Thuật toán CART (Classification and Regression Trees)

1. Cấu trúc cây quyết định:

- **Cây nhị phân**: Cây quyết định được biểu diễn dưới dạng cây nhị phân, trong đó mỗi nút chỉ có hai nhánh. Mỗi nút đại diện cho một câu hỏi hoặc một điều kiện quyết định dựa trên đặc trưng đã chọn.
- **Nút lá**: Đầu ra cuối cùng của cây. Trong phân loại, mỗi nút lá biểu thị một lớp mục tiêu cụ thể, còn trong hồi quy, nó đại diện cho giá trị dự đoán.

2. Quy trình xây dựng cây CART:

Bước 1: Khởi tạo

- Bắt đầu với toàn bộ tập dữ liệu tại nút gốc của cây.

Bước 2: Tính toán tiêu chí chia tách

- **Phân loại**: CART sử dụng **Gini impurity** hoặc **Entropy** để đánh giá độ không chắc chắn của dữ liệu trong các nhánh.
    - *Gini impurity*: Đo độ không chắc chắn của một nút, với công thức:

$$Gini = 1 - \sum p_i^2$$

trong đó *pi* là tỷ lệ các mẫu thuộc lớp *i* tại nút đó.

- *Entropy*: Đo lường độ không chắc chắn và đa dạng của các lớp trong tập dữ liệu, tính bằng:

$$Entropy = -\sum p_i \log_2 p_i$$

trong đó *pi* là tỷ lệ các mẫu thuộc lớp *i.*

Hồi quy: CART sử dụng **Mean Squared Error (MSE)** làm tiêu chí chia tách, giúp đánh giá độ chênh lệch giữa các giá trị dự đoán và thực tế. Công thức:

$$MSE = \frac{1}{|S|} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

trong đó *yi* là giá trị mục tiêu của mẫu *i*, $\bar{y}$ là giá trị trung bình của tập *S*.

Bước 3: Chia tách dữ liệu

- Chọn đặc trưng và ngưỡng tối ưu để chia dữ liệu thành hai nhánh sao cho giảm thiểu độ không chắc chắn (sử dụng Gini impurity, Entropy, hoặc MSE).

Bước 4: Lặp lại

- Thực hiện lại các bước trên cho mỗi nút con cho đến khi đạt điều kiện dừng. Điều kiện dừng có thể là chiều cao tối đa của cây, kích thước nhỏ nhất của nút, hoặc khi dữ liệu tại nút đồng nhất.

Bước 5: Cắt tỉa cây (Pruning)

- Sử dụng phương pháp **cost-complexity pruning** để loại bỏ các nhánh không cần thiết, giúp giảm overfitting. Bước này thường được thực hiện sau khi cây được xây dựng hoàn chỉnh.

3. Đánh giá mô hình

- **Độ chính xác**: Đo lường độ chính xác của mô hình bằng cách sử dụng tập kiểm tra.
- **Cross-validation**: Kiểm tra khả năng tổng quát hóa của mô hình và giảm nguy cơ overfitting bằng cách sử dụng cross-validation.

4. Ứng dụng

- CART thường được sử dụng trong nhiều lĩnh vực như y tế (phân loại bệnh), tài chính (đánh giá rủi ro), và marketing (phân tích hành vi khách hàng) để phân loại và hồi quy.

Thuật toán ID3 (Iterative Dichotomiser 3)

1. Cấu trúc cây quyết định:

- **Cây có thể có nhiều nhánh**: Mỗi nút trong cây có thể có nhiều nhánh, với mỗi nhánh tương ứng với một giá trị của đặc trưng được chọn.
- **Nút lá**: Đại diện cho lớp mục tiêu cuối cùng.

2. Quy trình xây dựng cây ID3:

Bước 1: Khởi tạo

- Sử dụng toàn bộ tập dữ liệu làm nút gốc.

Bước 2: Tính toán độ thông tin (Information Gain)

- ID3 sử dụng **Entropy** để đo độ không chắc chắn trong tập dữ liệu:

$$Entropy(S) = -\sum p_i \log_2 p_i$$

trong đó *pi* là tỷ lệ mẫu thuộc lớp *i* trong tập *S*.

- Tính **Information Gain** cho mỗi đặc trưng. Information Gain đo lường mức độ giảm độ không chắc chắn khi chia tách dữ liệu theo một đặc trưng:

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v)$$

- trong đó:
  - S là tập dữ liệu,
  - A là đặc trưng,
  - *Sv* là tập con của *S* trong đó đặc trưng *A* có giá trị *v*.
- Đặc trưng có **Information Gain** cao nhất sẽ được chọn làm đặc trưng chia tách cho nút hiện tại.

Bước 3: Chia tách dữ liệu

- Chia tập dữ liệu tại nút hiện tại dựa trên các giá trị của đặc trưng đã chọn. Mỗi giá trị của đặc trưng sẽ tạo thành một nhánh mới từ nút đó.

Bước 4: Lặp lại

- Thực hiện lại các bước trên cho mỗi nút con cho đến khi đạt tiêu chí dừng:
  - Khi tất cả mẫu trong nút thuộc cùng một lớp.
  - Khi không còn đặc trưng nào để chia tách hoặc không còn mẫu nào để phân chia.

3. Tiêu chí dừng

- **Đồng nhất dữ liệu**: Khi tất cả các mẫu thuộc cùng một lớp.
- **Không còn đặc trưng**: Khi không còn đặc trưng nào để chia tách thêm.
- **Dữ liệu trong nút ít hơn ngưỡng tối thiểu**: Thông thường, một số lượng nhỏ mẫu có thể không đủ để chia tách hợp lý.

4. Đánh giá mô hình

- **Độ chính xác**: Đánh giá trên tập kiểm tra để kiểm tra độ chính xác của cây quyết định.
- **Cross-validation**: Được sử dụng để kiểm tra khả năng tổng quát hóa của mô hình, đặc biệt là với các tập dữ liệu nhỏ.

5. Ứng dụng

- ID3 thường được dùng cho các bài toán phân loại như phân loại khách hàng, dự đoán bệnh tật, và phân loại văn bản.

So sánh tóm tắt giữa thuật toán CART và ID3:

Đặc điểm	CART (Classification and Regression Trees)	ID3 (Iterative Dichotomiser 3)
Cấu trúc cây	Cây nhị phân (mỗi nút có hai nhánh)	Cây đa nhánh (mỗi nút có nhiều nhánh tùy theo giá trị đặc trưng)
Tiêu chí chia tách	Gini impurity hoặc Entropy cho phân loại; MSE cho hồi quy	Entropy và Information Gain
Ứng dụng	Phân loại và hồi quy trong tài chính, y tế, marketing	Chủ yếu dùng cho phân loại, như phân tích hành vi khách hàng, phân loại văn bản
Cắt tỉa cây (Pruning)	Có cắt tỉa cây để tránh overfitting (cost-complexity pruning)	Không có cơ chế cắt tỉa cây tự động, dễ dẫn đến overfitting nếu không kiểm soát độ sâu
Khả năng áp dụng	Phân loại và hồi quy	Chỉ dùng cho phân loại