Học viện Công nghệ Bưu chính Viễn thông
Khoa Công nghệ thông tin 1

Nhập môn trí tuệ nhân tạo

Giới thiệu học máy

Ngô Xuân Bách

Nội dung

- Giới thiệu
- Học cây quyết định
- Phân loại Bayes đơn giản
- Học dựa trên ví dụ

Tài liệu tham khảo

- N. Nilsson. Introduction to machine learning http://ai.stanford.edu/people/nilsson/mlbook.html
- T. Mitchell. Machine learning. McGraw-Hill, 1997.
- E. Alpaydin. Introduction to machine learning. MIT Press, 2004.
- M. Mohri, A. Rostamizadeh, A. Talwalkar. Foundations of Machine Learning. MIT Press, 2012.

Công cụ và dữ liệu

- Bộ công cụ Weka
  - http://www.cs.waikato.ac.nz/~ml/weka
- Kho dữ liệu mẫu UC Irvine
  - http://www.ics.uci.edu/~mlearn/ML/Repository.html

Một số ứng dụng của học máy (1/3)

- Những ứng dụng khó lập trình theo cách thông thường do không tồn tại hoặc khó giải thích kinh nghiệm, kỹ năng của con người
  - Nhận dạng chữ viết, âm thanh, hình ảnh
  - Lái xe tự động, thám hiểm sao Hoả
- Chương trình máy tính có khả năng thích nghi: lời giải thay đổi theo thời gian hoặc theo tình huống cụ thể
  - Chương trình trợ giúp cá nhân
  - Định tuyến mạng

Một số ứng dụng của học máy (2/3)

- Khai phá (phân tích) dữ liệu
  - Hồ sơ bệnh án → tri thức y học
  - Dữ liệu bán hàng → quy luật kinh doanh

Một số ứng dụng của học máy (3/3)

- Hầu hết các ứng dụng trí tuệ nhân tạo ngày nay có sử dụng học máy

...	...
Web search	Product recommendation
Speech recognition	Advertisement placement
Handwriting recognition	Smart-grid energy optimization
Machine translation	Household robotics
Information extraction	Robotic surgery
Document summarization	Robot exploration
Question answering	Spam filtering
Spelling correction	Fraud detection
Image recognition	Fault diagnostics
3D scene reconstruction	AI for video games
Human activity recognition	Character animation
Autonomous driving	Financial trading
Music information retrieval	Protein folding
Automatic composition	Medical diagnosis
Social network analysis	Medical imaging
...	...

Học máy là gì?

- Học:
  - …thu thập kiến thức hoặc kỹ năng…
  - "*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*" Tom Mitchell (1997)

- Học máy:
  - Giải quyết vấn đề từ kinh nghiệm
  - …được thực hiện bởi chương trình máy tính có khả năng:
    - Thực hiện công việc T tốt hơn
    - Theo tiêu chí P
    - Nhờ sử dụng dữ liệu mẫu hoặc kinh nghiệm E

Ví dụ

- Học đánh cờ
  - T: đánh cờ
  - P: số ván thắng
  - E: kinh nghiệm tự chơi
- Học nhận dạng chữ
  - T: nhận dạng chữ cái từ ảnh
  - P: phần trăm chữ nhận dạng đúng
  - E: ảnh số của chữ và chữ tương ứng
- Dịch máy
  - T: dịch một câu tiếng Anh sang tiếng Việt
  - P: độ đo dịch máy (ví dụ số câu đúng, số mệnh đề đúng,...)
  - E: cặp câu tiếng Anh và tiếng Việt tương ứng

Vấn đề cần quan tâm (1/2)

- Kinh nghiệm cụ thể như thế nào?
  - Kinh nghiệm trực tiếp và gián tiếp
    - Trực tiếp: trạng thái cụ thể + nước đi đúng tương ứng
    - Gián tiếp: toàn bộ ván cờ và kết quả
  - Có giám sát (hướng dẫn) và không giám sát
    - Có giám sát
    - Không giám sát
    - Bán giám sát
- Cần phải học cái gì? Biểu diễn kiến thức học được thế nào?
  - Tri thức cần học được biểu diễn như một hàm đích, cần lựa chọn hàm đích cụ thể
  - Ví dụ đánh cờ:
    - Chọn_nước_đi: $trạng\ thái \rightarrow nước\ đi$
    - Điểm_số: $trạng\ thái \rightarrow điểm\ số$

Vấn đề cần quan tâm (2/2)

- Sử dụng thuật toán gì để học?
  - Sử dụng hàm
    - VD: $điểm\_số = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \cdots$
  - Sử dụng các luật
  - Sử dụng mạng nơ ron
  - Sử dụng cây quyết định
  - Sử dụng các mô hình xác suất
  - …

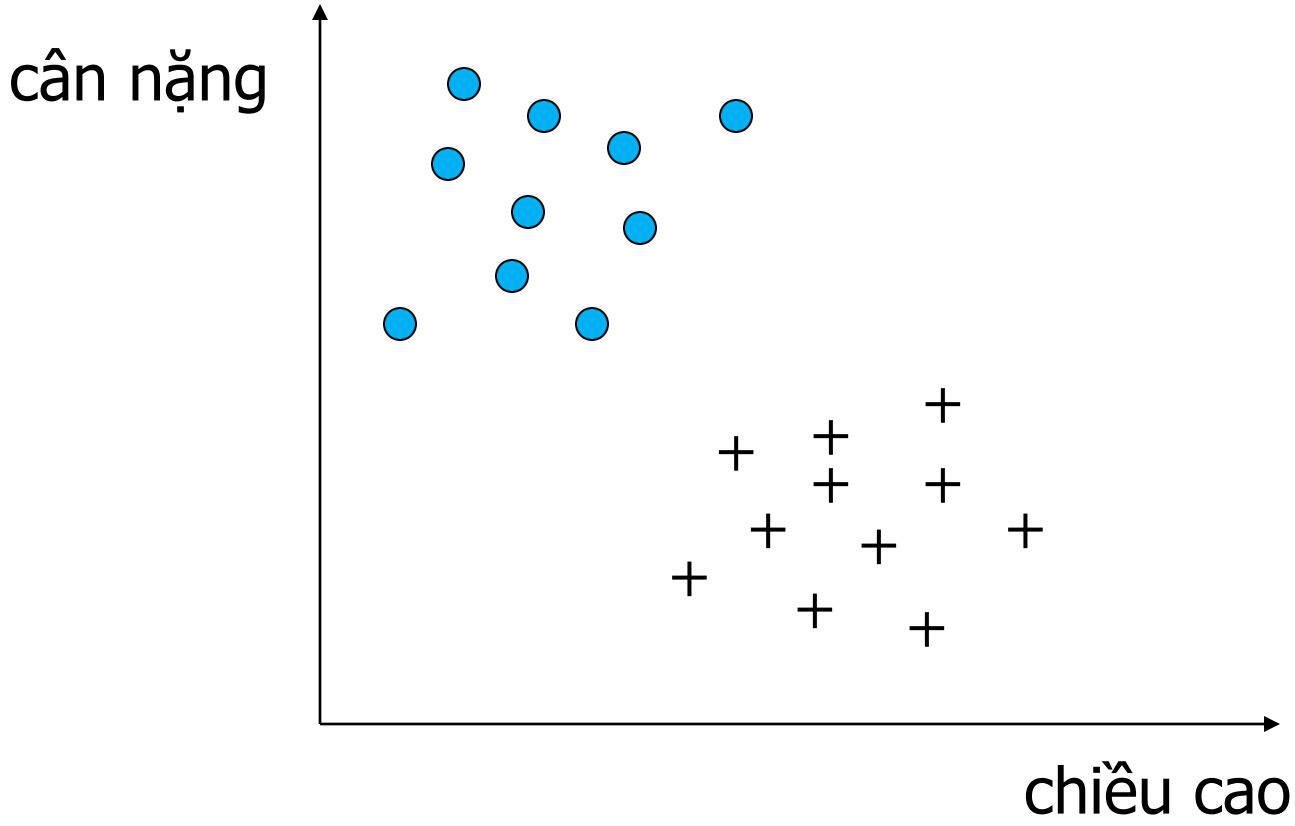
Một số khái niệm

- Mẫu, hay ví dụ (samples): là đối tượng cần xử lý (ví dụ phân loại)
  - Ví dụ: khi lọc thư rác thì mỗi thư là một mẫu
- Mẫu thường được mô tả bằng tập thuộc tính hay đặc trưng (features)
  - Ví dụ: trong chuẩn đoán bệnh, thuộc tính là triệu chứng của người bệnh, và các tham số khác như chiều cao, cân nặng, …
- Nhãn phân loại (label): thể hiện loại của đối tượng mà ta cần dự đoán
  - Ví dụ: nhãn phân loại thư rác có thể là "rác" hoặc "bình thường"

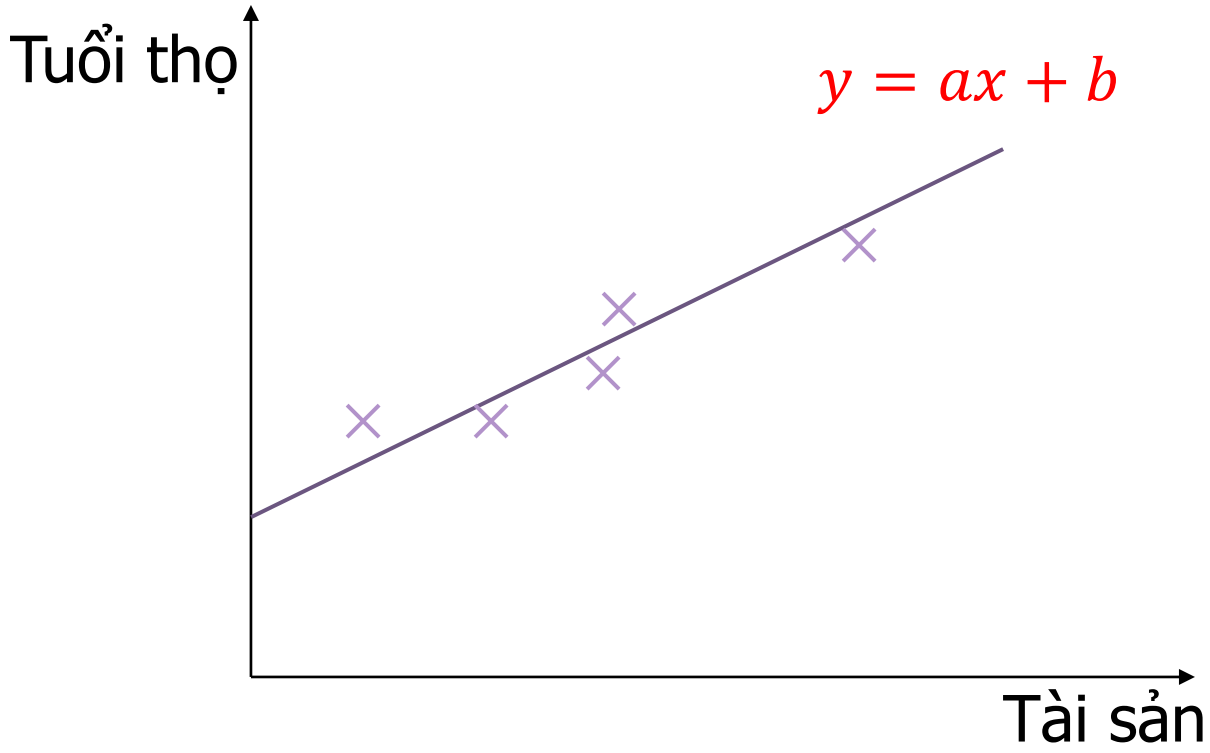
Một số dạng học máy phổ biến

- Học có giám sát (supervised learning)
  - Phân lớp (classification)
  - Hồi quy (regression)
- Học không giám sát (unsupervised learning)
  - Học luật kết hợp (association)
  - Phân cụm (clustering)
- Học bán giám sát (semi-supervised learning)
- Học tăng cường (reinforcement learning)

Phân lớp



Hồi quy (regression)



Ứng dụng: dự đoán giá cả, lái xe,...

Học luật kết hợp

- Ví dụ
  - Phân tích giao dịch, mua bán (hóa đơn mua hàng)
- $P(Y|X)$
  - Xác suất người mua hàng X còn mua hàng Y
- Ví dụ luật kết hợp
  - Người mua bánh mì thường mua bơ
  - Người mua lạc rang thường mua bia

Phân cụm

- Nhóm những trường hợp tương tự với nhau
- Không có giá trị đầu ra
- Ứng dụng
  - Phân cụm khách hàng, phân cụm sinh viên
  - Phân đoạn ảnh
  - Thiết kế vi mạch

Học tăng cường

- Kinh nghiệm không được cho trực tiếp dưới dạng đầu vào / đầu ra
- Hệ thống nhận được một giá trị thưởng (reward) là kết quả cho một chuỗi hành động nào đó
- Thuật toán cần học cách hành động để cực đại hóa giá trị thưởng
- Ví dụ: học đánh cờ
  - Hệ thống không được chỉ cho nước đi nào là hợp lý cho từng tình huống cụ thể
  - Chỉ biết kết quả thắng thua sau một chuỗi nước đi

Nội dung

- Giới thiệu
- Học cây quyết định (decision tree learning)
- Phân loại Bayes đơn giản
- Học dựa trên ví dụ

Dữ liệu huấn luyện

Ngày	Trời	Nhiệt độ	Độ ẩm	Gió	Chơi tennis
D1	nắng	nóng	cao	yếu	không
D2	nắng	nóng	cao	mạnh	không
D3	u ám	nóng	cao	yếu	có
D4	mưa	trung bình	cao	yếu	có
D5	mưa	lạnh	bình thường	yếu	có
D6	mưa	lạnh	bình thường	mạnh	không
D7	u ám	lạnh	bình thường	mạnh	có
D8	nắng	trung bình	cao	yếu	không
D9	nắng	lạnh	bình thường	yếu	có
D10	mưa	trung bình	bình thường	yếu	có
D11	nắng	trung bình	bình thường	mạnh	có
D12	u ám	trung bình	cao	mạnh	có
D13	u ám	nóng	bình thường	yếu	có
D14	mưa	trung bình	cao	mạnh	không

thuộc tính

nhãn

mẫu

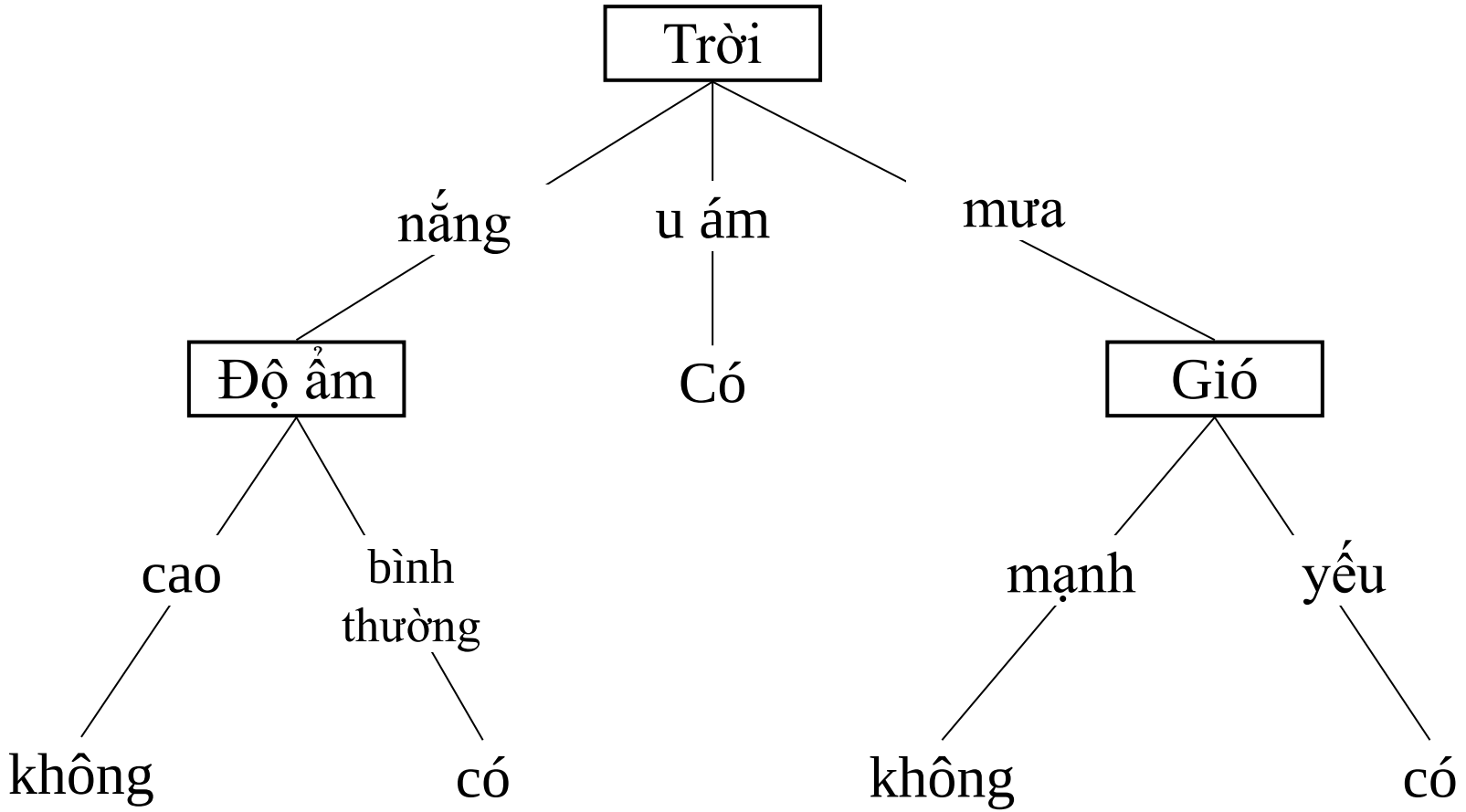
Ngày	Trời	Nhiệt độ	Độ ẩm	Gió	Chơi tennis
D1	nắng	nóng	cao	yếu	không
D2	nắng	nóng	cao	mạnh	không
D3	u ám	nóng	cao	yếu	có
D4	mưa	trung bình	cao	yếu	có
D5	mưa	lạnh	bình thường	yếu	có
D6	mưa	lạnh	bình thường	mạnh	không
D7	u ám	lạnh	bình thường	mạnh	có
D8	nắng	trung bình	cao	yếu	không
D9	nắng	lạnh	bình thường	yếu	có
D10	mưa	trung bình	bình thường	yếu	có
D11	nắng	trung bình	bình thường	mạnh	có
D12	u ám	trung bình	cao	mạnh	có
D13	u ám	nóng	bình thường	yếu	có
D14	mưa	trung bình	cao	mạnh	không

Dữ liệu

- n mẫu huấn luyện, mỗi mẫu là một cặp $< \boldsymbol{x}, y >$
  - $\boldsymbol{x}$ là vector các thuộc tính
  - y là nhãn phân loại, $y \in C$ (tập các nhãn)

- Ví dụ mẫu D4
  - $\boldsymbol{x} = (mưa,\ trung\ bình, cao, yếu)$
  - $y = có$

Ví dụ cây quyết định



Cây quyết định là gì?

- Là mô hình phân loại có dạng cây
  - Mỗi nút trung gian (không phải lá) ứng với một phép kiểm tra thuộc tính, mỗi nhánh của nút ứng với một giá trị của thuộc tính tại nút đó
  - Mỗi nút lá ứng với một nhãn phân loại
- Quá trình phân loại thực hiện như sau
  - Mẫu phân loại đi từ gốc cây xuống dưới
  - Tại mỗi nút trung gian, thuộc tính tương ứng với nút được kiểm tra, tùy giá trị thuộc tính, mẫu được chuyển xuống nhánh tương ứng
  - Khi tới nút lá, mẫu được nhận nhãn phân loại của nút

Biểu diễn dưới dạng quy tắc

- Cây quyết định có thể biểu diễn tương đương dưới dạng các quy tắc logic
- Mỗi cây là tuyển của các quy tắc, mỗi quy tắc bao gồm các phép hội
- Ví dụ

(Trời = nắng ∧ Độ ẩm = bình_thường)

∨ (Trời = u_ám)

∨ (Trời = mưa ∧ Gió = yếu)

Học cây quyết định

- Cây quyết định được học (xây dựng) từ dữ liệu huấn luyện
- Với mỗi bộ dữ liệu có thể xây dựng nhiều cây quyết định
  - Chọn cây nào?
- Quá trình học là quá trình tìm kiếm cây quyết định phù hợp với dữ liệu huấn luyện
  - Cho phép phân loại đúng dữ liệu huấn luyện

Thuật toán ID3

- Xây dựng lần lượt các nút của cây bắt đầu từ gốc

- Thuật toán
  - **Khởi đầu**: nút hiện thời là nút gốc chứa toàn bộ tập dữ liệu huấn luyện
  - Tại nút hiện thời n, lựa chọn thuộc tính
    - Chưa được sử dụng ở nút tổ tiên
    - Cho phép phân chia tập dữ liệu hiện thời thành các tập con **một cách tốt nhất**
    - Với mỗi giá trị thuộc tính được chọn thêm một nút con bên dưới
    - Chia các ví dụ ở nút hiện thời về các nút con theo giá trị thuộc tính được chọn
  - **Lặp** (đệ quy) cho tới khi
    - Tất cả các thuộc tính đã được sử dụng ở các nút phía trên, hoặc
    - Tất cả ví dụ tại nút hiện thời có cùng nhãn phân loại
    - Nhãn của nút được lấy theo đa số nhãn của ví dụ tại nút hiện thời

Lựa chọn thuộc tính tại mỗi nút thế nào?

Tiêu chuẩn chọn thuộc tính của ID3

- Tại mỗi nút n
  - Tập (con) dữ liệu ứng với nút đó
  - Cần lựa chọn thuộc tính cho phép phân chia tập dữ liệu tốt nhất
- Tiêu chuẩn:
  - Dữ liệu sau khi phân chia càng đồng nhất càng tốt
  - Đo bằng độ tăng thông tin (Information Gain - IG)
  - **Chọn thuộc tính có độ tăng thông tin lớn nhất**
  - IG dựa trên entropy của tập (con) dữ liệu

Entropy

- Trường hợp tập dữ liệu S có 2 loại nhãn: đúng (+) hoặc sai (-)

  $\boldsymbol{Entropy(S) = -p_+ log_2 p_+ - p_- log_2 p_-}$

  p_+: % số mẫu đúng, p_-: % số mẫu sai

- Trường hợp tổng quát: có C loại nhãn

  $\boldsymbol{Entropy(S) = \sum_{i=1}^{C} -p_i log_2 p_i}$

  p_i : % ví dụ của S thuộc loại i

- Ví dụ

  $\boldsymbol{Entropy([9^+, 5^-]) = -(9/14) log_2(9/14) - (5/14) log_2(5/14) = 0.94}$

Độ tăng thông tin IG

Với tập (con) mẫu S và thuộc tính A

$$IG(S,A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Trong đó:

values (*A*): tập các giá trị của A

S_v là tập con của S bao gồm các mẫu có giá trị của A bằng v

$|S|$ số phần tử của S

Ví dụ tính IG

- Tính $IG(S, Gió)$

$values(Gió) = \{yếu, mạnh\}$

$S = [9+, 5-], H(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$

$S_{yếu} = [6+, 2-], \mathrm{H}(S_{yếu}) = -\frac{6}{8} log_2 \frac{6}{8} - \frac{2}{8} log_2 \frac{2}{8} = 0.811$

$S_{mạnh} = [3+, 3-], \mathrm{H}(S_{mạnh}) = -\frac{3}{6} log_2 \frac{3}{6} - \frac{3}{6} log_2 \frac{3}{6} = 1$

$$IG(S, Gió) = H(S) - \frac{8}{14}\mathrm{H}(S_{yếu}) - \frac{6}{14}\mathrm{H}(S_{mạnh})$$
$$= 0.94 - \frac{8}{14} 0.811 - \frac{6}{14} 1$$
$$= 0.048$$

Các đặc điểm của ID3

- ID3 là thuật toán tìm kiếm cây quyết định phù hợp với dữ liệu huấn luyện
- Tìm kiếm theo kiểu tham lam, bắt đầu từ cây rỗng
- Hàm đánh giá là độ tăng thông tin
- ID3 có khuynh hướng (bias) lựa chọn cây đơn giản
  - Ít nút
  - Các thuộc tính có độ tăng thông tin lớn nằm gần gốc

Training error và Test error (1/2)

- Training error (lỗi huấn luyện)
  - Là lỗi đo được trên tập dữ liệu huấn luyện
  - Thường đo bằng sự sai khác giữa giá trị tính toán của mô hình và giá trị thực của dữ liệu huấn luyện
  - Trong quá trình học ta cố gắng làm giảm tới mức tối thiểu lỗi huấn luyện
- Test error (lỗi kiểm tra)
  - Là lỗi đo được trên tập dữ liệu kiểm tra
  - Là cái ta thực sự quan tâm

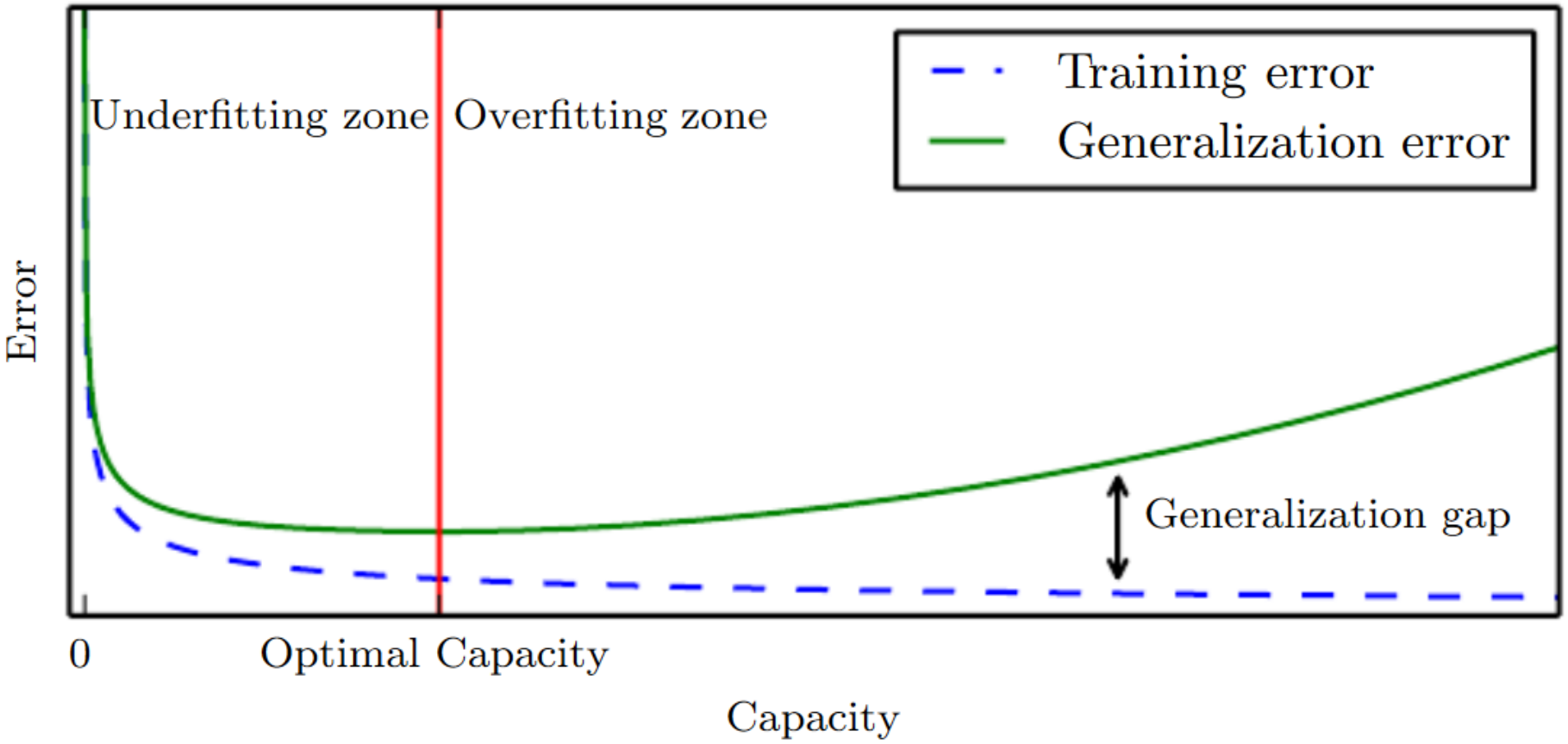
Làm sao ta có thể tác động tới hiệu quả của mô hình trên tập dữ liệu kiểm tra khi ta chỉ quan sát được tập dữ liệu huấn luyện?

Training error và Test error (2/2)

- i.i.d assumptions (independent, identically distributed)
  - Giả thiết rằng các mẫu dữ liệu (cả ở tập huấn luyện và tập kiểm tra) là độc lập, và các tập dữ liệu huấn luyện và kiểm tra có cùng phân phối
  - Nếu ta cố định các tham số của mô hình thì lỗi huấn luyện và lỗi kiểm tra sẽ bằng nhau
    - Trong quá trình huấn luyện tham số được tối ưu theo lỗi huấn luyện, do đó lỗi kiểm tra thường lớn hơn lỗi huấn luyện

- Hai yếu tố đánh giá độ tốt của một thuật toán học máy
  - Khả năng giảm thiểu lỗi huấn luyện
  - Khả năng giảm thiểu khoảng cách giữa lỗi huấn luyện và lỗi kiểm tra

Underfitting và Overfitting



Underfitting: dưới vừa; Overfitting: quá vừa

Generalization error = test error

Capacity: Khả năng của mô hình

Chống quá vừa bằng cách tỉa cây

- Chia dữ liệu thành hai phần
  - Huấn luyện
  - Kiểm tra
- Tạo cây đủ lớn trên dữ liệu huấn luyện
- Tính độ chính xác của cây trên tập kiểm tra
- Loại bỏ cây con sao cho kết quả trên dữ liệu kiểm tra được cải thiện nhất
- Lặp lại cho đến khi không còn cải thiện được kết quả nữa

Chống quá vừa dữ liệu bằng cách tỉa luật (C4.5)

- Biến đổi cây thành các luật
- Tỉa mỗi luật độc lập với các luật khác
  - Bỏ một số phần trong vế trái của luật
- Sắp xếp các luật sau khi tỉa theo mức độ chính xác của luật

Sử dụng thuộc tính có giá trị liên tục

- Tạo ra những thuộc tính rời rạc mới
- Ví dụ, với thuộc tính liên lục A, tạo ra thuộc tính rời rạc Ac như sau
  - $Ac = true$ nếu $A > c$
  - $Ac = false$ nếu $A \leq c$
- Xác định ngưỡng c thế nào?
  - Thường chọn sao cho Ac đem lại độ tăng thông tin lớn nhất
- Có thể chia thành nhiều khoảng với nhiều ngưỡng

Các độ đo khác

- Độ đo Information Gain (IG) ưu tiên thuộc tính có nhiều giá trị, ví dụ, thuộc tính ngày sẽ có độ tăng thông tin cao nhất

- Thông tin chia

$$SplitInformation(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

- Tiêu chuẩn đánh giá thuộc tính

$$GainRatio = \frac{InformationGain(S,A)}{SplitInformation(S,A)}$$

Nội dung

- Giới thiệu
- Học cây quyết định
- Phân loại Bayes đơn giản (Naïve Bayes classification)
- Học dựa trên ví dụ

Phương pháp phân loại Bayes (1/2)

- Trong giai đoạn huấn luyện ta có một tập mẫu, mỗi mẫu được cho bởi cặp $< \boldsymbol{x}_i, y_i >$, trong đó
  - $\boldsymbol{x}_i$ là vector đặc trưng (thuộc tính)
  - y_i là nhãn phân loại,$y_i \in C$ (C là tập các nhãn)
- Sau khi huấn luyện xong, bộ phân loại cần dự đoán nhãn y cho mẫu mới $\boldsymbol{x} =< x_1, x_2, \dots, x_n >$

$$y = argmax_{c_j \in C} P(c_j | x_1, x_2, \dots, x_n)$$

- Sử dụng quy tắc Bayes

$$y = argmax_{c_j \in C} \frac{P(x_1, x_2, \dots, x_n | c_j) P(c_j)}{P(x_1, x_2, \dots, x_n)}$$

$$= argmax_{c_j \in C} P(x_1, x_2, \dots, x_n | c_j) P(c_j)$$

Phương pháp phân loại Bayes (2/2)

Tần xuất quan sát thấy nhãn c_j trên tập dữ liệu D: $\frac{count(c_j)}{|D|}$

$$y = argmax_{c_j \in C} P(x_1, x_2, \dots, x_n | c_j) P(c_j)$$

Sử dụng giả thiết về tính độc lập (**Đơn giản!!!**)

$$P(x_1, x_2, \dots, x_n | c_j) = P(x_1 | c_j) P(x_2 | c_j) \dots P(x_n | c_j)$$

Số lần xuất hiện x_i cùng với c_j chia cho số lần xuất hiện c_j: $\frac{count(x_i, c_j)}{count(c_j)}$

Ví dụ

- Xác định nhãn phân loại cho mẫu sau

$$< Trời = nắng, Nhiệt\ độ = trung\ bình, Độ\ ẩm = cao, Gió = mạnh >$$

$$y = argmax_{c \in \{có, không\}} P(Trời = nắng|c) P(Nhiệt\ độ = trung\ bình|c)$$
$$P(Độ\ ẩm = cao|c) P(Gió = mạnh|c) P(c)$$

Nội dung

- Giới thiệu
- Học cây quyết định
- Phân loại Bayes đơn giản
- Học dựa trên ví dụ (Instance based learning)

Nguyên tắc chung

- Không xây dựng mô hình
- Chỉ lưu lại các mẫu huấn luyện
- Xác định nhãn cho mẫu mới dựa trên những mẫu giống mẫu mới nhất
- Gọi là học lười (lazy learning)

Thuật toán k hàng xóm gần nhất

- k-nearest neighbors (k-NN)
- Chọn k mẫu giống mẫu cần phân loại nhất, gọi là k hàng xóm
- Gán nhãn phân loại cho mẫu chỉ sử dụng thông tin của k hàng xóm này
  - Ví dụ lấy theo đa số trong số k hàng xóm
- Chọn hàng xóm thế nào?

Tính khoảng cách

- Giả sử mẫu x có giá trị thuộc tính là $< a_1(x), a_2(x), \ldots, an(x) >$, thuộc tính là số thực
- Khoảng cách giữa hai mẫu x_i và x_j là khoảng cách Euclidean

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{n} (a_l(x_i) - a_l(x_j))^2}$$

Thuật toán k-NN

Giai đoạn học (huấn luyện)

Lưu các mẫu huấn luyện có dạng $<x, f(x)>$ vào cơ sở dữ liệu

Giai đoạn phân loại

Đầu vào: tham số k

Với mẫu x cần phân loại:

1. Tính khoảng cách $d(x, x_i)$ từ x tới tất cả mẫu x_i trong cơ sở dữ liệu
2. Tìm k mẫu có $d(x, x_i)$ nhỏ nhất, giả sử k mẫu đó là $x_1, x_2, \ldots, x_k$.
3. Xác định nhãn phân loại $f'(x)$ là nhãn chiếm đa số trong tập $\{ x_1, x_2, \ldots, x_k \}$