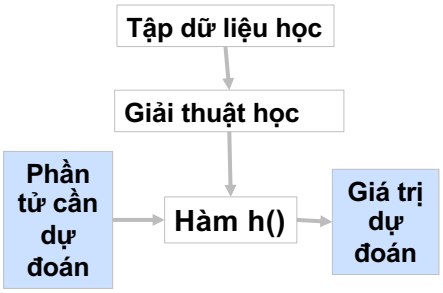
Phân loại học máy – học có giám sát

Từ tập dữ liệu huấn luyện $\{(x_1, y_1), (x_2,y_2),..,(x_m,y_m)$

- • Tìm hàm h (hypothesis) X=>Y sao cho h(x) dự báo được y từ x
- • Y là giá trị liên tục: sử dụng pp hồi quy (regression)
- • Y là giá trị rời rạc: sử dụng pp phân lớp (classification)



Quy ước

- Biến đầu vào (input variables)/đặc trưng (features), kí hiệu: $x^{(i)}$
- Biến đầu ra (output variable)/biến mục tiêu, kí hiệu y(i)
- Mẫu huấn luyện (training example)
  kí hiệu ($x^{(i)}$, $y^{(i)}$)
- Tập huấn luyện X = {($x^{(i)}$, $y^{(i)}$)}, i = 1..m

Square meters	Bedrooms	Floors	Age of building (years)	Price in 1000€
x1	*x2*	*x3*	*x4*	*y*
200	5	1	45	460
131	3	2	40	232
142	3	2	30	315
756	2	1	36	178
...	...	...	...	...

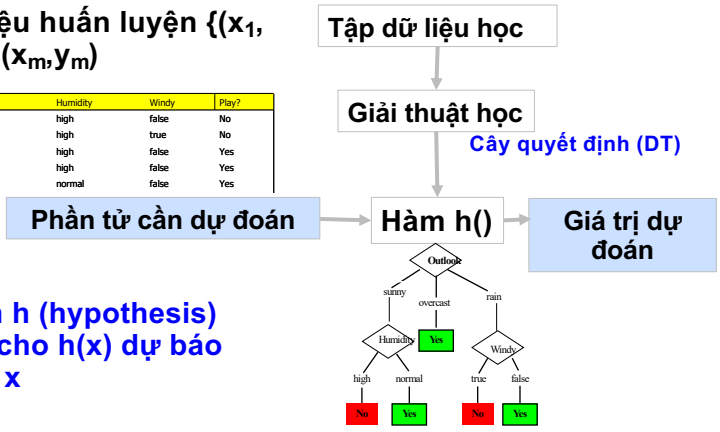
$$x^{(3)} = \begin{bmatrix} 142 \\ 3 \\ 3 \\ 0 \end{bmatrix}$$

$$x_1^{(4)} = 756$$

Phân loại học máy – học có giám sát

Từ tập dữ liệu huấn luyện {(x_1, y_1), (x_2,y_2),..,(x_m,y_m)

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes



- Tìm hàm h (hypothesis) X=>Y sao cho h(x) dự báo được y từ x

Cây quyết định

Từ tập dữ liệu học/huấn luyện {(x_1, y_1), (x_2,y_2),..,(x_m,y_m)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

[See: Tom M. Mitchell, *Machine Learning,* McGraw-Hill, 1997]

Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

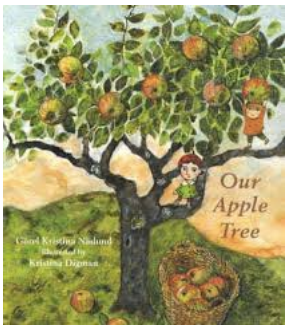
Phương pháp học cây quyết định
Decision Tree

Đỗ Thanh Nghị - *dtnghi@cit.ctu.eau.vn*
Trần Nguyễn Minh Thư – *tnmthu@cit.ctu.edu.vn*

Cần Thơ - 2015

Nội dung

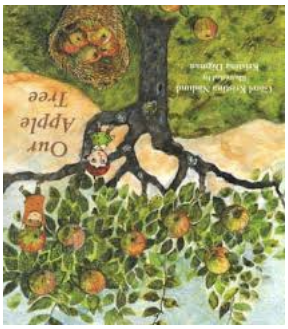
- Giới thiệu về cây quyết định
- Giải thuật học của cây quyết định
- Kết luận và hướng phát triển

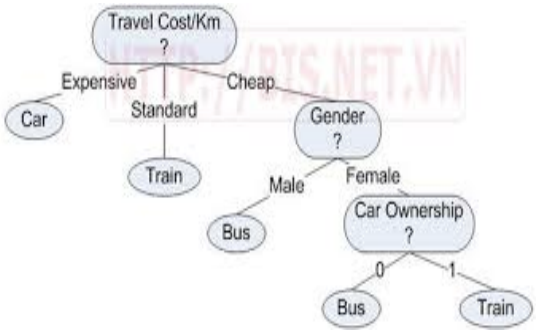


6

Nội dung

- **Giới thiệu về cây quyết định**
- Giải thuật học của cây quyết định
- Kết luận và hướng phát triển





7

Cây quyết định

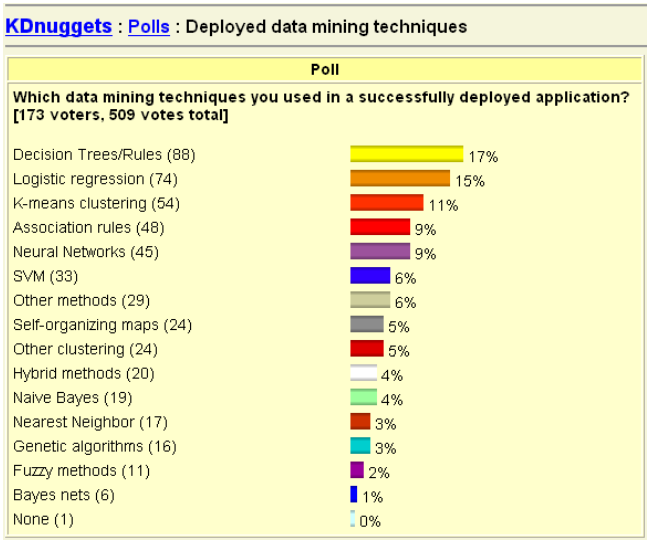
- lớp các giải thuật học
  - kết quả sinh ra dễ dịch **(if … then …)**
  - khá đơn giản, nhanh, hiệu quả được sử dụng nhiều
  - liên tục trong nhiều năm qua, cây quyết định được bình chọn là giải thuật được sử dụng nhiều nhất và thành công nhất
  - giải quyết các vấn đề của phân loại, hồi quy
  - làm việc cho **dữ liệu số và kiểu liệt kê**
  - được ứng dụng thành công trong hầu hết các lãnh vực về phân tích dữ liệu, phân loại text, spam, phân loại gien, etc

8

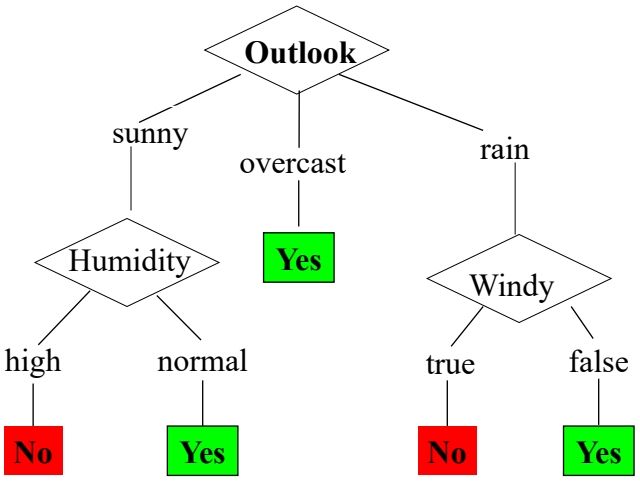
Cây quyết định

- Có rất nhiều giải thuật sẵn dùng
  - ID3 (Quinlan 79)
  - **CART – Classification and Regression Trees (Brieman et al. 84)**
  - Assistant (Cestnik et al. 87)
  - **C4.5 (Quinlan 93)**
  - See5 (Quinlan 97)
  - ...
  - Orange (Demšar, Zupan 98-03)

Kỹ thuật DM thành công trong ứng dụng thực (2004)



Example Decision Tree



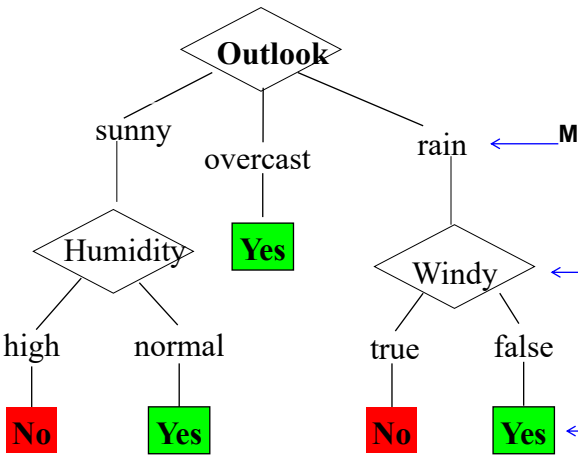
Cây quyết định

- Nút trong : được tích hợp với điều kiện để kiểm tra rẽ nhánh
- Nút lá : được gán nhãn tương ứng với lớp của dữ liệu
- 1 nhánh : trình bày cho dữ liệu thỏa mãn điều kiện kiểm tra, ví dụ : age < 25.
- ở mỗi nút, 1 thuộc tính được chọn để phân hoạch dữ liệu học sao cho tách rời các lớp tốt nhất có thể
- Một luật quyết định có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá.
- Dữ liệu mới đến được phân loại 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

12

Ví dụ Decision Tree

Một luật quyết định có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá.



Mỗi nhánh tương ứng với một giá trị của thuộc tính

Mỗi nút mang một thuộc tính (biến độc lập)

Mỗi nút lá là một lớp (biến phụ thuộc)