Báo cáo thực tập ngày 25/06/2024

Nguyễn Tiến Đạt - Thực tập sinh AI

1 Decision Tree

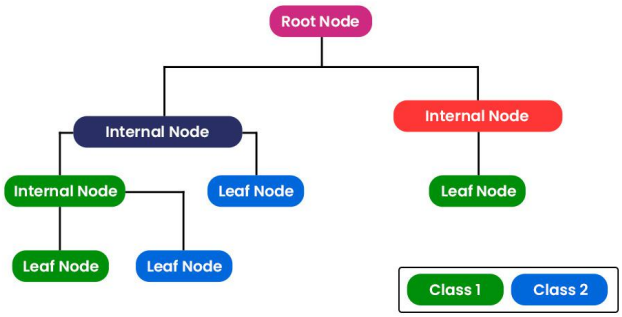
Cây quyết định (Decision Tree) là một đồ thị của các quyết định và các hậu quả có thể của nó (bao gồm rủi ro và hao phí tài nguyên). Cây quyết định được sử dụng để xây dựng một kế hoạch nhằm đạt được mục tiêu mong muốn. Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định. Cây quyết định là một dạng đặc biệt của cấu trúc cây.

1.1 Học máy

Trong lĩnh vực học máy, cây quyết định là một thuật toán thuộc loại Supervised Learning, phương pháp học có giám sát, kết quả hay biến mục tiêu của cây quyết định chủ yếu là biến phân loại. Các thuật toán được xây dựng giống hình dạng một các cây có ngọn cây, thân cây, lá cây kết nối bằng các cành cây, và mỗi thành phần đều có ý nghĩa riêng của nó, như các yếu tố tác động lên quyết định sau cùng.

Một cây quyết định bao gồm:

- "Root node": điểm ngọn chứa giá trị của biến đầu tiên được dùng để phân nhánh.
- "Internal node": các điểm bên trong thân cây là các biến chứa các giá trị dữ liệu được dùng để xét cho các phân nhánh tiếp theo
- "Leaf node": là các lá cây chứa giá trị của biến phân loại sau cùng.
- "Branch" là quy luật phân nhánh, nói đơn giản là mối quan hệ giữa giá trị của biến độc lập (Internal node) và giá trị của biến mục tiêu (Leaf node).



Hình 1: Minh hoạ cây quyết định

1.2 CART

CART (Classification And Regression Trees - Cây phân loại và hồi quy) là một biến thể của thuật toán cây quyết định. Nó có thể xử lý cả hai nhiệm vụ: phân loại và hồi quy. Scikit-Learn sử dụng thuật toán CART để huấn luyện cây quyết định.

Thuật ngữ CART được sử dụng như một thuật ngữ chung cho các loại cây quyết định sau:

- **Cây hồi quy (Regression tree)**: ước lượng các hàm giá có giá trị là số thực thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện)
- **Cây phân loại (Classification tree)**: nếu y là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua).

1.2.1 Triển khai CART

Thuật toán CART hoạt động theo quy trình sau:

- Điểm phân chia tốt nhất (best-split point) của mỗi đầu vào được xác định.
- Dựa trên các điểm phân chia tốt nhất của mỗi đầu vào trong Bước 1, điểm phân chia "tốt nhất" mới được xác định.
- Chia đầu vào đã chọn theo điểm phân chia "tốt nhất".
- Tiếp tục chia cho đến khi một quy tắc dừng được thỏa mãn hoặc không còn sự phân chia mong muốn nào khác có thể thực hiện.

Thuật toán CART sử dụng Gini Impurity để chia tập dữ liệu thành một cây quyết định. Nó thực hiện điều này bằng cách tìm kiếm độ đồng nhất tốt nhất cho các nút con, với sự trợ giúp của chỉ số Gini.

1.2.2 Gini Impurity

Chỉ số Gini là một chỉ số đo lường cho các nhiệm vụ phân loại trong CART. Nó lưu trữ tổng bình phương xác suất của mỗi lớp. Nó tính toán mức độ xác suất của một biến cụ thể bị phân loại sai khi được chọn ngẫu nhiên và là một biến thể của hệ số Gini. Nó hoạt động trên các biến phân loại, cung cấp kết quả là "thành công" hoặc "thất bại" và do đó chỉ thực hiện phân chia nhị phân.

Mức độ của chỉ số Gini dao động từ 0 đến 1:

- 0 biểu thị rằng tất cả các phần tử đều thuộc về một lớp nhất định, hoặc chỉ tồn tại một lớp.
- 1 biểu thị rằng tất cả các phần tử được phân bổ ngẫu nhiên qua các lớp khác nhau.
- 0.5 chỉ ra rằng các phần tử được phân bổ đều vào một số lớp.

Công thức của Gini Impurity:

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

trong đó p_i là xác suất của một đối tượng được phân loại vào một lớp cụ thể.

1.3 CART với cây phân loại

Cây phân loại là một thuật toán nơi biến mục tiêu là categorical. Thuật toán CART được sử dụng để xác định "Lớp" mà biến mục tiêu có khả năng rơi vào. Cây phân loại được sử dụng khi tập dữ liệu cần được chia thành các lớp thuộc về biến phản hồi (như có hoặc không).

Cây bao gồm các nút (node), tượng trưng cho các điểm quyết định khác nhau, và các nhánh (branch), tượng trưng cho kết quả có thể của những quyết định đó. Nhãn lớp dự đoán có mặt tại mỗi nút lá (leaf node) của cây.

Thuật toán CART dùng trong cây phân loại hoạt động bằng cách tách dữ liệu huấn luyện thành các tập con nhỏ hơn và nhỏ hơn dựa trên một số tiêu chí. Mục tiêu là tách dữ liệu theo cách giảm thiểu độ không tinh khiết trong mỗi tập con. Độ không tinh khiết là một thước đo mức độ lẫn lộn của dữ liệu trong một tập con cụ thể.

- **Độ không tinh khiết Gini (Gini Impurity)**: đo xác suất phân loại nhầm một thực thể ngẫu nhiên từ một tập con được dán nhãn theo lớp đa số. Độ không tinh khiết Gini càng thấp có nghĩa là độ tinh khiết của tập con càng cao.
- **Tiêu chí phân chia**: Thuật toán CART đánh giá tất cả các phân chia tiềm năng tại mỗi nút và chọn phân chia làm giảm tốt nhất độ không tinh khiết Gini của các tập

con kết quả. Quá trình này tiếp tục cho đến khi đạt được tiêu chí dừng, như độ sâu tối đa của cây hoặc số lượng thực thể tối thiểu trong một nút lá.

1.4 CART với cây hồi quy

Cây hồi quy là một thuật toán nơi biến mục tiêu là liên tục và cây được sử dụng để dự đoán giá trị của nó. Cây hồi quy được sử dụng khi biến phản hồi là liên tục. Ví dụ, nếu biến phản hồi là nhiệt độ của ngày.

Cây bao gồm các nút đại diện cho các điểm quyết định khác nhau và các nhánh đại diện cho kết quả có thể của những quyết định đó. Các giá trị dự đoán cho biến mục tiêu được lưu trữ trong mỗi nút lá của cây.

Thuật toán CART dùng trong cây hồi quy hoạt động bằng cách chia dữ liệu huấn luyện một cách đệ quy thành các tập con nhỏ hơn dựa trên các tiêu chí cụ thể. Mục tiêu là chia dữ liệu theo cách giảm thiểu sự giảm dư lượng trong mỗi tập con.

- **Giảm dư lượng**: là một thước đo mức độ giảm trung bình bình phương của sự khác biệt giữa các giá trị dự đoán và các giá trị thực tế cho biến mục tiêu do phân chia tập con. Giảm dư lượng càng thấp, mô hình càng phù hợp với dữ liệu.
- **Tiêu chí phân chia**: CART đánh giá mọi phân chia có thể tại mỗi nút và chọn phân chia dẫn đến sự giảm lỗi dư lượng lớn nhất trong các tập con kết quả. Quá trình này được lặp lại cho đến khi đạt được tiêu chí dừng, chẳng hạn như đạt đến độ sâu tối đa của cây hoặc có quá ít thực thể trong một nút lá.

1.5 Ưu và nhược điểm

1.5.1 Ưu điểm

- **Dễ dàng hình dung và giải thích:** Biểu diễn đồ họa của nó rất trực quan để hiểu và không yêu cầu bất kỳ kiến thức nào về thống kê để giải thích.
- **Hữu ích trong khám phá dữ liệu:** Chúng ta có thể dễ dàng xác định biến quan trọng nhất và mối quan hệ giữa các biến với một cây quyết định. Nó có thể giúp chúng ta tạo ra các biến mới hoặc đặt một số tính năng vào một nhóm.
- **Ít yêu cầu làm sạch dữ liệu:** Nó khá miễn nhiễm với ngoại lệ và dữ liệu thiếu, do đó ít cần làm sạch dữ liệu hơn.
- **Kiểu dữ liệu không phải là rào cản:** Nó có thể xử lý cả dữ liệu phân loại và số liệu.

1.5.2 Nhược điểm

- **Overfitting**
- **Không phù hợp tuyệt đối cho dữ liệu liên tục:** Mất mát một số thông tin liên quan đến các biến số khi phân loại chúng vào các danh mục khác nhau do làm việc với đầu ra là trung bình cộng nhiều.
- **Biến động nhỏ trong dữ liệu có thể dẫn đến kết quả khác nhau trong cây quyết định.**