks.
- CART algorithm
  - Uses a new metric named gini index to create decision points for classification tasks.
  - Decision rules will be found by GINI index value.
- Data set
  - There are 14 instances of golf playing decisions based on outlook, temperature, humidity and wind factors.
- Gini index
  - Gini index is a metric for classification tasks in CART.
  - It stores sum of squared probabilities of each class.
  - Gini = 1 – Σ (Pi)2 for i=1 to number of classes

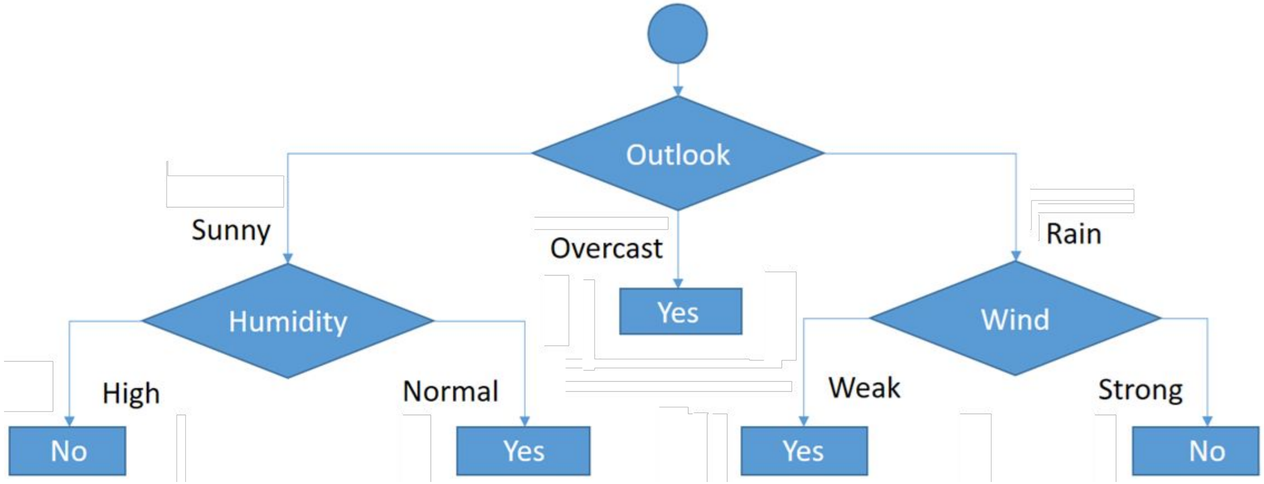
Classification and Regression Trees (CART)

- Example

Day	Outlook	Temp.	Humidity	Wind	Class
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Classification and Regression Trees (CART)

- The final decision tree is built using Gini Index as shown here.



References

Text books:

1. Ethem Alpaydin, “Introduction to Machine Learning , 4th Edition, The MIT Press, 2020.
2. Peter Harrington, “Machine Learning in Action”, 1st Edition, Dreamtech Press, 2012."
3. Tom Mitchell, “Machine Learning”, 1st Edition, McGraw Hill, 2017.
4. Andreas C, Müller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, 1ed, O'reilly, 2016.
5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, 1st Edition, MIT Press, 2012."

Reference Books:

6. Aurélien Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, 2nd Edition, Shroff/O'Reilly, 2019.
7. Witten Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal., “Data Mining: Practical machine learning tools and techniques”, 1st Edition, Morgan Kaufmann, 2016.
8. Han, Kamber, “Data Mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann, 2012.
9. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, “Foundations of Machine Learning”, 1ed, MIT Press, 2012.
10. H. Dunham, “Data Mining: Introductory and Advanced Topics”, 1st Edition, Pearson Education, 2006.

Thank You.