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

**CART** là một thuật toán cây quyết định phát triển bởi Breiman và cộng sự, được sử dụng cho cả bài toán phân loại (Classification) và hồi quy (Regression). Nó xây dựng một cây nhị phân (binary tree), nghĩa là mỗi nút phân nhánh ra tối đa hai nhánh con.

* **Tiêu chí phân chia**: CART sử dụng **Chỉ số Gini** (Gini Index) cho bài toán phân loại và **Sai số Bình phương Trung bình** (Mean Squared Error) cho bài toán hồi quy để đánh giá mức độ "tinh khiết" của các tập con sau khi phân chia.
  + **Chỉ số Gini** đo lường xác suất mà một điểm dữ liệu ngẫu nhiên sẽ bị phân loại sai nếu được gán nhãn ngẫu nhiên từ tập dữ liệu.
  + Giá trị của Chỉ số Gini dao động từ 0 đến 0,5. Một tập dữ liệu có Chỉ số Gini là 0 thể hiện tính thuần nhất tuyệt đối (tất cả dữ liệu thuộc một lớp). Ngược lại, Chỉ số Gini càng cao thì độ tinh khiết của dữ liệu sau phân chia càng thấp.
* **Phương pháp phân chia**: CART thực hiện chia nhị phân tại mỗi nút, có nghĩa là nó chỉ tạo ra hai nhánh con cho mỗi quyết định. Do đó, dù tập dữ liệu gốc có bao nhiêu lớp (class) thì mỗi lần phân chia sẽ chỉ tạo ra hai tập con.
* **Quy trình xây dựng cây**:
  + Tại mỗi nút, CART duyệt qua tất cả các đặc trưng và các ngưỡng cắt để tìm ra điểm phân chia tối ưu dựa trên Chỉ số Gini hoặc Sai số Bình phương Trung bình.
  + Sau đó, cây tiếp tục được xây dựng bằng cách chia đôi dữ liệu dựa trên đặc trưng và ngưỡng cắt được chọn cho đến khi đạt đến độ sâu tối đa hoặc một tiêu chí dừng khác.
* **Ưu điểm và Nhược điểm**:
  + **Ưu điểm**: Cấu trúc nhị phân của CART đơn giản hơn cho việc triển khai và có thể hoạt động tốt trên cả dữ liệu số và dữ liệu phân loại. Nó cũng dễ dàng sử dụng trong việc dự báo cả các giá trị liên tục và phân loại.
  + **Nhược điểm**: CART có thể dẫn đến các cây rất sâu và phức tạp nếu không có tiêu chí dừng tốt, dẫn đến hiện tượng overfitting (quá khớp).

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

**ID3** là một thuật toán cây quyết định được phát triển bởi Ross Quinlan, chủ yếu được sử dụng cho bài toán phân loại. ID3 sử dụng khái niệm **Độ lợi Thông tin** (Information Gain), một thước đo dựa trên Entropy, để lựa chọn cách chia dữ liệu tại mỗi nút của cây.

* **Tiêu chí phân chia**: ID3 chọn đặc trưng tại mỗi nút dựa trên **Độ lợi Thông tin**. Độ lợi Thông tin xác định sự giảm độ bất định trong phân phối lớp khi một đặc trưng được sử dụng để phân chia dữ liệu.
  + **Entropy** là một thước đo mức độ hỗn loạn hoặc không chắc chắn trong dữ liệu. Giá trị entropy cao biểu thị tập dữ liệu hỗn loạn, không thuần nhất, và ngược lại, entropy bằng 0 biểu thị dữ liệu thuần nhất (tất cả điểm dữ liệu thuộc cùng một lớp).
  + **Độ lợi Thông tin** là mức giảm của entropy khi phân chia dựa trên một đặc trưng. Đặc trưng nào có Độ lợi Thông tin lớn nhất sẽ được chọn để phân chia.
* **Phương pháp phân chia**: ID3 có thể thực hiện phân chia đa nhánh (multisplits) tại mỗi nút, tức là số nhánh có thể lớn hơn hai nếu đặc trưng được sử dụng để phân chia có nhiều giá trị.
* **Quy trình xây dựng cây**:
  + ID3 tính entropy cho tập dữ liệu hiện tại.
  + Sau đó, nó tính Độ lợi Thông tin cho mỗi đặc trưng và chọn đặc trưng có Độ lợi Thông tin lớn nhất để phân chia tại nút.
  + Quy trình này tiếp tục lặp lại cho đến khi cây đạt đến một tiêu chí dừng, như tất cả các điểm dữ liệu thuộc một lớp, hoặc không còn đặc trưng nào để phân chia.
* **Ưu điểm và Nhược điểm**:
  + **Ưu điểm**: ID3 có thể tạo ra cây quyết định đa nhánh, giúp giảm độ sâu của cây và tạo cấu trúc cây phân loại nhanh chóng.
  + **Nhược điểm**: ID3 thường gặp khó khăn khi làm việc với các đặc trưng số, vì nó không hỗ trợ phân chia nhị phân. Ngoài ra, vì Độ lợi Thông tin ưa chuộng các đặc trưng có nhiều giá trị, ID3 dễ dẫn đến hiện tượng overfitting nếu không được xử lý cẩn thận.

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **CART** | **ID3** |
| **Tiêu chí phân chia** | Chỉ số Gini (phân loại), MSE (hồi quy) | Độ lợi Thông tin |
| **Phương pháp phân chia** | Nhị phân | Đa nhánh |
| **Ứng dụng** | Phân loại và Hồi quy | Chủ yếu cho Phân loại |
| **Kiểu cây tạo ra** | Cây nhị phân | Cây đa nhánh |
| **Ưu điểm** | Đơn giản, dễ triển khai | Tạo cây nhanh, đa nhánh ít phức tạp |
| **Nhược điểm** | Cây phức tạp nếu không kiểm soát độ sâu | Overfitting với đặc trưng nhiều giá trị |