


Phân loại dữ liệu

Phân loại dữ liệu

- Tổng quan về phân loại dữ liệu
- Phân loại dữ liệu với cây quyết định
- Phân loại dữ liệu với máy học vector hỗ trợ
- Phương pháp k láng giềng (K nearest neighbors)
- Đánh giá mô hình phân loại dữ liệu
- Cải thiện độ chính xác với các phương pháp tập hợp mô hình
- Tóm tắt

Phân loại dữ liệu

- Tổng quan về phân loại dữ liệu 
- Phân loại dữ liệu với cây quyết định
- Phân loại dữ liệu với máy học vector hỗ trợ
- Phương pháp k láng giềng (K nearest neighbors)
- Đánh giá mô hình phân loại dữ liệu
- Cải thiện độ chính xác với các phương pháp tập hợp mô hình
- Tóm tắt

Tổng quan về phân loại dữ liệu

- Dạng phân tích dữ liệu nhằm rút trích các mô hình mô tả các lớp dữ liệu hoặc dự đoán xu hướng dữ liệu.
- Ví dụ: phần mềm lọc thư rác
 - Cần có 1 tập dữ liệu thư điện tử. Mỗi thư có nhãn là thư rác hay thư bình thường.
 - Xây dựng mô hình phân loại thư điện tử để khi có một thư điện tử mới đến thì mô hình dự báo được thư này có phải là thư rác hay không.

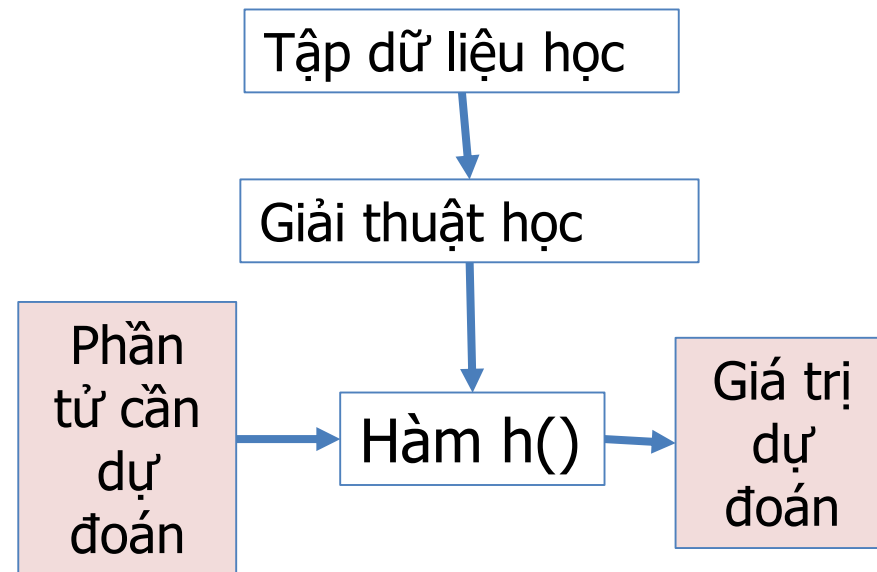
Tổng quan về phân loại dữ liệu

- Quá trình gồm 2 bước
 - Bước xây dựng mô hình: xây dựng bộ phân loại (classifier) bằng việc phân tích/học tập huấn luyện.
 - Bước phân loại: phân loại dữ liệu/đối tượng mới nếu độ chính xác của bộ phân loại được đánh giá là có thể chấp nhận được.

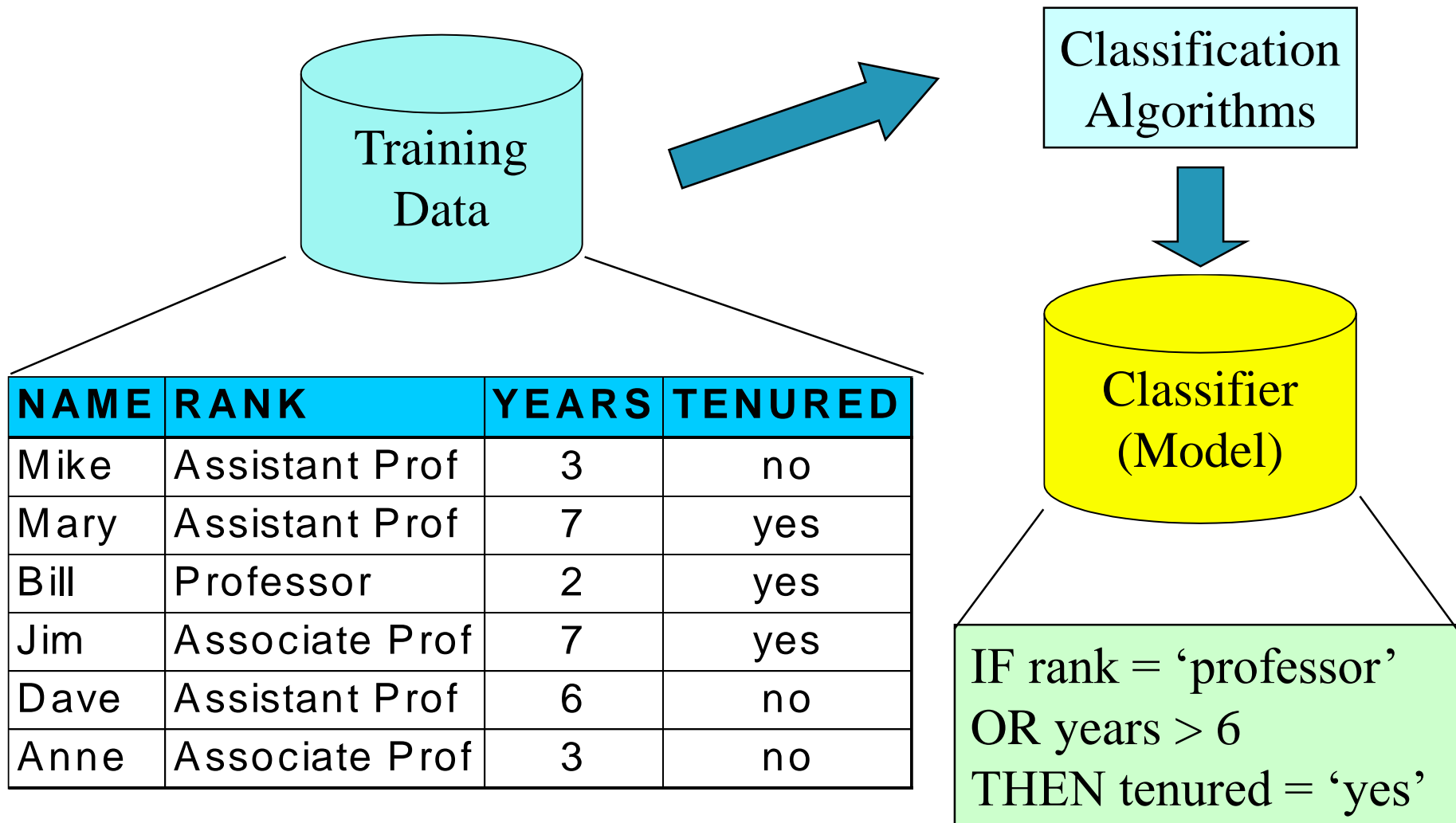
Tổng quan về phân loại dữ liệu

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

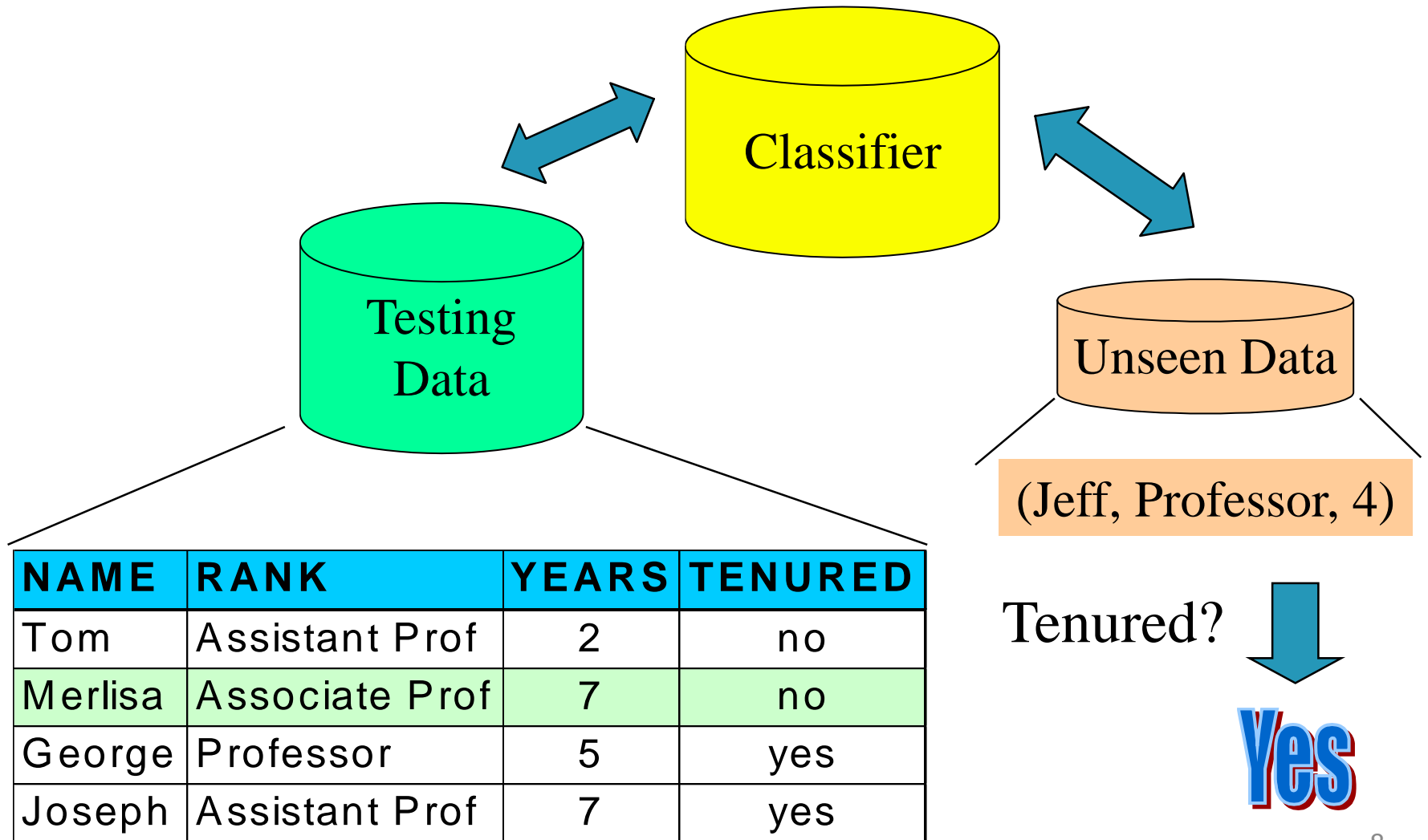
- Tìm hàm h (hypothesis) $X \Rightarrow Y$ sao cho $h(x)$ dự báo được y từ x
- Y là giá trị liên tục: sử dụng pp hồi quy (regression)
- Y là giá trị rời rạc: sử dụng pp phân lớp (classification)



Bước 1: xây dựng mô hình



Bước 2: phân loại dữ liệu



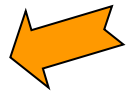
Tổng quan về phân loại dữ liệu

- Các giải thuật phân loại dữ liệu
 - Cây quyết định (decision tree)
 - Máy học vectơ hỗ trợ (svm)
 - Mạng Bayesian
 - Bayes thơ ngây
 - Mạng neural
 - K phần tử cận gần nhất (k-nearest neighbor)
 - Suy diễn dựa trên tình huống (case-based reasoning)
 - Giải thuật di truyền (genetic algorithms)
 - ...

Tổng quan về phân loại dữ liệu

- Các tiêu chí lựa chọn giải thuật
 - Mô hình cần dễ hiểu hay không
 - Độ chính xác của mô hình
 - Thời gian xây dựng mô hình
 - Thời gian dự đoán
 - Đặc tính của dữ liệu như kiểu dữ liệu, số lượng phần tử, số lượng chiều.

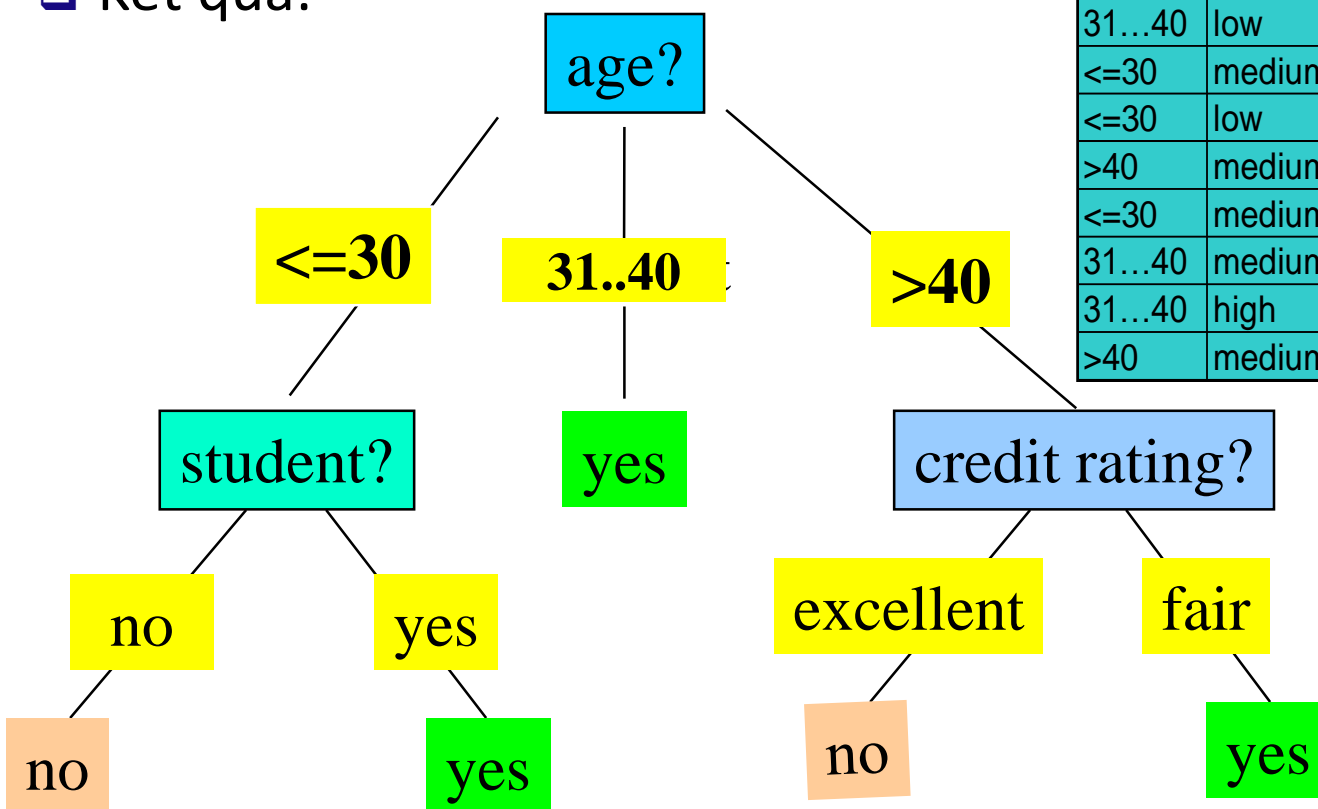
Phân loại dữ liệu

- Tổng quan về phân loại dữ liệu
- Phân loại dữ liệu với cây quyết định 
- Phân loại dữ liệu với máy học vector hỗ trợ
- Phương pháp k láng giềng (K nearest neighbors)
- Đánh giá mô hình phân loại dữ liệu
- Cải thiện độ chính xác với các phương pháp tập hợp mô hình
- Tóm tắt

Cây quyết định: ví dụ

- ❑ Tập dữ liệu huấn luyện:
Buys_computer
- ❑ Giải thuật: Quinlan's ID3
- ❑ Kết quả:

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no



Cây quyết định

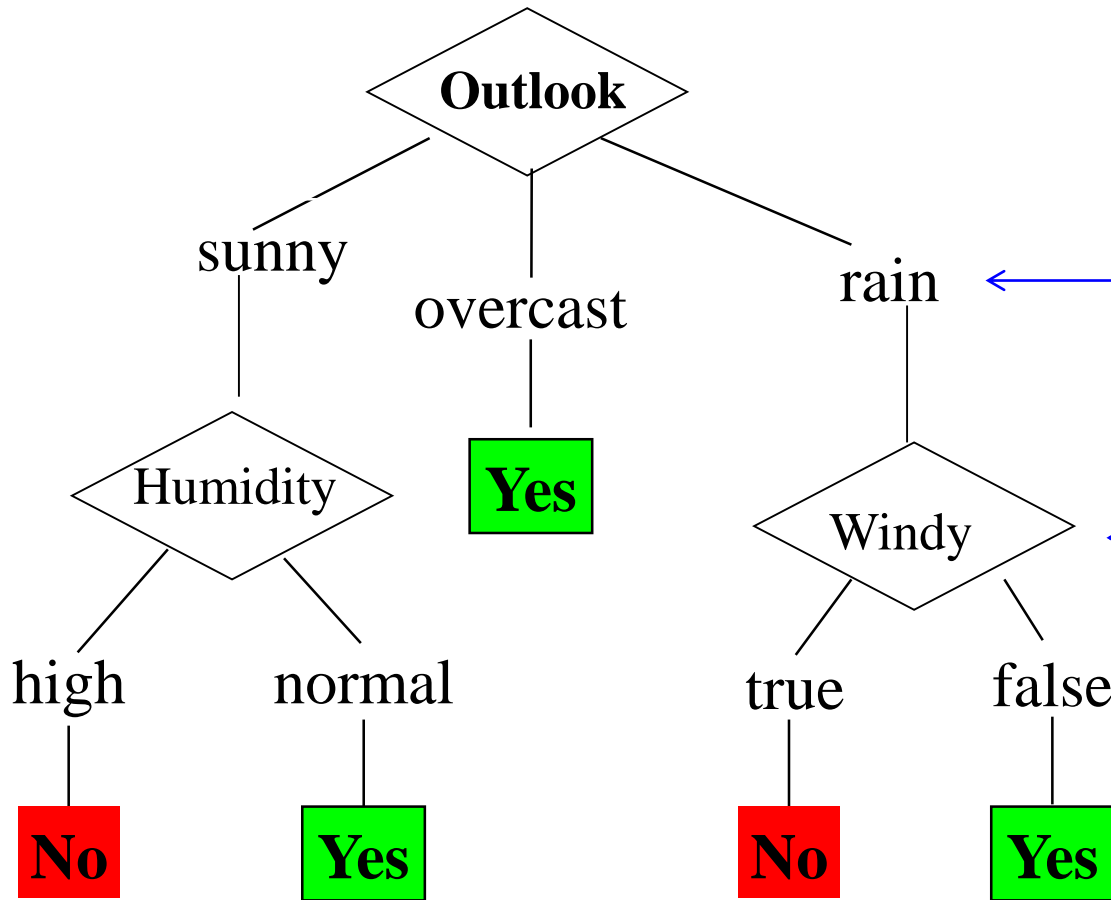
- Kết quả sinh ra dễ diễn dịch (if ... then ...)
- Khá đơn giản, nhanh, hiệu quả, được sử dụng nhiều.
- Trong nhiều năm qua, cây quyết định được bình chọn là giải thuật được sử dụng nhiều nhất và thành công nhất
- Làm việc đối với kiểu dữ liệu số và liệt kê.
- Được ứng dụng thành công trong hầu hết các lĩnh vực về phân tích dữ liệu, phân loại văn bản, thư rác, phân loại gen, ...

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: $\text{age} < 25$.
- 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 dừng đến nút lá, từ đó rút ra lớp của đối tượng cần xét

Cây quyết định

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)


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)

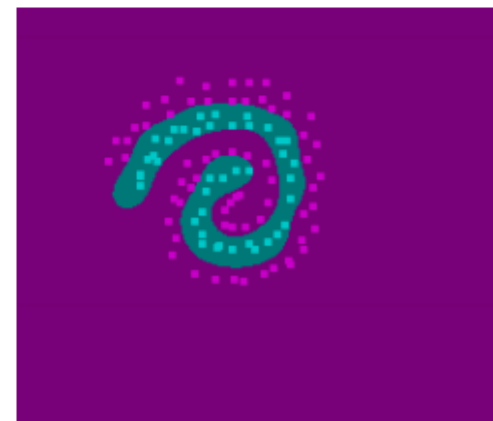
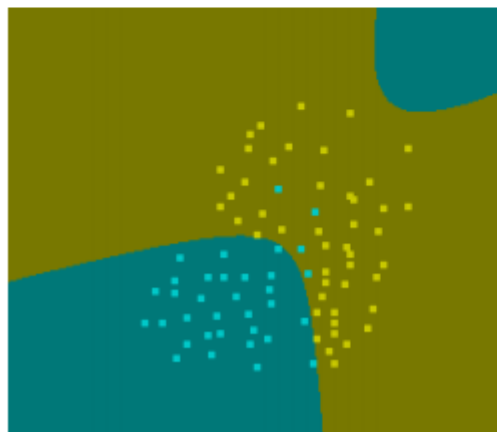
Sinh viên ôn lại giải thuật cây quyết định đã học trong năm Nguyên lý máy học.

Phân loại dữ liệu

- Tổng quan về phân loại dữ liệu
- Phân loại dữ liệu với cây quyết định
- Phân loại dữ liệu với máy học vector hỗ trợ 
- Phương pháp k láng giềng (K nearest neighbors)
- Đánh giá mô hình phân loại dữ liệu
- Cải thiện độ chính xác với các phương pháp tập hợp mô hình
- Tóm tắt

Máy học vectơ hỗ trợ

- Máy học vectơ hỗ trợ (Support vector machines)
 - Tìm siêu phẳng trong không gian N-dim để phân loại dữ liệu

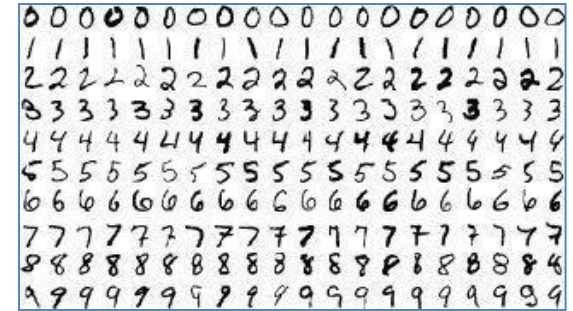


Xem chi tiết giải thuật tại slide
“Máy học vectơ hỗ trợ”
Support vector machines.

Máy học vectơ hỗ trợ

- Ứng dụng

- Nhận dạng: tiếng nói, ảnh, chữ viết tay
- Phân loại văn bản, khai mỏ dữ liệu văn bản
- Phân tích dữ liệu gen, nhận dạng bệnh, công nghệ bào chế thuốc
- Phân tích dữ liệu marketing
- ...



Happy



Sad



Surprised



Angry

Máy học vectơ hỗ trợ


- Ưu điểm

- Cho kết quả rất tốt trong thực tế, mô hình có độ chính xác cao
- Chịu đựng được nhiễu
- Hiệu quả khi xử lý dữ liệu có số thuộc tính lớn
- Thành công trong nhiều ứng dụng

- Hạn chế

- Khó dịch kết quả
- Quá trình học mô hình SVM tốn nhiều thời gian do độ phức tạp cao
- Chỉ làm việc với dữ liệu số
- Tham số SVM và hàm nhân khó điều chỉnh

Phân loại dữ liệu


- Tổng quan về phân loại dữ liệu
- Phân loại dữ liệu với cây quyết định
- Phân loại dữ liệu với máy học vector hỗ trợ
- Phương pháp k láng giềng (K nearest neighbors) 
- Đánh giá mô hình phân loại dữ liệu
- Cải thiện độ chính xác với các phương pháp tập hợp mô hình
- Tóm tắt

PP k láng giềng

- Rất đơn giản, không có quá trình học
- Khi phân loại mất nhiều thời gian, do quá trình tìm kiếm k dữ liệu lân cận. Sau đó phân loại dựa trên majority vote (hồi quy dựa trên giá trị trung bình)
- Kết quả phụ thuộc vào việc chọn khoảng cách sử dụng
- Có thể làm việc trên nhiều loại dữ liệu khác nhau.
- Được ứng dụng thành công trong hầu hết các lĩnh vực tìm kiếm thông tin, nhận dạng, phân tích dữ liệu, etc.

Xem chi tiết giải thuật tại slide
PP k láng giềng – k nearest neighbors

Phân loại dữ liệu


- Tổng quan về phân loại dữ liệu
- Phân loại dữ liệu với cây quyết định
- Phân loại dữ liệu với máy học vector hỗ trợ
- Phương pháp k láng giềng (K nearest neighbors)
- Đánh giá mô hình phân loại dữ liệu 
- Cải thiện độ chính xác với các phương pháp tập hợp mô hình
- Tóm tắt

Phương pháp đánh giá hiệu quả

- Nghi thức kiểm tra
 - K-fold
 - Hold-out
- Độ đo hiệu quả của giải thuật:
 - Độ chính xác (Accuracy)
 - Ma trận confusion
 - Đường cong Receiver Operating Characteristic
- Trường hợp dữ liệu không cân bằng (unbalanced)
- Xem chi tiết tại slide “Các phương pháp đánh giá hiệu quả phân loại dữ liệu”.

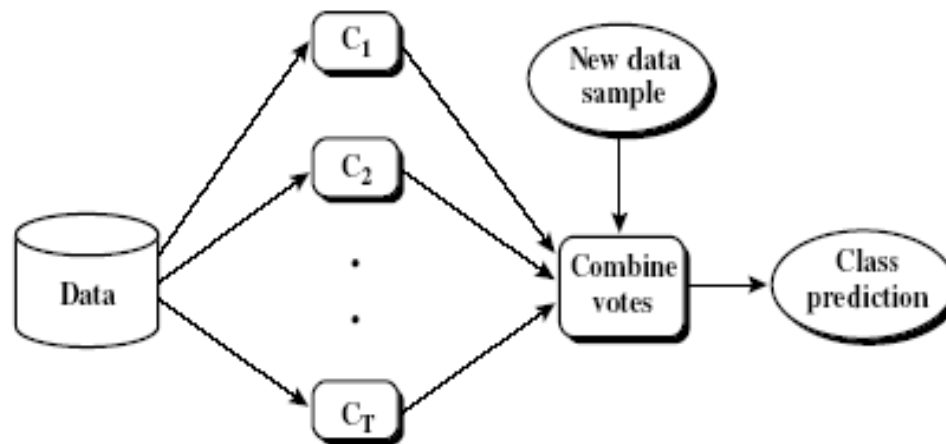
		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Phân loại dữ liệu


- Tổng quan về phân loại dữ liệu
- Phân loại dữ liệu với cây quyết định
- Phân loại dữ liệu với máy học vector hỗ trợ
- Phương pháp k láng giềng (K nearest neighbors)
- Đánh giá mô hình phân loại dữ liệu
- Cải thiện độ chính xác với các phương pháp tập hợp mô hình 
- Tóm tắt

PP tập hợp mô hình

- Cải thiện độ chính xác bằng cách kết hợp nhiều mô hình
- Các giải thuật phổ biến
 - Bagging
 - Boosting
 - Random forest (rừng ngẫu nhiên)
- Xem chi tiết các giải thuật trong slide “Phương pháp tập hợp mô hình”



Phân loại dữ liệu

- Tổng quan về phân loại dữ liệu
- Phân loại dữ liệu với cây quyết định
- Phân loại dữ liệu với máy học vector hỗ trợ
- Phương pháp k láng giềng (K nearest neighbors)
- Đánh giá mô hình phân loại dữ liệu
- Cải thiện độ chính xác với các phương pháp tập hợp mô hình
- Tóm tắt 

Tóm tắt

- Phân loại dữ liệu: xây dựng phân loại dựa trên tập dữ liệu học có nhãn (lớp).
- Lựa chọn giải thuật phù hợp
 - Một giải thuật khai mở dữ liệu được biết là tốt trong trường hợp cụ thể này thì trong trường hợp khác lại xử lý kém hiệu quả so với phương pháp khác.
 - Hầu hết các giải thuật khai mở dữ liệu hiện nay được phát triển rất đặc thù cho trường hợp cụ thể nào đó (ad-hoc).
- Dò tìm các bộ thông số phù hợp của mỗi giải thuật:
 - Sử dụng các nghi thức kiểm tra chéo
 - Thường lặp đi lặp lại cho đến khi tìm được bộ tham số của mô hình tương ứng cho kết quả cao nhất.
- Đánh giá hiệu quả phân loại dữ liệu
- Cải tiến hiệu quả phân loại

Bài đọc thêm

- Nghị, Đỗ Thanh, et al. "PHÁT HIỆN MÔN HỌC QUAN TRỌNG ẢNH HƯỞNG ĐẾN KẾT QUẢ HỌC TẬP SINH VIÊN NGÀNH CÔNG NGHỆ THÔNG TIN." *Tạp chí Khoa học Trường Đại học Cần Thơ* (2014): 49-57.
- Duong Trung, Nghia & Tan, Dang & Luu, Tien-Dao & Huynh, Hiep. (2019). Black Friday Sale Prediction via Extreme Gradient Boosted Trees.
- Trần Thanh Điện, Thái Nhựt Thanh và Nguyễn Thái Nghe, 2019. Giải pháp phân loại bài báo khoa học bằng kỹ thuật máy học. *Tạp chí Khoa học Trường Đại học Cần Thơ*. 55(4A): 29-37.
- Ngân, Võ Tuyết, and Đỗ Thanh Nghị. "PHÂN LOẠI Ý KIẾN TRÊN TWITTER." *Tạp chí Khoa học Trường Đại học Cần Thơ* (2015): 32-38.
- Đệ, Trần Cao, and Phạm Nguyên Khang. "Phân loại văn bản với Máy học vector hỗ trợ và Cây quyết định." *Tạp chí Khoa học Trường Đại học Cần Thơ* (2012): 52-63.
- Huỳnh Phụng Toàn, Nguyễn Minh Trung, Đỗ Thanh Nghị, Nguyễn Vũ Lâm, 2011. PHÂN LOẠI THƯ RÁC VỚI GIẢI THUẬT BOOSTING CÂY QUYẾT ĐỊNH NGẪU NHIÊN XIÊN PHÂN ĐƠN GIẢN. *Tạp chí Khoa học Trường Đại học Cần Thơ*. 19b: 1-9
- Toàn, Huỳnh Phụng, et al. "RỪNG NGẪU NHIÊN CẢI TIẾN CHO PHÂN LOẠI DỮ LIỆU GIEN." *Tạp chí Khoa học Trường Đại học Cần Thơ* (2012): 9-17.