

Cây quyết định (Decision tree)

Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh
Tài liệu nội bộ

Tháng 2 năm 2020



Nội dung

- ➊ Ví dụ mở đầu
- ➋ Xây dựng cây quyết định
- ➌ Thực hành với python-Xây dựng và vẽ cây quyết định
- ➍ Điều kiện dừng
- ➎ Xử lý một số dạng dữ liệu khác

Nội dung trình bày

① Ví dụ mở đầu

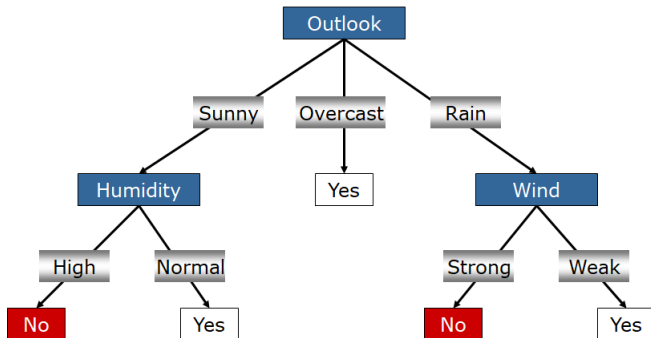
David là quản lý của một câu lạc bộ đánh golf. Anh nhận thấy: Có ngày đông người muốn chơi golf nhưng số nhân viên câu lạc bộ lại không đủ phục vụ, Có hôm lại quá ít (hoặc không có) người đến chơi dẫn đến câu lạc bộ lại thừa nhân viên phục vụ, và việc này rõ ràng bị ảnh hưởng lớn từ yếu tố thời tiết.

Do vậy, David muốn dựa vào dữ liệu thời tiết để tối ưu hóa số nhân viên phục vụ mỗi ngày. Trong hai tuần, anh ta thu thập thông tin về: Trời (outlook) (nắng (sunny), nhiều mây (overcast) hoặc mưa (raining)); nhiệt độ (temperature) bằng độ F; độ ẩm (humidity); có gió mạnh (wind) hay không; và số người chơi trong ngày (yes=đông, no=ít). David thu được một bộ dữ liệu gồm 14 dòng và 5 cột.

Day	Outlook	Temp	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Hình 1: Dữ liệu của David

Hình ảnh cây quyết định



Hình 2: Sơ đồ phân tích của anh David

**KHÁI NIỆM "GỐC", "NÚT", "NHÁNH", "LÁ", "ĐỘ SÂU", THỂ NÀO LÀ CÂY QĐ TỐT?
RA QĐ THỂ NÀO? LUẬT**

- Cây quyết định càng đơn giản càng tốt
- Cây quyết định được sử dụng trong phân lớp bằng cách duyệt từ nút gốc của cây cho đến khi đụng đến nút lá, từ đó rút ra lớp của đối tượng cần xét
- Mỗi đường dẫn đến lá trong cây tạo thành một luật
VD: (Outlook=Sunny AND Humidity=Normal) OR (Outlook=Overcast)
OR (Outlook=Rain AND Wind=Weak) = nhãn "yes"
- Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định

② Xây dựng cây quyết định

TỪ DỮ LIỆU THƠ ĐẾN CÂY QĐ?

DỰA VÀO Đâu ĐỂ XÁC ĐỊNH GỐC- Đo độ "không thuần khiết" (impurity)

DÙNG GINI HAY ENTROPY?

ĐỘ ĐO GINI - Thuật toán CART (Classification And Regression Tree)

Breiman et al., 1984

ĐỘ ĐO ENTROPY- thuật toán ID3 (Iterative Dichotomiser)

Ross Quilan, 1993

<https://link.springer.com/article/10.1007/BF00993309>

Equation 6-1. Gini impurity

$$G_i = 1 - \sum_{k=1}^n p_{i,k}^2$$

Equation 6-3. Entropy

$$H_i = - \sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^n p_{i,k} \log_2 (p_{i,k})$$

- Hai thuật toán như nhau trong đa số trường hợp
- Gini thường chạy nhanh hơn nên được mặc định trong Scikit-Learn
- Entropy thường cho cây cân bằng hơn

Hàm mất mát trong thuật toán CART

Chọn nút gốc có hàm mất mát nhỏ nhất:

$$J(k, t_k) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}}$$

where $\begin{cases} G_{\text{left/right}} & \text{measures the impurity of the left/right subset,} \\ m_{\text{left/right}} & \text{is the number of instances in the left/right subset} \end{cases}$

TÍNH ĐỘ KHÔNG THUẦN KHIẾT CỦA THUỘC TÍNH CÓ DỮ LIỆU SỐ ?

Weight	Heart Disease
220	Yes
180	Yes
225	Yes
190	No
155	No

Dữ liệu số

Hình 3

TÍNH ĐỘ KHÔNG THUẦN KHIẾT CỦA THUỘC TÍNH CÓ DỮ LIỆU SỐ

Weight	Heart Disease		Weight	Heart Disease		Weight	Heart Disease	
220	Yes		155	No	Lowest	155	No	
180	Yes		180	Yes		180	Yes	Gini impurity = 0.3
225	Yes		190	No		185	No	Gini impurity = 0.47
190	No		220	Yes		205	Yes	Gini impurity = 0.27
155	No		225	Yes	Highest	222.5	Yes	Gini impurity = 0.4

Dữ liệu số

Bước 1. Sắp xếp

Bước 2. Tính giá trị trung bình giữa các số liệu

Bước 3. Tính độ đo Gini và chọn giá trị Gini nhỏ nhất

Hình 4

ƯỚC LƯỢNG XÁC SUẤT MỘT MẪU THUỘC PHÂN LỚP

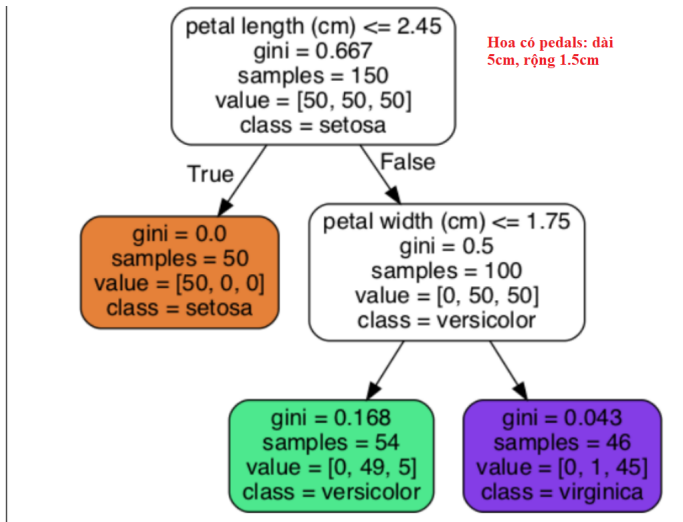


Figure 6-1. Iris Decision Tree

Hình 5

- 3 Thực hành với python-Xây dựng và vẽ cây quyết định

Sinh viên thực hành đoạn code trang 177, 178

④ Điều kiện dừng

KHI NÀO THÌ DỪNG QUÁ TRÌNH XÂY DỰNG CÂY? HIỆU CHỈNH SIÊU THAM SỐ ĐỂ TRÁNH HIỆN TƯỢNG QUÁ KHỚP (OVERFITTING)

HIỆU CHỈNH SIÊU THAM SỐ ĐỂ TRÁNH HIỆN TƯỢNG QUÁ KHỚP (OVERFITTING)

DecisionTreeClassifier có các siêu tham số (tr 184)

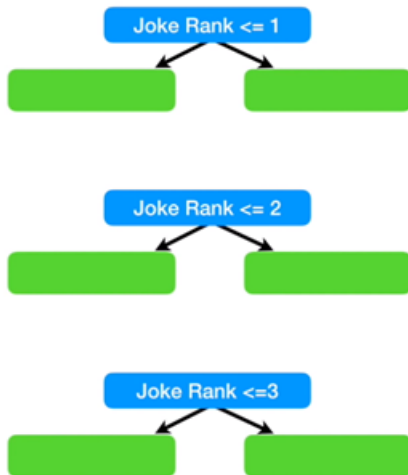
- max_depth:
- min_samples_leaf:
- min_samples_split:
-

Quá trình này gọi là thêm ràng buộc cho các siêu tham số- Regularization
Hyperparameters

- 5 Xử lý một số dạng dữ liệu khác

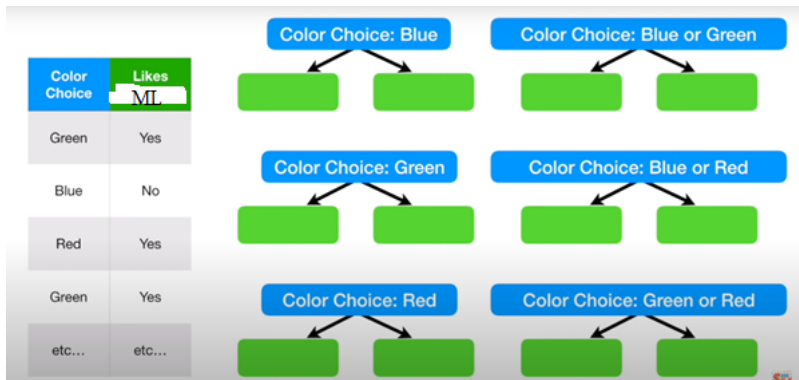
TÍNH ĐỘ KHÔNG THUẦN KHIẾT CỦA THUỘC TÍNH CÓ DỮ LIỆU ĐƯỢC XẾP HẠNG (RANKED DATA)

Rank my jokes...	Like ML
1	Yes
1	No
3	Yes
1	Yes
etc...	etc...



Hình 6

TÍNH ĐỘ KHÔNG THUẦN KHIẾT CỦA THUỘC TÍNH CÓ DỮ LIỆU NHIỀU CHỌN LỰA



Hình 7

Nhận xét Cây quyết định không đòi hỏi nhiều việc xử lý (chuẩn bị) dữ liệu, không cần co dẫn/ chuẩn hóa dữ liệu.

- 1 Giải thích ý nghĩa và cách dùng các siêu tham số của lệnh `DecisionTreeClassifier` ở trang 184
- 2 Nếu mô hình quá khớp/quá kém thì cần hiệu chỉnh tham số như thế nào?
- 3 Hiệu chỉnh đoạn CT tr 177,178 để xây dựng cây quyết định từ dữ liệu của anh David ở phần đầu tài liệu này.
- 4 Làm bài tập 7 trang 189.
- 5 Đọc thêm về kỹ thuật tỉa cành (pruning)
- 6 Đọc thêm bài toán hồi quy dựa trên cây quyết định (tr 185)

- Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd Edition của tác giả Aurélien Géron.
- <https://www.kaggle.com/bbose71/svm-non-linear-classification>
- <https://machinelearningcoban.com/>
- <https://www.youtube.com/watch?v=7VeUPuFGJHk>