Lancaster University Leipzig

CART Algorithm

LZMSCI521M | Week 1

Dr Nono Saha

Content

1.1
What is CART Algorithm?

Dr Nono Saha

University of Leipzig/ ScaDS.AI

What is the CART Algorithm?

CART stands for **C**lassification and **R**egression **T**rees:

- **Classification Trees** are used when the target variable is categorical (e.g., classifying if a patient has a disease or not).
- **Regression Trees** are used when the target variable is continuous (e.g., predicting a house price).

> The CART algorithm works by recursively splitting the data into smaller and smaller subsets.

Key points

- It builds a binary tree, where each node asks a **yes/no** question about a feature, splitting the data to reduce uncertainty or error in prediction.
- The goal is to create "pure" leaf nodes that contain the most homogeneous outcomes

Lancaster University
Leipzig

1.2

How CART works and Key Ingredients

Dr Nono Saha

University of Leipzig/ ScaDS.AI

How CART works

1. **Splitting**: Starting with the entire dataset, the algorithm picks the best feature and threshold to split the data into two groups. It chooses splits that result in the greatest reduction in **impurity** for classification or **variance** for regression.
   - For classification, it measures impurity using metrics like **Gini index** or **entropy**.
   - For regression, it often uses **mean squared error** (MSE).
2. **Stopping Criteria**: CART continues splitting the data until a stopping condition is met, such as reaching a minimum node size or when further splitting doesnt significantly improve accuracy.
3. **Pruning**: CART can prune the tree to avoid overfitting. It cuts branches with little impact on the prediction, resulting in a simpler, more generalizable model.
4. **Prediction**: Once the tree is built:
   - For classification, the majority class in the leaf node is the predicted class.
   - For regression, the average value of the target variable in the leaf node is the predicted value.

Key Concepts

1. **Gini Index (for Classification)**: it measures how pure a node is. A node is pure when all of its data points belong to one class. The formula is:

$$Gini(\mathcal{D}) = 1 - \sum_{k=1}^{N} p_k^2 \tag{1}$$

where p_k is the proportion of instances in class k. The goal is to minimize the Gini index when splitting the data.

2. **Mean Squared Error (for Regression)**: it is used in regression tasks to evaluate splits. The formula is:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

where y_i is the true value, and $\hat{y}_i$ is the predicted value. The algorithm seeks to minimize MSE by finding splits that reduce error the most.

1.3
Advantages and Disadvantages

Dr Nono Saha

University of Leipzig/ ScaDS.AI

Strengths and Weaknesses of CART

Stregths:

1. Easy to interpret: CART models can be visualized as a decision tree, making them understandable to non-experts.
2. Non-parametric: CART doesnt assume a specific distribution of the data, making it flexible for many datasets.
3. Handles both classification and regression: CART can be applied to a wide range of problems.

Weaknesses:

1. Overfitting: Without Pruning, CART can create overly complex trees that don't generalize well to new data.
2. Instability: Small changes in the data can lead to a completely different tree structure.
3. Bias towards features with more splits: CART may favour features with more potential splitting points, even if they arent the most predictive.

1.4
Practical Example and Homework

Dr Nono Saha

University of Leipzig/ ScaDS.AI

Practical Example

Let us consider the following tabular data set:

CGPA (C)	Interactiveness (I)	PracticalKnowledge (P)	CommSkills (K)	Label (L)
≥ 9	Yes	Very good	Good	Yes
≥ 8	No	Good	Moderate	Yes
≥ 9	No	Average	Poor	No
< 8	No	Average	Poor	No
≥ 8	Yes	Good	Moderate	Yes
≥ 9	Yes	Good	Moderate	Yes
< 8	Yes	Good	Poor	No
≥ 9	No	Very good	Good	Yes
≥ 8	Yes	Very good	Good	Yes
≥ 8	Yes	Average	Good	Yes

Question: Construct a classification tree model using the CART algorithm.

1. **Step 1**: Compute the Gini index of the whole data $(\mathcal{D})_{1 \leq i \leq 10}$ with respect the target "Label".

$$
\begin{aligned}
Gini(\mathcal{D}) &= 1 - \sum_{k=1}^{2} p_k^2 \\
&= 1 - \left[(\frac{7}{10})^2 + (\frac{2}{10})^2 \right] \\
&= 1 - \frac{58}{100} = \frac{42}{100} = 0.42
\end{aligned}
$$

2. **Step 2**: Compute the Gini index of each feature and all possible two groups.

Solution to the Practical Example (2)

1. **Step 2**: Compute the Gini index of each feature and all possible two groups.
   Let us consider the feature: **CGPA**

CGPA (C)	Num Class	$L = Yes$	$L = No$
>=9	0	03	01
>=8	1	04	00
>8	2	00	02

We have three possible value for the feature **CGPA**: $C = \{0, 1, 2\}$ All the possible subsets (the power set of C) are:

$$2^C = \{(), (0), (1), (2), (0,1), (0,2), (1,2), (0,1,2)\}$$

Now, we want to find the best binary partitioning $P^* = \{S_1, S_2\}$ such that

$$Gini_{P*}(\mathcal{D}) = \min_{\forall P} Gini_P(\mathcal{D})$$

Solution to the Practical Example (2)

We have

$$Gini_P(\mathcal{D}) = \frac{|S_1|}{|\mathcal{D}|} Gini(S_i) + \frac{|S_2|}{|\mathcal{D}|} Gini(S_j) \tag{3}$$

Note that $S_1 \cup S_2 = C$ and ecah partition $P = (S_1, S_2)$ corresponds to a particular splitting. The possible binary partitions of C are:

- $S_1 = \{0\}, S_2 = \{1, 2\}, P_1 = \{S_1, S_2\}$

$$\begin{aligned} Gini(S_1) &= 1 - \sum_{k=1}^{2} p_k^2 \\ &= 1 - \left[(\frac{3}{4})^2 + (\frac{1}{4})^2 \right] \\ &= 1 - \frac{10}{16} = \frac{6}{16} = 0.375 \end{aligned}$$

Solution to the Practical Example (2)

- $S_1 = \{0\}, S_2 = \{1, 2\}, P_1 = \{S_1, S_2\}$

$$
\begin{aligned}
Gini(S_2) &= 1 - \sum_{k=1}^{2} p_k^2 \\
&= 1 - \left[(\frac{4}{6})^2 + (\frac{2}{6})^2 \right] \\
&= 1 - \frac{20}{36} = \frac{16}{36} = 0.44
\end{aligned}
$$

Then,

$$
\begin{aligned}
Gini_{P_1}(\mathcal{D}) &= \frac{|S_i|}{|\mathcal{D}|} Gini(S_i) + \frac{|S_j|}{|\mathcal{D}|} Gini(S_j) \\
&= \frac{4}{10} \times 0.375 + \frac{6}{10} \times 0.44 = 0.414
\end{aligned}
$$

- $S_1 = \{0, 1\}, S_2 = \{2\}, P_2 = \{S_1, S_2\}$

$$\begin{aligned} Gini(S_1) &= 1 - \textstyle\sum_{k=1}^{2} p_k^2 \\ &= 1 - \left[(\tfrac{7}{8})^2 + (\tfrac{1}{8})^2\right] \\ &= 1 - \tfrac{50}{64} = \tfrac{14}{64} = 0.218 \end{aligned} \qquad \Bigg| \qquad \begin{aligned} Gini(S_2) &= 1 - \textstyle\sum_{k=1}^{2} p_k^2 \\ &= 1 - \left[(\tfrac{0}{2})^2 + (\tfrac{2}{2})^2\right] \\ &= 1 - 1 = 0 \end{aligned}$$

Then,

$$\begin{aligned} Gini_{P_2}(\mathcal{D}) &= \frac{|S_1|}{|\mathcal{D}|} Gini(S_2) + \frac{|S_1|}{|\mathcal{D}|} Gini(S_2) \\ &= \frac{8}{10} \times 0.0.218 + \frac{2}{10} \times 0 = 0.175 \end{aligned}$$

- $S_1 = \{1\}, S_2 = \{0, 2\}, P_3 = \{S_1, S_2\}$

$$\begin{aligned} Gini(S_1) &= 1 - \textstyle\sum_{k=1}^{2} p_k^2 \\ &= 1 - \left[(\tfrac{4}{4})^2 + (\tfrac{0}{4})^2\right] \\ &= 1 - 1 = 0 \end{aligned} \qquad \Bigg| \qquad \begin{aligned} Gini(S_2) &= 1 - \textstyle\sum_{k=1}^{2} p_k^2 \\ &= 1 - \left[(\tfrac{3}{6})^2 + (\tfrac{3}{6})^2\right] \\ &= 1 - \tfrac{1}{2} = 0.5 \end{aligned}$$

Then,

$$\begin{aligned} Gini_{P_3}(\mathcal{D}) &= \frac{|S_1|}{|\mathcal{D}|} Gini(S_1) + \frac{|S_j|}{|\mathcal{D}|} Gini(S_2) \\ &= \frac{4}{10} \times 0 + \frac{6}{10} \times 0.5 = 0.3 \end{aligned}$$

3 **Step 3**: Choose the best splitting subset for the feature **CGPA**.

Therefore, we have:

Partitions	Gini Index ($G_{P_i}(\mathcal{D})$)
$P_1 = \{\{0\}, \{1,2\}\}$	0.414
$P_2 = \{\{0,1\}, \{2\}\}$	0.175
$P_3 = \{\{1\}, \{0,2\}\}$	0.3

We can conclude that the best possible splitting from node **CGPA** is $S_1 = \{0, 1\}$, $S_2 = \{2\}$ since

$$Gini_{P*}(\mathcal{D}) = \min_{\forall i} Gini_{P_i}(\mathcal{D}) = 0.175$$

4. **Step 4**: Compute the $\Delta Gini$ respect to the best splitting subset for the feature **CGPA**. We use the following formula

$$
\begin{aligned}
\Delta Gini_C(\mathcal{D}) &= Gini(\mathcal{D}) - Gini_{P*}(\mathcal{D}) \\
&= 0.42 - 0.175 = 0.245
\end{aligned}
$$

Similarly, we need to calculate the **Gini Index** of the features: Interactiveness, PracticalKnowledge, and CommonSkills.

HomeWork:

1. Compute the **Gini Index** of the features: Interactiveness, PracticalKnowledge, and CommonSkills using the previous four steps
2. Apply the following steps to derive the Classification Tree
3. Compare your results to the one produced by the Python Library `scikit-learn`

Solution to the Practical Example (7)

5. **Step 5**: Choose the feature with the maximum $\Delta Gini$ After computations, we will have a table similar to this

Features	Gini Index	$\Delta Gini$
CGPA (C)	0.175	0.245
Interactiveness (I)	0.368	0.052
PracticalKnowledge (P)	0.3058	0.1146
CommonSkills (K)	0.175	0.245

6. **Step 6**: Set the feature with the maximum $\Delta Gini$ as the root and set the best splitting subsets as its direct children.
7. **Step 7**: For each child nodes :
   - stop splitting if the node is pure and remove all the data points that belong to that node from ($\mathcal{D}$)
8. **Step 8**: For the reminding data points and feature repeat steps 1,2,3,4,7 till one stop criterion is satisfied.

What did we learn?

1. What is a CART algorithm
2. Example of Impurities such as Gini index
3. Practical Classification Problem and CART algorithm
4. Implementation of CART in Python

1.5
What Next?

Dr Nono Saha

University of Leipzig/ ScaDS.AI

Beyond CART algorithm

1. Random Forest: Ensemble methods that use multiple decision trees (often CART) to improve predictive performance and reduce overfitting.
2. C4.5 and C5.0: Extensions of CART that allow for multiway splits and handle categorical variables more effectively.
3. Gradient Boosting Machines (GBM): Boosting algorithms that also use decision trees (often CART) as base learners, sequentially improving model performance.

Some Important Materials

- Youtube Tutorial on CART Algorithm
- Code and Course Material