Trịnh Bá Ngọc – 20212523

**1. CART (Classification and Regression Trees)**

**a. Tổng quan về CART**

CART (Cây phân loại và hồi quy) là một trong những phương pháp cây quyết định, được đề xuất bởi Leo Breiman vào năm 1984. CART tạo ra một cây nhị phân, trong đó mỗi nút phân chia dữ liệu dựa trên một thuộc tính duy nhất và theo một ngưỡng giá trị cụ thể.

CART có thể được sử dụng trong:

* **Phân loại**: Phân lớp đối tượng thành các nhóm (nhãn) khác nhau.
* **Hồi quy**: Dự đoán các giá trị liên tục.

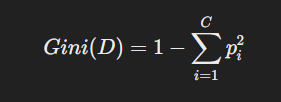
**b. Cách hoạt động của CART**

1. **Phân chia nút**:
   * CART xem xét mọi thuộc tính và mọi giá trị ngưỡng (threshold) có thể trong thuộc tính đó để phân tách tập dữ liệu tại mỗi nút.
   * CART sử dụng **Gini Index** làm tiêu chí chọn thuộc tính và ngưỡng để giảm độ hỗn loạn của nút nhiều nhất có thể.
2. **Tạo cây nhị phân**:
   * Ở mỗi nút, tập dữ liệu được chia thành hai nhánh con dựa trên thuộc tính và ngưỡng giá trị đã chọn, tạo ra cấu trúc cây nhị phân.
   * Quá trình này tiếp tục đệ quy cho đến khi đạt điều kiện dừng hoặc các nhánh chỉ còn dữ liệu thuộc một lớp duy nhất.

**c. Gini Index**

**Định nghĩa và Cách tính toán**

**Gini Index** là một độ đo trong CART để đo độ không thuần khiết của một nút. Công thức tính Gini Index cho một tập dữ liệu D là:



Trong đó:

* C là số lượng lớp trong tập dữ liệu.
* pi​ là xác suất của lớp i trong tập dữ liệu D.

**Ý nghĩa**

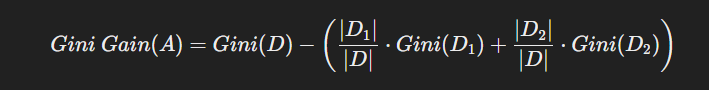
* **Gini Index = 0**: Tất cả các mẫu trong nút thuộc cùng một lớp, tức là nút này hoàn toàn thuần nhất.
* **Gini Index > 0**: Các mẫu trong nút phân phối vào nhiều lớp khác nhau, thể hiện mức độ hỗn loạn.
* **Gini Index = 0.5**: Khi dữ liệu được phân phối đồng đều giữa các lớp, Gini Index đạt giá trị tối đa.

**Cách CART sử dụng Gini Index**

* **Giảm Gini Index**: CART chọn thuộc tính và ngưỡng giá trị sao cho Gini Index của các nhánh con là nhỏ nhất.
* **Gini Gain**: Đây là mức độ giảm Gini Index khi thực hiện một phân chia. Gini Gain càng cao, mức độ thuần khiết của các nhánh con càng cao.

**Công thức Gini Gain**

Nếu một thuộc tính A chia tập D thành hai tập D1​ và D2 thì **Gini Gain** của A được tính bằng:



**d. Điều kiện dừng và Cắt tỉa cây**

1. **Điều kiện dừng**:
   * Không còn thuộc tính để phân chia.
   * Độ sâu của cây đạt ngưỡng tối đa.
   * Số lượng mẫu trong nút nhỏ hơn ngưỡng tối thiểu.
2. **Cắt tỉa cây (Pruning)**:
   * Cắt tỉa là quá trình loại bỏ các nhánh dư thừa để giảm độ phức tạp của cây và giảm quá khớp.
   * **Cắt tỉa hậu nghiệm (Post-pruning)** là phương pháp phổ biến, thực hiện sau khi xây dựng xong cây.

**2. ID3 (Iterative Dichotomiser 3)**

**a. Tổng quan về ID3**

ID3 là một thuật toán cây quyết định phân loại do Ross Quinlan phát triển vào năm 1986. ID3 xây dựng cây bằng cách chọn thuộc tính có **Information Gain** cao nhất tại mỗi nút.

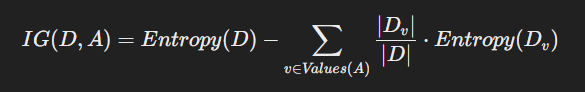
**b. Cách hoạt động của ID3**

1. **Lựa chọn thuộc tính**:
   * ID3 tính toán **Information Gain** cho từng thuộc tính và chọn thuộc tính có Information Gain cao nhất tại mỗi nút.
2. **Phân chia dữ liệu**:
   * ID3 chia tập dữ liệu thành nhiều nhánh con, mỗi nhánh con tương ứng với một giá trị của thuộc tính được chọn.
3. **Điều kiện dừng**:
   * Tất cả các mẫu trong nút thuộc cùng một lớp.
   * Tất cả các thuộc tính đã được sử dụng để phân chia.

**c. Information Gain**

**Định nghĩa và Cách tính toán**

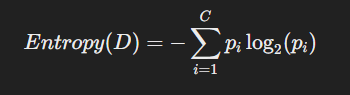
**Information Gain** là lượng thông tin thu được khi chia tập dữ liệu dựa trên một thuộc tính. Công thức tính Information Gain cho thuộc tính A trong tập dữ liệu D là:

Trong đó:

* Entropy(D): Độ đo độ hỗn loạn của tập dữ liệu D.
* Values(A) Tập hợp các giá trị có thể của thuộc tính A.
* Dv ​: Tập con của D khi thuộc tính A có giá trị v.

**Entropy**

**Entropy** là một độ đo độ không chắc chắn, được tính bằng công thức:



Trong đó:

* C: Số lớp trong tập dữ liệu.
* pi ​: Xác suất của lớp i trong tập dữ liệu D.

**Ý nghĩa của Information Gain**

* **Information Gain cao**: Cho thấy thuộc tính này giúp giảm đáng kể độ hỗn loạn của dữ liệu, tức là thuộc tính có giá trị trong việc phân chia dữ liệu.
* **Chọn thuộc tính có Information Gain cao nhất**: ID3 chọn thuộc tính có Information Gain cao nhất tại mỗi bước để phân tách dữ liệu.

**d. Các vấn đề về quá khớp và giải pháp**

1. **Ưu tiên thuộc tính có nhiều giá trị**:
   * ID3 có xu hướng chọn thuộc tính với nhiều giá trị (các thuộc tính danh mục) vì chúng thường có Information Gain cao hơn.
   * **Giải pháp**: Sử dụng tiêu chí khác như **Gain Ratio**, là biến thể của Information Gain trong thuật toán C4.5.
2. **Quá khớp (Overfitting)**:
   * ID3 dễ bị quá khớp, đặc biệt là khi cây trở nên phức tạp.
   * **Giải pháp**: Áp dụng cắt tỉa cây (pruning) hoặc giới hạn độ sâu của cây.

### Tóm tắt

* **CART với Gini Index**: Phù hợp với các bài toán phân loại và hồi quy, cây nhị phân, hiệu quả cho dữ liệu lớn nhưng dễ quá khớp.
* **ID3 với Information Gain**: Thích hợp cho phân loại, dễ hiểu, nhưng nhạy cảm với các thuộc tính có nhiều giá trị và cũng dễ quá khớp.

Hai thuật toán này, với các độ đo tương ứng, đều có ưu điểm và hạn chế riêng, phù hợp với các loại dữ liệu và mục đích khác nhau.