Lý thuyết

1. Thuật toán CART (Classification and Regression Trees)

Cấu trúc Cây Quyết Định

- **Cây nhị phân**: Mỗi nút trong cây chỉ có hai nhánh.
- **Nút**: Mỗi nút đại diện cho một quyết định hoặc một câu hỏi về một đặc trưng.
- **Nút lá**: Đầu ra cuối cùng (kết quả phân loại hoặc giá trị hồi quy).

Quy trình Xây dựng Cây

1. **Khởi tạo**: Bắt đầu với toàn bộ tập dữ liệu như một nút gốc.
2. **Tính toán Tiêu chí Chia**:
    - Đối với phân loại, CART sử dụng **Gini impurity** hoặc **Entropy**.
        - **Gini impurity**:

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

Trong đó pip_ipi là tỷ lệ mẫu thuộc lớp iii.
 Entropy (tương tự như trong ID3):

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Đối với hồi quy, CART sử dụng **Mean Squared Error (MSE)**:

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

1.
    - Trong đó y¯\bar{y}y¯ là giá trị trung bình của các giá trị mục tiêu trong tập SSS.
2. **Chia tách Dữ liệu**:
    - Chọn đặc trưng và ngưỡng tối ưu để chia tách dữ liệu thành hai tập con sao cho giảm thiểu độ không chắc chắn (Gini impurity hoặc MSE).
3. **Lặp lại**:

    - Thực hiện lại bước 2 và 3 cho mỗi nút con cho đến khi đạt điều kiện dừng.
4. **Cắt tỉa Cây**:
    - Sử dụng phương pháp như chi phí cắt tỉa (cost-complexity pruning) để loại bỏ các nhánh không cần thiết nhằm giảm overfitting. Cắt tỉa thường được thực hiện sau khi cây đã được xây dựng hoàn chỉnh.

Đánh giá Cây

- **Độ chính xác (Accuracy)**: Sử dụng tập kiểm tra để đánh giá độ chính xác của mô hình.
- **Cross-validation**: Thực hiện để kiểm tra khả năng tổng quát của mô hình.

Ứng dụng

- CART được sử dụng trong nhiều lĩnh vực như tài chính, y tế, marketing, và khoa học dữ liệu để phân loại và dự đoán.

---

2. Thuật toán ID3 (Iterative Dichotomiser 3)

Cấu trúc Cây Quyết Định

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

Quy trình Xây dựng Cây

1. **Khởi tạo**: Tập dữ liệu đầu vào.
2. **Tính toán Độ Thông tin**:
    - Sử dụng **Entropy** để đo lường độ không chắc chắn trong tập dữ liệu.
    - **Gain Information:**

$$Gain(S, A) = H(S) - \sum_{v \in A} \frac{|S_v|}{|S|} H(S_v)$$

    - Chọn đặc trưng AAA có Gain Information cao nhất để chia tách.
2. **Chia tách Dữ liệu**:
    - Chia tập dữ liệu dựa trên giá trị của đặc trưng đã chọn.
3. **Lặp lại**:

- Thực hiện lại quy trình cho từng nút cho đến khi đạt một tiêu chí dừng, chẳng hạn như:
  - Nút lá có tất cả mẫu cùng lớp.
  - Không còn đặc trưng nào để chia tách.

Tiêu chí Dừng

- Không còn dữ liệu hoặc không còn đặc trưng.
- Nút chứa một số lượng mẫu nhỏ hơn ngưỡng tối thiểu.

Đánh giá Cây

- **Độ chính xác**: Kiểm tra độ chính xác bằng cách sử dụng tập kiểm tra.
- **Cross-validation**: Đánh giá khả năng tổng quát.

Ứng dụng

- ID3 thường được sử dụng trong các ứng dụng phân loại như phân tích khách hàng, dự đoán bệnh tật, và phân loại văn bản.

Tóm tắt So sánh CART và ID3

Đặc điểm	CART	ID3
Kiểu cây	Cây nhị phân	Cây có thể có nhiều nhánh
Phương pháp phân chia	Gini impurity hoặc MSE	Gain Information (Entropy)
Khả năng hồi quy	Có	Không
Cắt tỉa	Có	Không (dễ bị overfitting)
Ứng dụng	Rộng rãi trong nhiều lĩnh vực	Chủ yếu cho phân loại

Thực hành

https://github.com/truonggxuan/thuchanh.git