

# HỌC MÁY ỨNG DỤNG

(Applied Machine Learning)

## CHƯƠNG 1 : GIỚI THIỆU HỌC MÁY



GVGD: PHAN HỒ VIẾT TRƯỜNG

# NỘI DUNG

- 01. GIỚI THIỆU HỌC MÁY**
- 02. HỌC KHÁI NIỆM CƠ BẢN**
- 03. GIẢI THUẬT FIND-S**
- 04. GIẢI THUẬT CANDIDATE ELIMINATION**
- 05. BÀI TẬP**

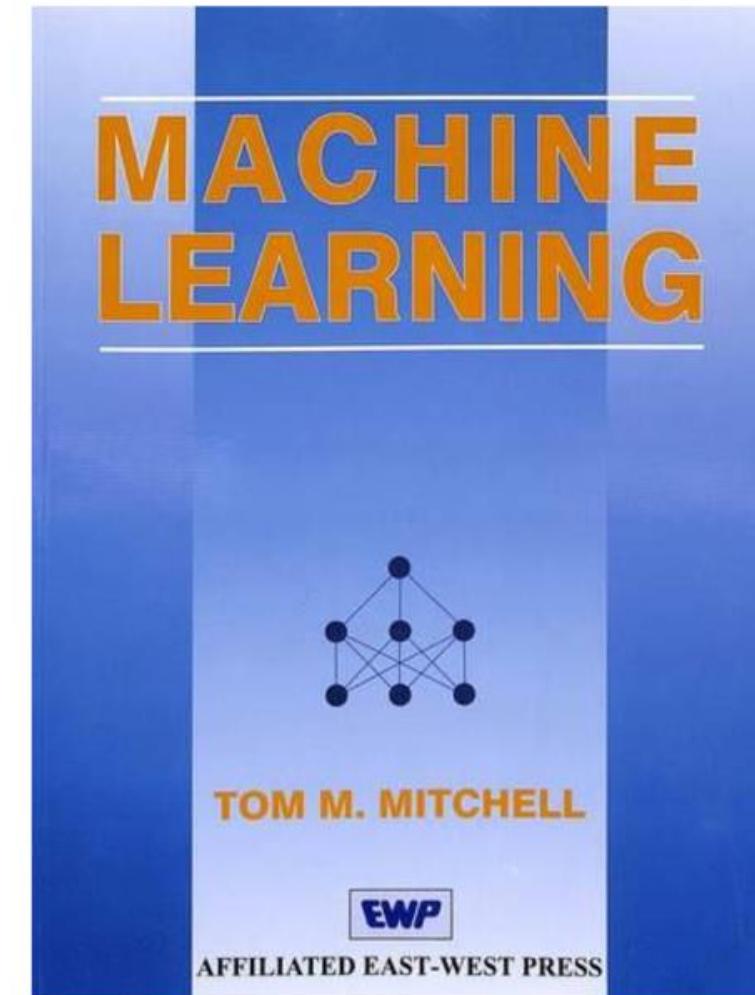


# GIỚI THIỆU HỌC MÁY

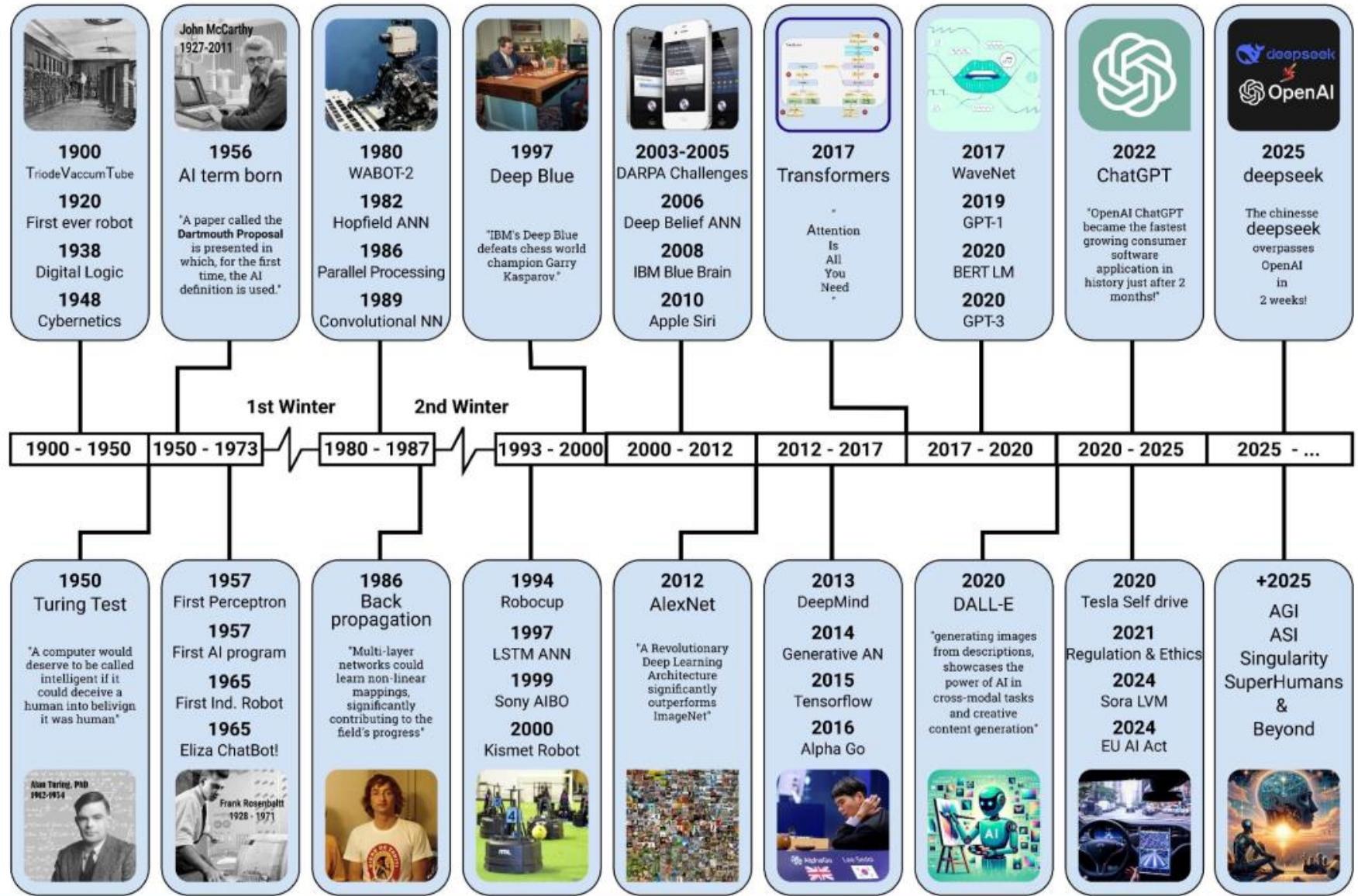
## What is Machine Learning?

"A computer program is said to learn from experience E, with respect to some task T, and some performance measure P, if its performance on T as measured by P improves with experience E"

- Tom Mitchell

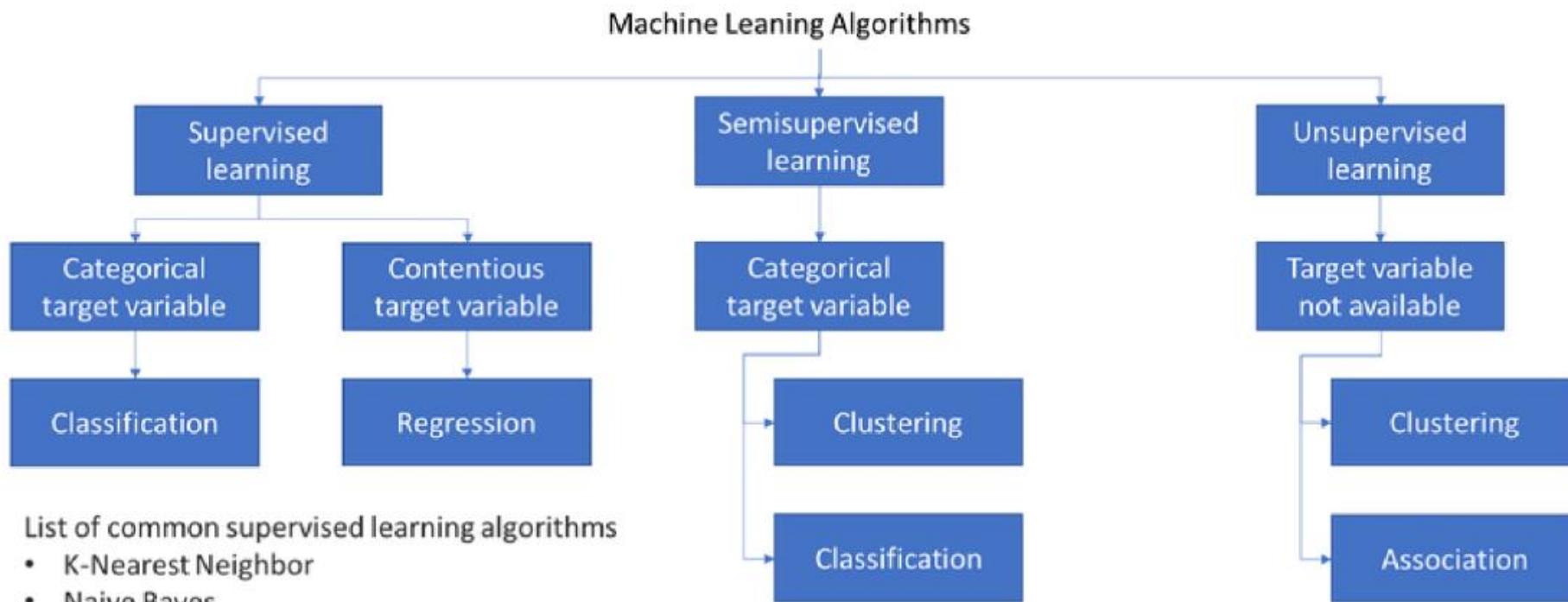


# GIỚI THIỆU HỌC MÁY





# GIỚI THIỆU HỌC MÁY



List of common supervised learning algorithms

- K-Nearest Neighbor
- Naive Bayes
- Decision Trees
- Linear Regression
- Support Vector Machines (SVM)
- Neural Networks
- Classification and Regression Trees
- Gradient Boosted Regression Tree
- Perceptron Back-Propagation
- Random Forest

List of common semi supervised learning algorithms:

- Linear Regression
- Logistic Regression

List of common unsupervised learning algorithms:

- k-means clustering and classification
- Association Rules



# XÁC ĐỊNH VÂN ĐỀ HỌC TRÊN MÁY

- Có 3 đặc trưng task ( $T$ ) , performance ( $P$ ), experience source ( $E$ )

Ví dụ 1: Vấn đề máy học chơi cờ

**T:** Chơi cờ

**P:** Tỉ lệ máy thắng

**E:** Chơi thực tế với chính máy và con người





# XÁC ĐỊNH VÂN ĐỀ HỌC TRÊN MÁY

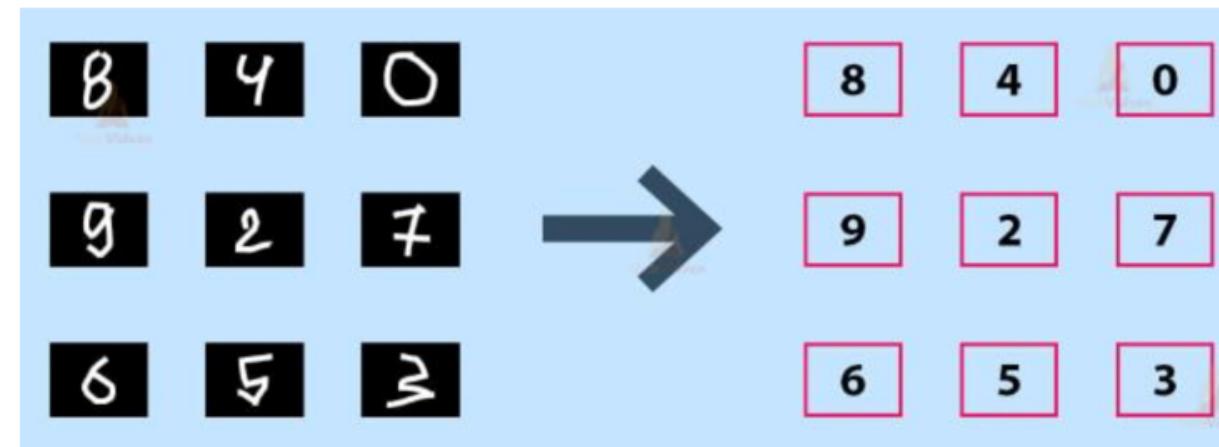
- Có 3 đặc trưng task (*T*) , performance (*P*), experience source (*E*)

Ví dụ 2: Vân đề máy học nhận dạng chữ viết tay

**T**: nhận dạng và phân lớp từ viết tay từ hình ảnh

**P**: Tỉ lệ máy nhận dạng đúng

**E**: Tập dữ liệu từ viết tay cùng vớp lớp cho trước





# HỌC MÁY NHƯ THẾ NÀO ?

- Chia làm 3 giai đoạn
- Data Input: Dữ liệu quá khứ được sử dụng dự báo tương lai
- Abstraction: Dữ liệu đầu vào được biểu diễn theo cách rộng hơn thông qua giải thuật.
- Generalization: Dữ liệu abstraction được tổng quát thành một framework ra quyết định.



# HỌC KHÁI NIỆM TỪ TỔNG QUÁT ĐẾN CỤ THỂ

- Học máy yêu cầu khái niệm tổng quát từ tập mẫu học cụ thể.
- Xét vấn đề tự động suy diễn khái niệm tổng quát từ tập mẫu có gán nhãn
- Cho tập mẫu

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- → Làm sao dự báo giá trị EnjoySport cho ngày thứ 5 ?



# HỌC KHÁI NIỆM TỪ TỔNG QUÁT ĐẾN CỤ THỂ

- Cho tập mẫu

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- Gọi giả thuyết tổng quát mỗi ngày là 1 mẫu positive (?, ?, ?, ?, ?, ?)
- Gọi giả thuyết cụ thể không có ngày nào là positive (0, 0, 0, 0, 0, 0)
- ? Là giá trị chấp nhận của thuộc tính, 0 là giá trị không được chấp nhận



# HỌC KHÁI NIỆM TỪ TỔNG QUÁT ĐẾN CỤ THỂ

- Cho tập mẫu

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- Học khái niệm EnjoySport là việc học trên tập mẫu có EnjoySport = Yes.
- Nếu ngày x thỏa các giá trị của giả thuyết h thì  $h(x) = 1$  phân lớp Yes.



# HỌC KHÁI NIỆM TỪ TỔNG QUÁT ĐẾN CỤ THỂ

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- Gọi  $\{Sky, AirTemp, Humidity, Wind, Water, Forecast\}$ : thuộc tính của 1 mẫu  $x \in X$ .
- Hàm mục tiêu (khái niệm mục tiêu) ký hiệu là  $c: X \rightarrow \{0, 1\}$ .
- Giá trị  $c(x)$  tương ứng giá trị của  $EnjoySport$
- $c(Sky, AirTemp, Humidity, Wind, Water, Forecast) = 1$  nếu  $EnjoySport = Yes$  (Positive)
- $c(Sky, AirTemp, Humidity, Wind, Water, Forecast) = 0$  nếu  $EnjoySport = No$  (Negative)



# HỌC KHÁI NIỆM TỪ TỔNG QUÁT ĐẾN CỤ THỂ

---

- Given:

- Instances  $X$ : Possible days, each described by the attributes
  - *Sky* (with possible values *Sunny*, *Cloudy*, and *Rainy*),
  - *AirTemp* (with values *Warm* and *Cold*),
  - *Humidity* (with values *Normal* and *High*),
  - *Wind* (with values *Strong* and *Weak*),
  - *Water* (with values *Warm* and *Cool*), and
  - *Forecast* (with values *Same* and *Change*).
- Hypotheses  $H$ : Each hypothesis is described by a conjunction of constraints on the attributes *Sky*, *AirTemp*, *Humidity*, *Wind*, *Water*, and *Forecast*. The constraints may be “?” (any value is acceptable), “ $\emptyset$ ” (no value is acceptable), or a specific value.
- Target concept  $c$ :  $EnjoySport : X \rightarrow \{0, 1\}$
- Training examples  $D$ : Positive and negative examples of the target function (see Table 2.1).

- Determine:

- A hypothesis  $h$  in  $H$  such that  $h(x) = c(x)$  for all  $x$  in  $X$ .
-



# HỌC KHÁI NIỆM TỪ TỔNG QUÁT ĐẾN CỤ THỂ

- Học khái niệm xem như là bài toán tìm kiếm tập mẫu giả thiết ban đầu cho việc huấn luyện.
- Tuy nhiên, không gian tìm kiếm mẫu rất lớn → cần giải thuật hiệu quả hoặc giới hạn tập mẫu → Huấn luyện học khái niệm tốt nhất
- Sử dụng cấu trúc General-to-Specific (Tổng quát đến cụ thể) thiết kế giải thuật vét cạn tập không gian mẫu lớn.



# HỌC KHÁI NIỆM TỪ TỔNG QUÁT ĐẾN CỤ THỂ

- Xét ví dụ sau:

$$H1 = (\text{Sunny}, ?, ?, \text{Strong}, ?, ?)$$

$$H2 = (\text{Sunny}, ?, ?, ?, ?, ?, ?)$$

→ H2 ít ràng buộc hơn H1 → H2 tổng quát hơn H1

→ Định nghĩa quan hệ **GENERAL\_EQUAL**

“Gọi  $h_j$  và  $h_k$  là 2 giả thuyết,  $h_j$  được gọi là **GENERAL\_EQUAL**  $h_k$  nếu và chỉ nếu có bất kỳ mẫu nào thỏa điều kiện  $h_k$  cũng sẽ thỏa  $h_j$ ”

$$h_j \geq_g h_k : (\forall x \in X)[(h_k(x) = 1) \rightarrow (h_j(x) = 1)]$$



# GIẢI THUẬT FIND-S TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

---

1. Initialize  $h$  to the most specific hypothesis in  $H$
  2. For each positive training instance  $x$ 
    - For each attribute constraint  $a_i$  in  $h$ 
      - If the constraint  $a_i$  is satisfied by  $x$ 
        - Then do nothing
      - Else replace  $a_i$  in  $h$  by the next more general constraint that is satisfied by  $x$
  3. Output hypothesis  $h$
-



# GIẢI THUẬT FIND-S TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- B1: khởi động giả thuyết và chỉ xét mẫu positive (EnjoySport = Yes)

$$h \leftarrow (0, 0, 0, 0, 0, 0) \quad (\text{Specific})$$



# GIẢI THUẬT FIND-S TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- B2: với mỗi giá trị thuộc tính trong mẫu  $x_1$  không thỏa h thì thay bằng giá trị tổng quát hơn thỏa  $x_1$   
 $h \leftarrow (\text{Sunny}, \text{Warm}, \text{Normal}, \text{Strong}, \text{Warm}, \text{Same})$  (Còn specific)



# GIẢI THUẬT FIND-S TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- B2: với mỗi giá trị thuộc tính trong mẫu  $x_2$  không thỏa h thì thay bằng giá trị tổng quát hơn thỏa  $x_2$

$h \leftarrow (\text{Sunny}, \text{Warm}, ?, \text{Strong}, \text{Warm}, \text{Same})$



# GIẢI THUẬT FIND-S TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- B2:  $x_3$  là negative nên qua mẫu tiếp theo  
 $h \leftarrow (\text{Sunny}, \text{Warm}, ?, \text{Strong}, \text{Warm}, \text{Same})$



# GIẢI THUẬT FIND-S TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- B2: Với mỗi giá trị thuộc tính trong  $x_4$  (positive) không thỏa h thì thay bằng giá trị tổng quát hơn  
 $h \leftarrow (\text{Sunny}, \text{Warm}, ?, \text{Strong}, ?, ?)$  (DÙNG GIẢI THUẬT)



# BÀI TẬP GIẢI THUẬT FIND-S

P ( Temp =Cool Or Wind = Weak | Outlook= Rain)

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No



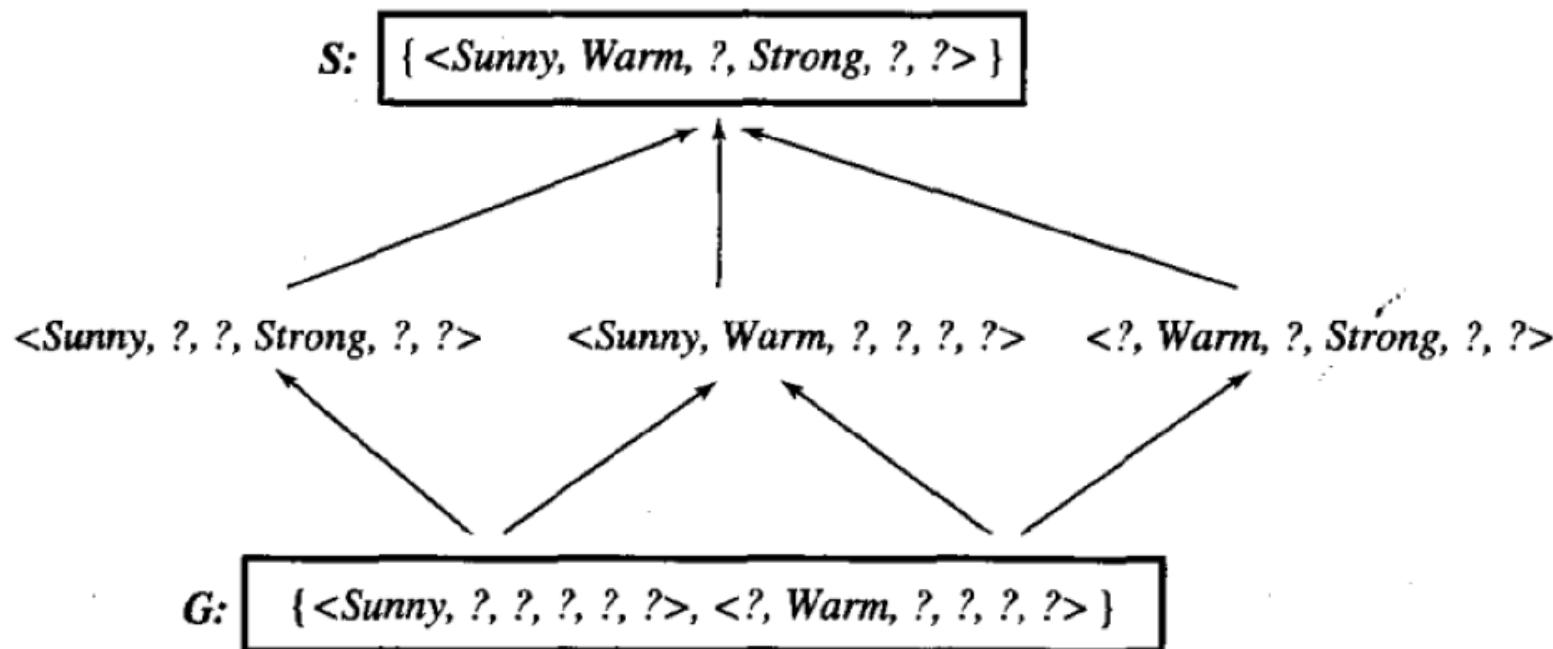


# BÀI TẬP GIẢI THUẬT FIND-S

rec	Age	Income	Student	Credit_rating	Buys_computer
r1	<=30	Hight	No	Fair	No
r2	<=30	Hight	No	Excellent	No
r3	31..40	Hight	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31..40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	=30	Low	Yes	Fair	Yes
r10	>30	Medium	Yes	Fair	Yes
r11	<=30	Medium	Yes	Excellent	Yes
r12	31..40	Medium	No	Excellent	Yes
r13	31..40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No

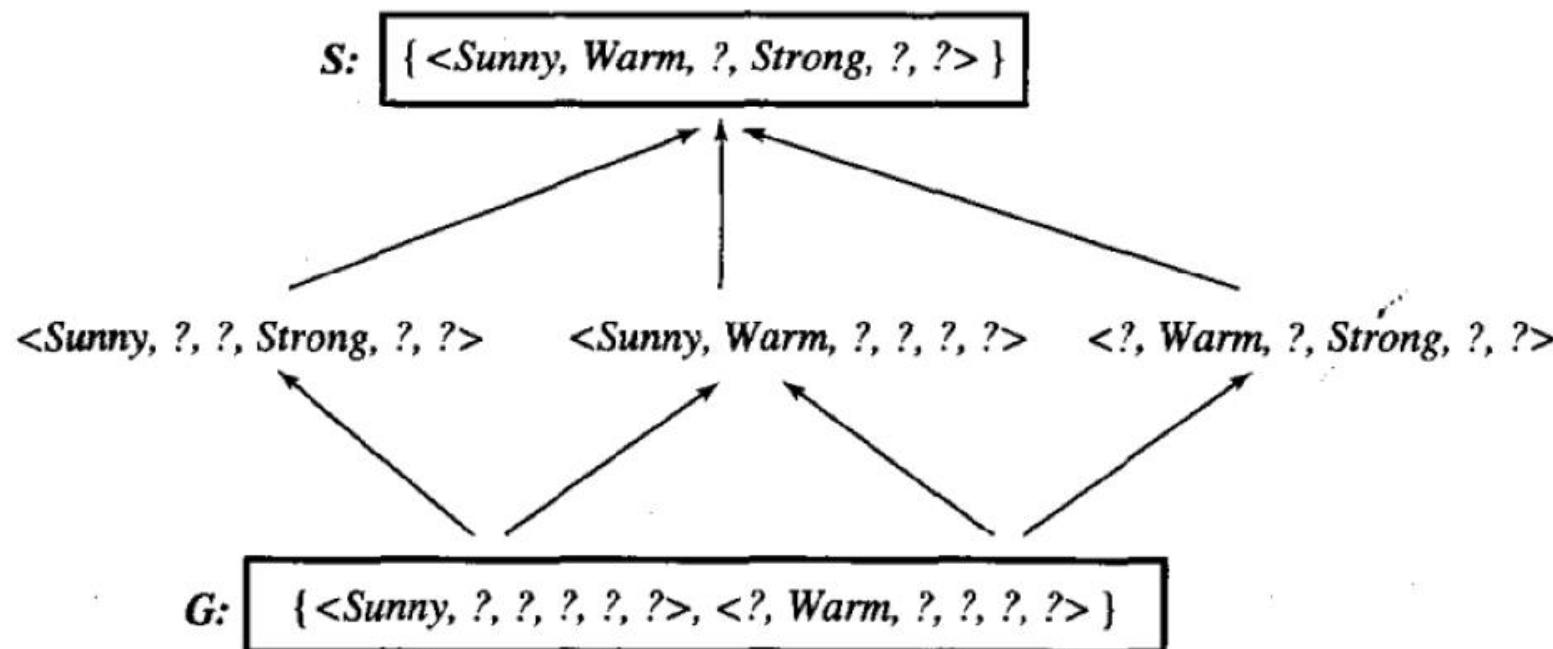
# GIẢI THUẬT CANDIDATE-ELIMINATION TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

- FIND-S chỉ tìm được 1 mẫu chung giả thuyết  $h$  trên tập huấn luyện
- CANDIDATE-ELIMINATION tìm một tập mẫu chung giả thuyết  $h$  trên tập huấn luyện.



# GIẢI THUẬT CANDIDATE-ELIMINATION TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

- Biên tổng quát (General boundary): là tập các giả thuyết tổng quát nhất H trên tập mẫu D
- Biên cụ thể (Specific boundary): là tập giả thuyết cụ thể nhất H trên tập D





# GIẢI THUẬT CANDIDATE-ELIMINATION TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

---

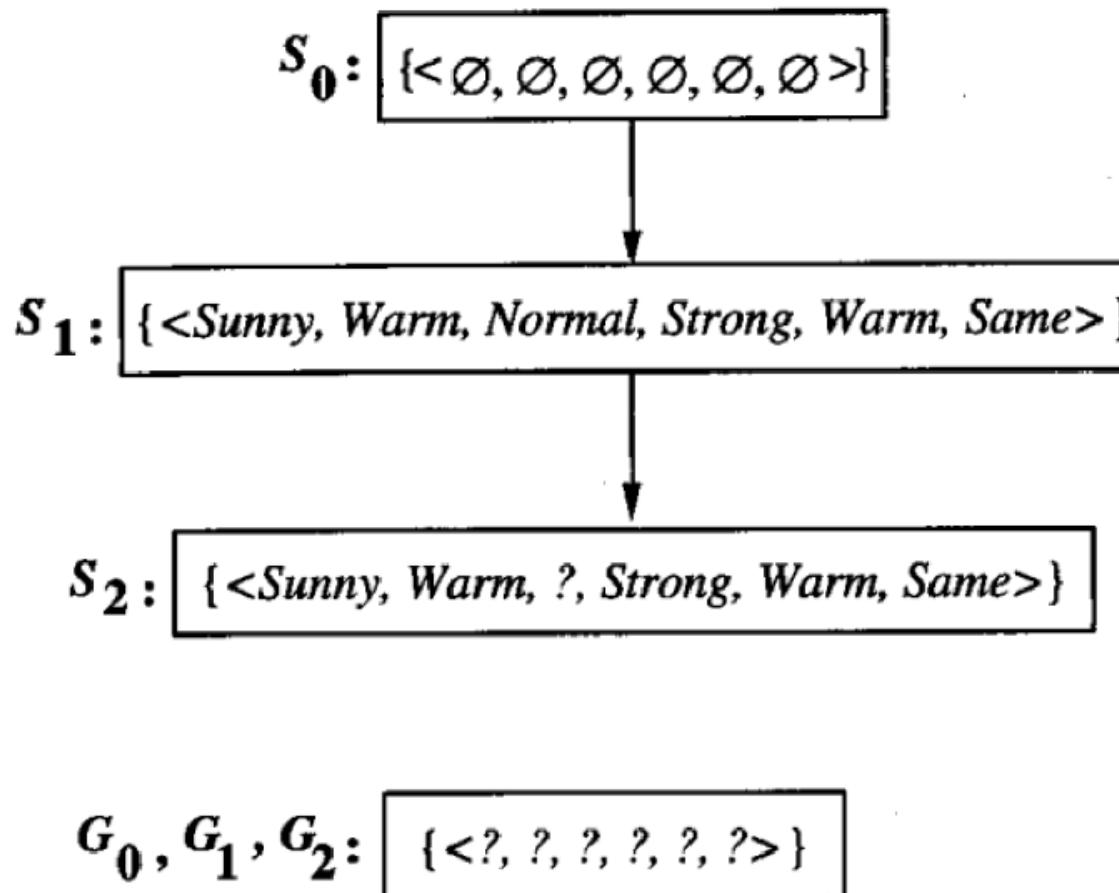
Initialize  $G$  to the set of maximally general hypotheses in  $H$

Initialize  $S$  to the set of maximally specific hypotheses in  $H$

For each training example  $d$ , do

- If  $d$  is a positive example
  - Remove from  $G$  any hypothesis inconsistent with  $d$
  - For each hypothesis  $s$  in  $S$  that is not consistent with  $d$ 
    - Remove  $s$  from  $S$
    - Add to  $S$  all minimal generalizations  $h$  of  $s$  such that
      - $h$  is consistent with  $d$ , and some member of  $G$  is more general than  $h$
    - Remove from  $S$  any hypothesis that is more general than another hypothesis in  $S$
- If  $d$  is a negative example
  - Remove from  $S$  any hypothesis inconsistent with  $d$
  - For each hypothesis  $g$  in  $G$  that is not consistent with  $d$ 
    - Remove  $g$  from  $G$
    - Add to  $G$  all minimal specializations  $h$  of  $g$  such that
      - $h$  is consistent with  $d$ , and some member of  $S$  is more specific than  $h$
    - Remove from  $G$  any hypothesis that is less general than another hypothesis in  $G$

# GIẢI THUẬT CANDIDATE-ELIMINATION TÌM TẬP GIẢ THUYẾT TỔNG QUÁT



Training examples:

1. *<Sunny, Warm, Normal, Strong, Warm, Same>, Enjoy Sport = Yes*
2. *<Sunny, Warm, High, Strong, Warm, Same>, Enjoy Sport = Yes*

Tương tự FIND-S: Tìm giả thuyết tổng quát S2 dự báo các mẫu positive

# GIẢI THUẬT CANDIDATE-ELIMINATION TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

$S_2, S_3:$  { <Sunny, Warm, ?, Strong, Warm, Same> }

$G_3:$  { <Sunny, ?, ?, ?, ?, ?>   <?, Warm, ?, ?, ?, ?>   <?, ?, ?, ?, ?, Same> }

$G_2:$  { <?, ?, ?, ?, ?, ?> }

Tìm và loại giả thuyết tổng quát  $S$  dự báo các mẫu negative

<Sunny, ?, ?, ?, ?, ?, ?>  
<?, Warm, ?, ?, ?, ?, ?>  
<?, ?, ?, Strong, ?, ?, ?> (Negative → Loại)  
<?, ?, ?, ?, Warm, ?, ?> (Negative → Loại)  
<?, ?, ?, ?, ?, Same>

Training Example:

3. <Rainy, Cold, High, Strong, Warm, Change>, EnjoySport=No

# GIẢI THUẬT CANDIDATE-ELIMINATION TÌM TẬP GIẢ THUYẾT TỔNG QUÁT

S<sub>3</sub>: {<Sunny, Warm, ?, Strong, Warm, Same>}



S<sub>4</sub>: {<Sunny, Warm, ?, Strong, ?, ?>}

- FIND-S tìm S<sub>4</sub> <Sunny, Warm, ?, Strong, ?, ?>
- Xóa trong G<sub>3</sub> những h làm cho dự báo negative <?, ?, ?, ?, ?, Same>

G<sub>4</sub>: {<Sunny, ?, ?, ?, ?, ?> <?, Warm, ?, ?, ?, ?>}

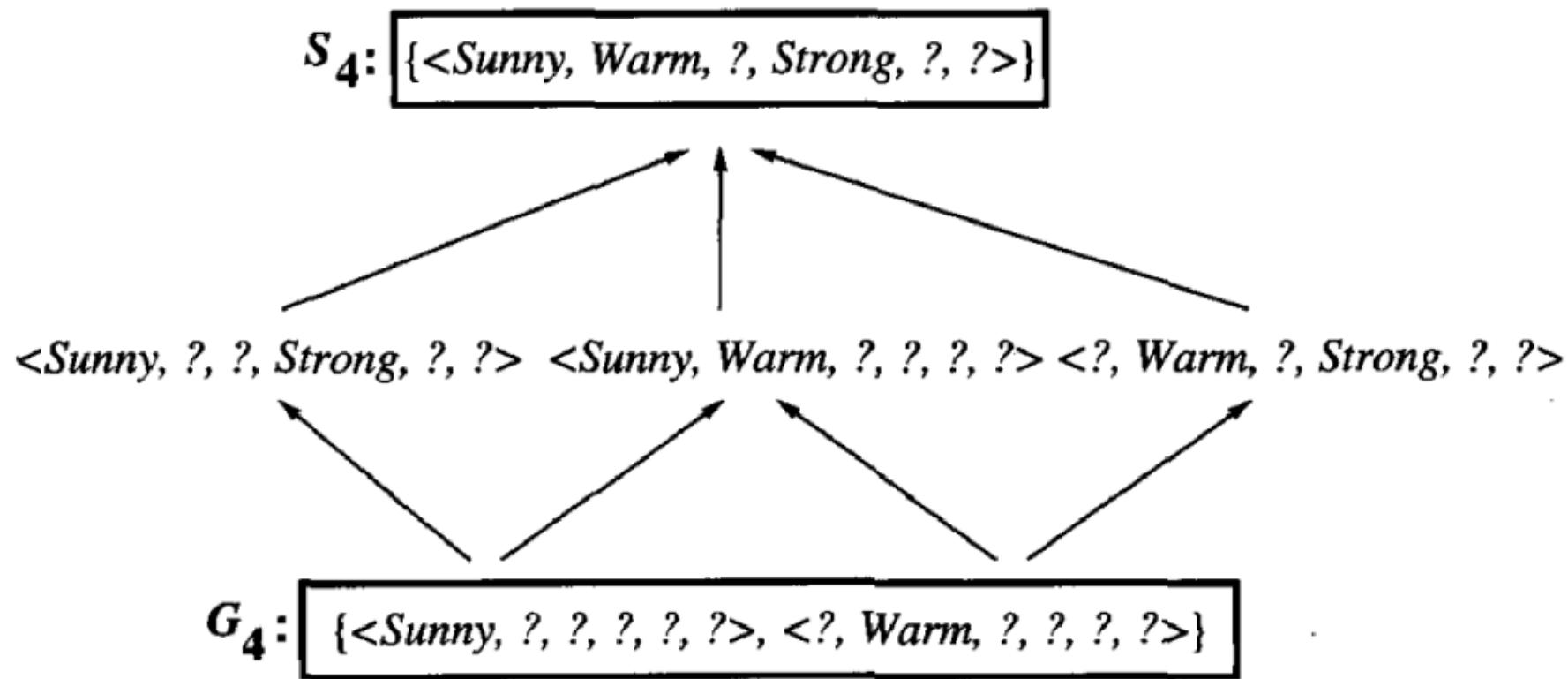


G<sub>3</sub>: {<Sunny, ?, ?, ?, ?, ?> <?, Warm, ?, ?, ?, ?> <?, ?, ?, ?, ?, Same>}

Training Example:

4. <Sunny, Warm, High, Strong, Cool, Change>, EnjoySport = Yes

# GIẢI THUẬT CANDIDATE-ELIMINATION TÌM TẬP GIẢ THUYẾT TỔNG QUÁT





# BÀI TẬP

## GIẢI THUẬT CANDIDATE-ELIMINATION

P ( Temp =Cool Or Wind = Weak | Outlook= Rain)

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No



# BÀI TẬP

## GIẢI THUẬT CANDIDATE-ELIMINATION

rec	Age	Income	Student	Credit_rating	Buys_computer
r1	<=30	Hight	No	Fair	No
r2	<=30	Hight	No	Excellent	No
r3	31..40	Hight	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31..40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	=30	Low	Yes	Fair	Yes
r10	>30	Medium	Yes	Fair	Yes
r11	<=30	Medium	Yes	Excellent	Yes
r12	31..40	Medium	No	Excellent	Yes
r13	31..40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No



# BÀI TẬP

- Thế nào là concept trong học máy? Bạn có thể đưa ra một ví dụ đơn giản về một concept trong đời sống hằng ngày không?
- Tại sao người ta thường nói concept learning là quá trình “tìm một giả thuyết phù hợp nhất” với tập dữ liệu huấn luyện?
- Ưu điểm và hạn chế của việc coi concept learning như một bài toán tìm kiếm trong không gian giả thuyết là gì?
- FIND-S giả định gì về dữ liệu huấn luyện (ví dụ: tất cả ví dụ huấn luyện đều được gán nhãn chính xác)? Hệ quả nếu giả định này sai thì sao?
- Trong thực tế, khi nào thì FIND-S có thể cho ra một giả thuyết “quá hẹp”? Bạn sẽ khắc phục bằng cách nào?
- So sánh FIND-S với cách tiếp cận “naive memorization” (chỉ ghi nhớ tất cả ví dụ huấn luyện). Điểm mạnh – yếu của mỗi cách?
- Candidate-Elimination sử dụng tập S (giả thuyết đặc hiệu nhất) và G (giả thuyết tổng quát nhất). Ý nghĩa trực giác của hai tập này là gì?
- Khi thêm một ví dụ huấn luyện mới, tại sao cả S và G đều có thể bị điều chỉnh? Cho ví dụ minh họa.
- Candidate-Elimination có thể dẫn đến tình huống “phiên không kết luận được” (no consistent hypothesis). Khi nào điều này xảy ra?
- Nếu dữ liệu huấn luyện có nhiễu (noise), Candidate-Elimination sẽ gặp khó khăn như thế nào? Bạn đề xuất giải pháp gì để xử lý?



**Q&A**

