HỆ THỐNG HỌC (MÁY HỌC)

Dẫn nhập

- Các hệ thống thông minh đã học: lập luận từ tri thức đã có sẵn (do người cung cấp)
- Hôm nay: **hệ thống học** (máy học) - tự động học tri thức từ dữ liệu

Ví dụ: dự đoán bệnh cúm

- Cho các triệu chứng của một bệnh nhân:
  - Ớn lạnh: có / không
  - Sổ mũi: có / không
  - Đau đầu: không / hơi-đau / đau
  - Sốt: có / không
- Yêu cầu: dự đoán bệnh nhân có bị bệnh cúm (flu) hay không?

Máy/người-không-biết-gì-về-bệnh-cúm có thể dự đoán được không?

Không. Ta cần dạy/huấn-luyện cho máy/người-không-biết-gì-về-bệnh-cúm học cách dự đoán

Ví dụ: dự đoán bệnh cúm

Dạy/huấn-luyện cho máy học cách dự đoán bệnh cúm như thế nào?

- Dạy bằng một tập dữ liệu (gọi là **tập huấn luyện**) gồm các ví dụ mẫu có dạng (input, output đúng)
- Máy sẽ tự động học từ tập dữ liệu này cách dự đoán ouput từ input
- Học không phải là ghi nhớ các mẫu trong tập dữ liệu, mà phải rút ra được qui luật chung ở bên dưới để có thể dự đoán được với một mẫu mới mà không có trong tập dữ liệu

Input				**Output**
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
⋮	⋮	⋮	⋮	⋮

Giới thiệu sơ bộ về lĩnh vực máy học

Máy học (machine learning) - lĩnh vực nghiên cứu về việc làm cho máy có thể tự động học từ dữ liệu - là một lĩnh vực "hot" hiện nay; lý do: dữ liệu ngày càng nhiều và máy tính ngày cành mạnh

- Ở các công ty IT lớn trên thế giới (Google, Facebook, …), máy học là một mảng quan trọng
- Ở các công ty IT Việt Nam, máy học đang dần nổi lên như là một xu hướng tất yếu

Giới thiệu sơ bộ về lĩnh vực máy học

Trong máy học, có nhiều thể loại học

- **Học có giám sát (supervised learning)**
  - Tập dữ liệu gồm các mẫu có dạng (input, output đúng)
  - Trong học có giám sát, lại chia thành 2 nhóm con:
    - **Hồi qui (regression)**: output có giá trị liên tục (vd, dự đoán giá chứng khoán)
    - **Phân lớp (classification)**: output có giá trị rời rạc (vd, dự đoán xu hướng giá chứng khoán: tăng, giữ nguyên, giảm)

Giới thiệu sơ bộ về lĩnh vực máy học

Trong máy học, có nhiều thể loại học

- **Học có giám sát (supervised learning)**
  - Tập dữ liệu gồm các mẫu có dạng (input, output đúng)
  - Trong học có giám sát, lại chia thành 2 nhóm con:
    - **Hồi qui** ục (vd, dự đoán giá chứng k
    - **Phân lớp (classification)**: output có giá trị rời rạc (vd, dự đoán xu hướng giá chứng khoán: tăng, giữ nguyên, giảm)

Đây là loại học mà ta sẽ tập trung tìm hiểu trong môn học này

- Ngoài ra còn có các loại học khác như:
  - Học không giám sát (unsupervised learning)
  - Học tăng cường (reinforcement learning)
  - …

Nội dung tiếp theo

Thuật toán học Naïve Bayes: một thuật toán học-cách-phân-lớp đơn giản

Trước tiên: ký hiệu

- X: véc-tơ đầu vào nói chung
- Y: đầu ra nói chung; trong bài toán phân lớp, Y còn được gọi là lớp/nhãn
- X_i: thuộc tính thứ i của véc-tơ đầu vào
- n: số lượng thuộc tính của véc-tơ đầu vào

X ≡ Input				Y ≡ Output
X_1 ≡ Ớn lạnh	X_2 ≡ Sổ mũi	X_3 ≡ Đau đầu	X_4 ≡ Sốt	Y ≡ Cúm
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
⋮	⋮	⋮	⋮	⋮

Thuật toán học Naïve Bayes

- Với một véc-tơ đầu vào X bất kỳ, Naïve Bayes sẽ dự đoán lớp Y của X như sau:
  - Với các giá trị v có thể có của Y, tính xác suất $p(Y = v|X)$
  - Dự đoán lớp Y của X là v mà có xác suất tương ứng lớn nhất
- Vd: với $X = [\text{có (ớn lạnh)}, \text{không (sổ mũi)}, \text{hơi đau (đầu)}, \text{không (sốt)}]$, dự đoán Y (Cúm) là có/không?
  - Tính $p(Y = \text{có}|X)$ và tính $p(Y = \text{không}|X)$
  - Dự đoán lớp của X là lớp mà có xác suất tương ứng lớn hơn

Thuật toán học Naïve Bayes

- Để dự đoán, đầu tiên với mỗi lớp v, ta cần tính xác suất $p(Y = v|X)$; theo luật **Bayes** trong xác suất:

$$p(Y = v|X) = \frac{p(X|Y = v)p(Y = v)}{p(X)}$$

- Với các lớp v khác nhau, $p(Y = v|X)$ có phần mẫu $p(X)$ giống nhau → để chọn ra lớp v có $p(Y = v|X)$ lớn nhất, ta chỉ cần so sánh phần tử $p(X|Y = v)p(Y = v)$ của các lớp v với nhau
- Như vậy, để dự đoán được với một véc-tơ đầu vào X bất kỳ, trong quá trình học (huấn luyện), với mỗi lớp v, máy cần học:
  - $p(Y = v)$
  - $p(X|Y = v)$ với mọi giá trị có thể có của X

Thuật toán học Naïve Bayes

Như vậy, để dự đoán được với một véc-tơ đầu vào X bất kỳ, trong quá trình học (huấn luyện), với mỗi lớp v, máy cần học:

- $p(Y = v)$
- $p(X|Y = v)$ với mọi giá trị có thể có của X
  - Có bao nhiêu giá trị có thể có của X?
    - Giả sử X có 10 thuộc tính đầu vào, mỗi thuộc tính có 2 giá trị có thể có → X có 2^{10} giá trị có thể có ☹; có nhiều giá trị sẽ không có trong tập dữ liệu → không học được
  - Để đơn giản hóa, Naïve Bayes giả định một cách "**naïve**" rằng: *với điều kiện Y, các thuộc tính đầu vào X_i độc lập với nhau*

$$p(X|Y = v) = \prod_{i=1}^{n} p(X_i|Y = v)$$

  - Chỉ cần học $p(X_i|Y = v)$ với mọi X_i và với mọi giá trị có thể có của mỗi X_i
    - Giả sử X có 10 thuộc tính đầu vào, mỗi thuộc tính có 2 giá trị có thể có → tổng cộng sẽ có 2×10 giá trị xác suất cần học ☺

Để dự đoán được với một véc-tơ đầu vào X bất kỳ, máy cần học:

- $p(\text{Cúm} = \text{không})$
- $p\big(\text{Ớn lạnh} = \text{không} \big| \text{Cúm} = \text{không}\big)$, $p(\text{Ớn lạnh} = \text{có} | \text{Cúm} = \text{không})$
- $p(\text{Sổ mũi} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Sổ mũi} = \text{có} | \text{Cúm} = \text{không})$
- $p(\text{Đau đầu} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Đau đầu} = \text{hơi đau} | \text{Cúm} = \text{không})$, $p(\text{Đau đầu} = \text{đau} | \text{Cúm} = \text{không})$
- $p(\text{Sốt} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Sốt} = \text{có} | \text{Cúm} = \text{không})$
- Tương tự cho lớp Cúm = có

$X \equiv$ Input				$Y \equiv$ Output
$X_1 \equiv$ Ớn lạnh	$X_2 \equiv$ Sổ mũi	$X_3 \equiv$ Đau đầu	$X_4 \equiv$ Sốt	$Y \equiv$ Cúm
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
⋮	⋮	⋮	⋮	⋮

Để dự đoán được với một véc-tơ đầu vào X bất kỳ, máy cần học:

- $p(\text{Cúm} = \text{không})$
- $p\big(\text{Ớn lạnh} = \text{không} \big| \text{Cúm} = \text{không}\big)$, $p(\text{Ớn lạnh} = \text{có} | \text{Cúm} = \text{không})$
- $p(\text{Sổ mũi} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Sổ mũi} = \text{có} | \text{Cúm} = \text{không})$
- $p(\text{Đau đầu} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Đau đầu} = \text{hơi đau} | \text{Cúm} = \text{không})$, $p(\text{Đau đầu} = \text{đau} | \text{Cúm} = \text{không})$
- $p(\text{Sốt} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Sốt} = \text{có} | \text{Cúm} = \text{không})$
- $p(\text{Cúm} = \text{có})$
- $p\big(\text{Ớn lạnh} = \text{không} \big| \text{Cúm} = \text{có}\big)$, $p(\text{Ớn lạnh} = \text{có} | \text{Cúm} = \text{có})$
- $p(\text{Sổ mũi} = \text{không} | \text{Cúm} = \text{có})$, $p(\text{Sổ mũi} = \text{có} | \text{Cúm} = \text{có})$
- $p(\text{Đau đầu} = \text{không} | \text{Cúm} = \text{có})$, $p(\text{Đau đầu} = \text{hơi đau} | \text{Cúm} = \text{có})$, $p(\text{Đau đầu} = \text{đau} | \text{Cúm} = \text{có})$
- $p(\text{Sốt} = \text{không} | \text{Cúm} = \text{có})$, $p(\text{Sốt} = \text{có} | \text{Cúm} = \text{có})$

Để dự đoán được với một véc-tơ đầu vào X bất kỳ, máy cần học:

- $p(\text{Cúm} = \text{không})$
- $p(\text{Ớn lạnh} = \text{không} | \text{Cúm} = \text{không}),\ p(\text{Ớn lạnh} = \text{có} | \text{Cúm} = \text{không})$
- $p(\text{Sổ mũi} = \text{không} | \text{Cúm} = \text{không}),\ p(\text{Sổ mũi} = \text{có} | \text{Cúm} = \text{không})$
- $p(\text{Đau đầu} = \text{không} | \text{Cúm} = \text{không}),\ p(\text{Đau đầu} = \text{hơi đau} | \text{Cúm} = \text{không}),$ $p(\text{Đau đầu} = \text{đau} | \text{Cúm} = \text{không})$
- $p(\text{Sốt} = \text{không} | \text{Cúm} = \text{không}),\ p(\text{Sốt} = \text{có} | \text{Cúm} = \text{không})$
- Tương tự cho lớp $\text{Cúm} = \text{có}$

Khi đã có các giá trị xác suất này ta có thể dự đoán với một véc-tơ đầu vào X bất kỳ

Vd: thử dự đoán với $X = [\text{có (ớn lạnh)}, \text{không (sổ mũi)}, \text{hơi đau (đầu)}, \text{không (sốt)}]$

Để dự đoán được với một véc-tơ đầu vào X bất kỳ, máy cần học:

- $p(\text{Cúm} = \text{không})$
- $p(\text{Ớn lạnh} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Ớn lạnh} = \text{có} | \text{Cúm} = \text{không})$
- $p(\text{Sổ mũi} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Sổ mũi} = \text{có} | \text{Cúm} = \text{không})$
- $p(\text{Đau đầu} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Đau đầu} = \text{hơi đau} | \text{Cúm} = \text{không})$, $p(\text{Đau đầu} = \text{đau} | \text{Cúm} = \text{không})$
- $p(\text{Sốt} = \text{không} | \text{Cúm} = \text{không})$, $p(\text{Sốt} = \text{có} | \text{Cúm} = \text{không})$
- Tương tự cho ... $\text{Cúm} = \text{có}$

Học các giá trị xác suất này từ tập dữ liệu như thế nào?

Khi đã có các giá trị xác suất ... ta có thể dự đoán với một véc-tơ đầu vào X bất kỳ

Vd: thử dự đoán với $X = [\text{có (ớn lạnh)}, \text{không (sổ mũi)}, \text{hơi đau (đầu)}, \text{không (sốt)}]$

Thuật toán học Naïve Bayes

Để dự đoán được với một véc-tơ đầu vào X bất kỳ, trong quá trình học (huấn luyện), với mỗi lớp v, máy cần học:

- $p(Y = v) = \frac{\text{Số mẫu có } Y=v}{\text{Tổng số mẫu}}$
- $p(X_i|Y = v)$ với mọi X_i và với mọi giá trị có thể có của mỗi X_i
  - Xét một thuộc tính X_i và một giá trị u của X_i:

$$p(X_i = u|Y = v) = \frac{\text{Số mẫu vừa có } X_i = u \text{ vừa có } Y = v}{\text{Số mẫu có } Y = v}$$

Tính giá trị các xác suất từ tập dữ liệu

- $p(\text{Cúm} = \text{không}) = {}^{3}/_{8}$
- $p(\text{Ớn lạnh} = \text{không} | \text{Cúm} = \text{không}) = {}^{2}/_{3}$
- $p(\text{Ớn lạnh} = \text{có} | \text{Cúm} = \text{không}) = {}^{1}/_{3}$

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Tính giá trị các xác suất từ tập dữ liệu

- $p(\text{Sổ mũi} = \text{không} | \text{Cúm} = \text{không}) = {}^{2}/_{3}$
- $p(\text{Sổ mũi} = \text{có} | \text{Cúm} = \text{không}) = {}^{1}/_{3}$

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Tính giá trị các xác suất từ tập dữ liệu

- $p(\text{Đau đầu} = \text{không} | \text{Cúm} = \text{không}) = {}^{1}/_{3}$
- $p(\text{Đau đầu} = \text{hơi đau} | \text{Cúm} = \text{không}) = {}^{1}/_{3}$
- $p(\text{Đau đầu} = \text{đau} | \text{Cúm} = \text{không}) = {}^{1}/_{3}$

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Tính giá trị các xác suất từ tập dữ liệu

- $p(\text{Sốt} = \text{không}|\text{Cúm} = \text{không}) = {}^{2}/_{3}$
- $p(\text{Sốt} = \text{có}|\text{Cúm} = \text{không}) = {}^{1}/_{3}$

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Bài tập

- Tính giá trị các xác suất còn lại của lớp $\text{Cúm} = \text{có}$
- Dự đoán bệnh nhân có $X = [\text{có (ớn lạnh)}, \text{không (sổ mũi)}, \text{hơi đau (đầu)}, \text{không (sốt)}]$ có bị cúm hay không?

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Tính giá trị các xác suất từ tập dữ liệu

- $p(\text{Cúm} = \text{có}) = {}^{5}/_{8}$
- $p(\text{Ớn lạnh} = \text{không} | \text{Cúm} = \text{có}) = {}^{2}/_{5}$
- $p(\text{Ớn lạnh} = \text{có} | \text{Cúm} = \text{có}) = {}^{3}/_{5}$

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Tính giá trị các xác suất từ tập dữ liệu

- $p(\text{Sổ mũi} = \text{không} | \text{Cúm} = \text{có}) = {}^{1}/_{5}$
- $p(\text{Sổ mũi} = \text{có} | \text{Cúm} = \text{có}) = {}^{4}/_{5}$

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Tính giá trị các xác suất từ tập dữ liệu

- $p(\text{Đau đầu} = \text{không}|\text{Cúm} = \text{có}) = {}^{1}/_{5}$
- $p(\text{Đau đầu} = \text{hơi đau}|\text{Cúm} = \text{có}) = {}^{2}/_{5}$
- $p(\text{Đau đầu} = \text{đau}|\text{Cúm} = \text{có}) = {}^{2}/_{5}$

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Tính giá trị các xác suất từ tập dữ liệu

- $p(\text{Sốt} = \text{không}|\text{Cúm} = \text{có}) = {}^{1}/_{5}$
- $p(\text{Sốt} = \text{có}|\text{Cúm} = \text{có}) = {}^{4}/_{5}$

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Dự đoán

Dự đoán bệnh nhân có $X = [\text{có (ớn lạnh)}, \text{không (sổ mũi)}, \text{hơi đau (đầu)}, \text{không (sốt)}]$ có bị cúm hay không?

Tử của $p(\text{Cúm} = \text{không}|X) = p(X|\text{Cúm} = \text{không}) \times p(\text{Cúm} = \text{Không})$

$= p(\text{Ớn lạnh} = \text{có}|\text{Cúm} = \text{không}) \times p(\text{Sổ mũi} = \text{không}|\text{Cúm} = \text{không})$

$\times p(\text{Đau đầu} = \text{hơi đau}|\text{Cúm} = \text{không}) \times p(\text{Sốt} = \text{có}|\text{Cúm} = \text{không}) \times p(\text{Cúm} = \text{Không})$

$= \frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{3}{8} \approx 0.019$

Dự đoán: không cúm!

Tử của $p(\text{Cúm} = \text{có}|X) = p(X|\text{Cúm} = \text{có}) \times p(\text{Cúm} = \text{có})$

$= p(\text{Ớn lạnh} = \text{có}|\text{Cúm} = \text{có}) \times p(\text{Sổ mũi} = \text{không}|\text{Cúm} = \text{có})$

$\times p(\text{Đau đầu} = \text{hơi đau}|\text{Cúm} = \text{có}) \times p(\text{Sốt} = \text{có}|\text{Cúm} = \text{có}) \times p(\text{Cúm} = \text{có})$

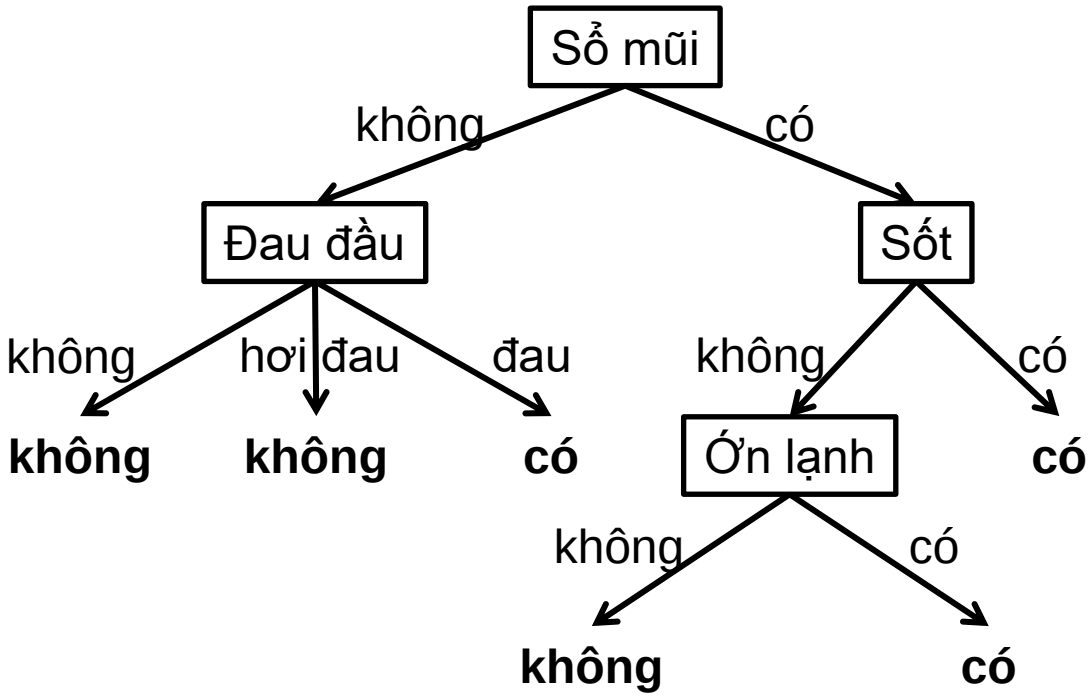
$= \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{5}{8} = 0.006$

Nội dung tiếp theo

Cây quyết định ID3: một thuật toán học-cách-phân-lớp đơn giản khác

Cây quyết định

- Cây quyết định cho biết các luật để dự đoán
- Vd, với cây quyết định ở dưới, ta sẽ có các luật để dự đoán:
  - NẾU Sổ mũi = không VÀ Đau đầu = không THÌ Cúm = **không**
  - NẾU Sổ mũi = không VÀ Đau đầu = hơi đau THÌ Cúm = **không**
  - NẾU Sổ mũi = không VÀ Đau đầu = đau THÌ Cúm = **có**
  - …



Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

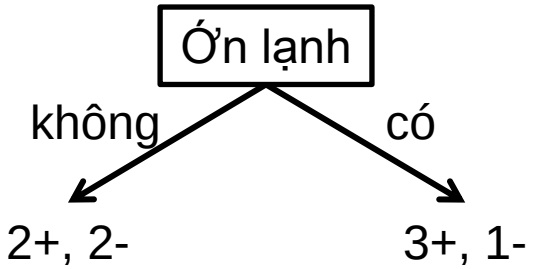
Đầu tiên, chọn một thuộc tính làm thuộc tính gốc của cây (nên chọn thuộc tính nào?)

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

Đầu tiên, chọn một thuộc tính làm thuộc tính gốc của cây (nên chọn thuộc tính nào?)

– Giả sử chọn đại "Ớn lạnh"



Chưa phân lớp hoàn toàn các mẫu → cần làm tiếp

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

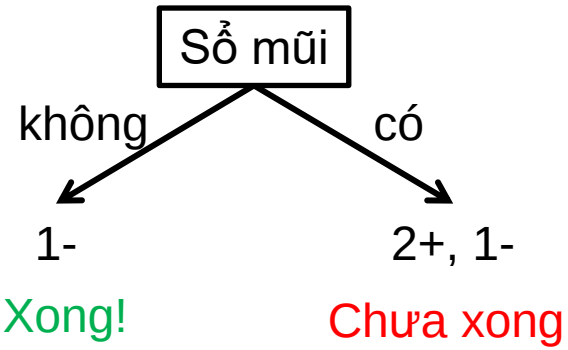
Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

Chọn thuộc tính kế tiếp cho nhánh "Ớn lạnh = không" (nên chọn thuộc tính nào?)

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

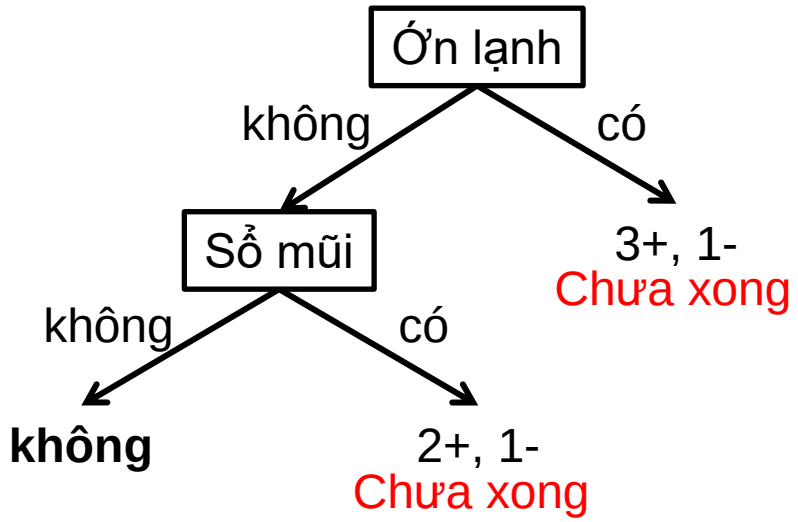
Chọn thuộc tính kế tiếp cho nhánh "Ớn lạnh = không" (nên chọn thuộc tính nào?)



– Giả sử chọn đại "Sổ mũi"

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Cây đến thời điểm hiện tại



Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

Chọn thuộc tính kế tiếp cho nhánh

"Ớn lạnh = không" và "Sổ mũi = có"

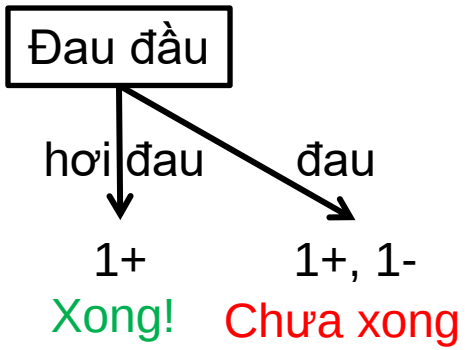
(nên chọn thuộc tính nào?)

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

Chọn thuộc tính kế tiếp cho nhánh "Ớn lạnh = không" và "Sổ mũi = có"

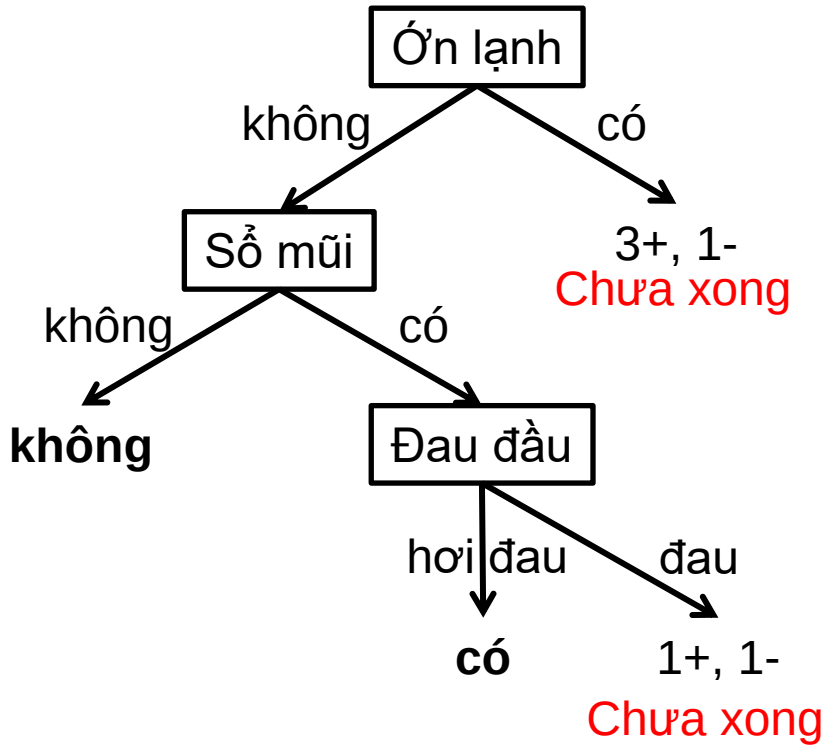
(nên chọn thuộc tính nào?)



– Giả sử chọn đại "Đau đầu"

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Cây đến thời điểm hiện tại



Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

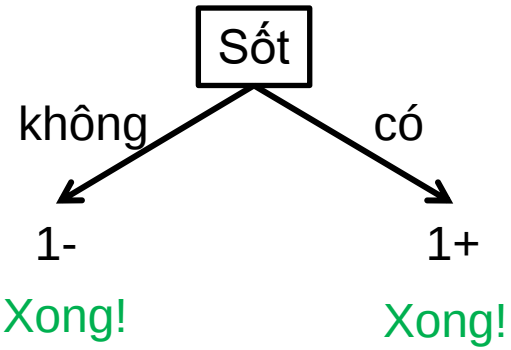
Chọn thuộc tính kế tiếp cho nhánh "Ớn lạnh = không" và "Sổ mũi = có" và "Đau đầu = đau" (nên chọn thuộc tính nào?)

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

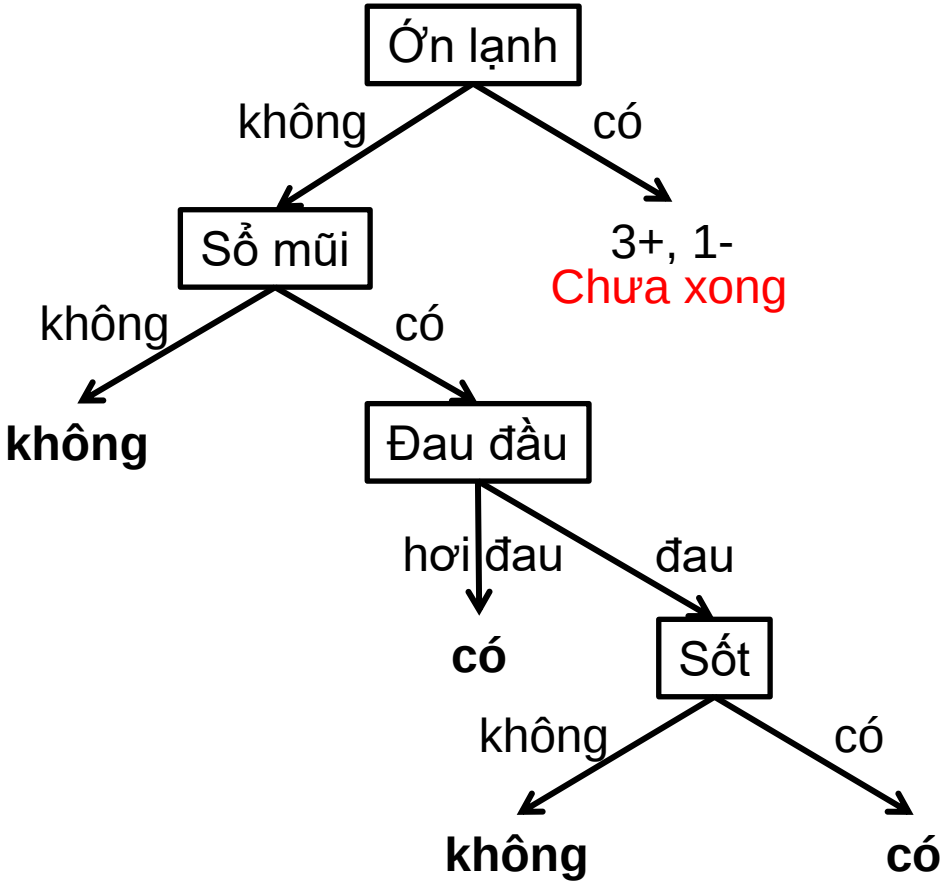
Chọn thuộc tính kế tiếp cho nhánh "Ớn lạnh = không" và "Sổ mũi = có" và "Đau đầu = đau" (nên chọn thuộc tính nào?)

– Chỉ còn một thuộc tính để chọn là "Sốt"



Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Cây đến thời điểm hiện tại



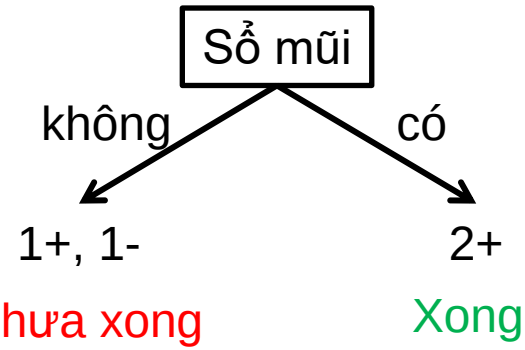
Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

Chọn thuộc tính kế tiếp cho nhánh "Ớn lạnh = có" (nên chọn thuộc tính nào?)

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

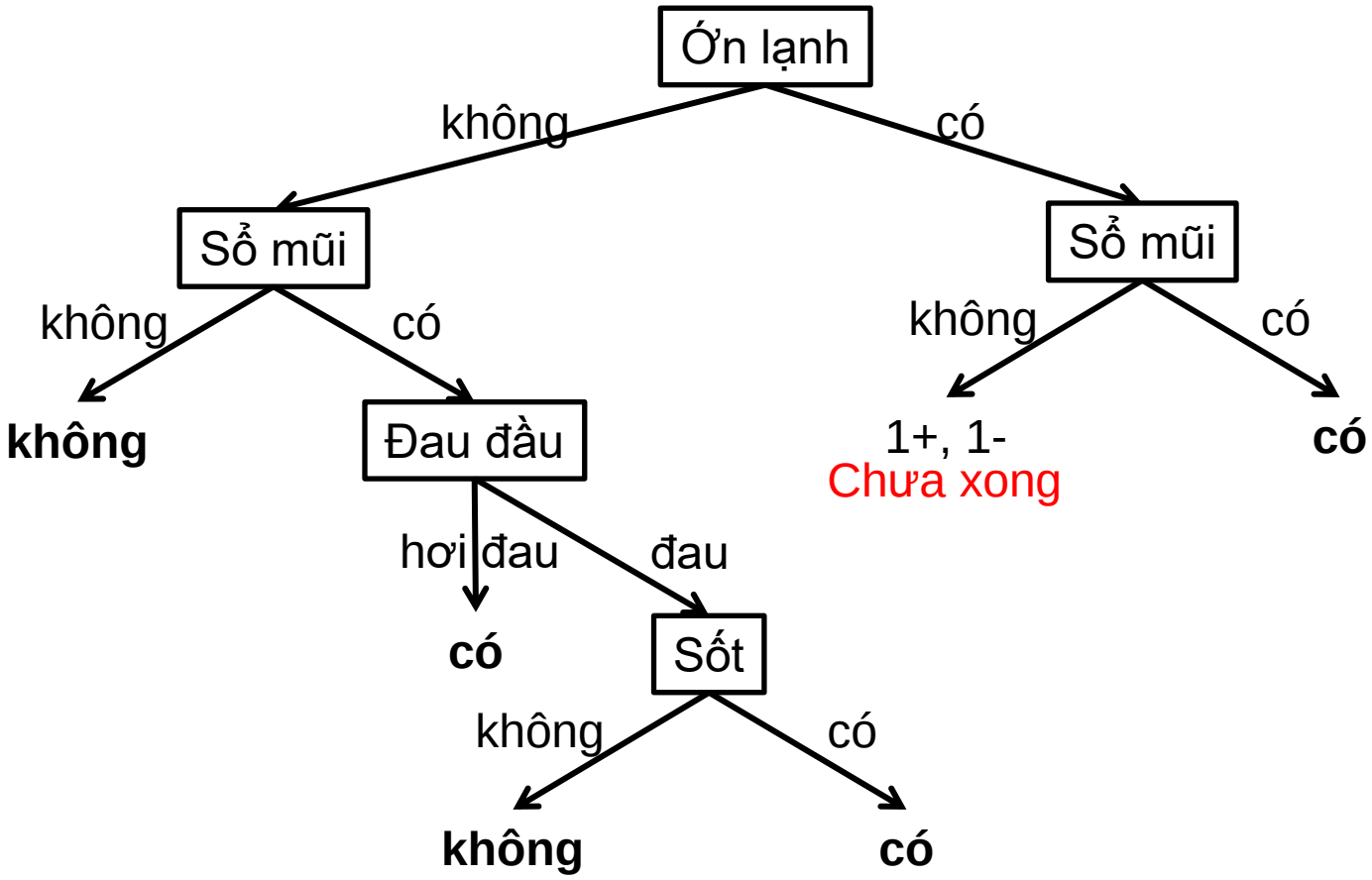
Chọn thuộc tính kế tiếp cho nhánh "Ớn lạnh = có" (nên chọn thuộc tính nào?)



– Giả sử chọn đại "Sổ mũi"

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Cây đến thời điểm hiện tại



Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

Chọn thuộc tính kế tiếp cho nhánh

"Ớn lạnh = có" và "Sổ mũi = không"

(nên chọn thuộc tính nào?)

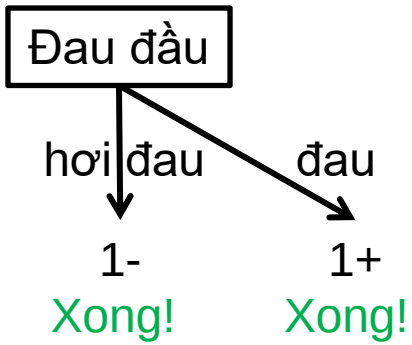
Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Thuật toán ID3: thuật toán học cây quyết định từ tập dữ liệu

Chọn thuộc tính kế tiếp cho nhánh "Ớn lạnh = có" và "Sổ mũi = không"

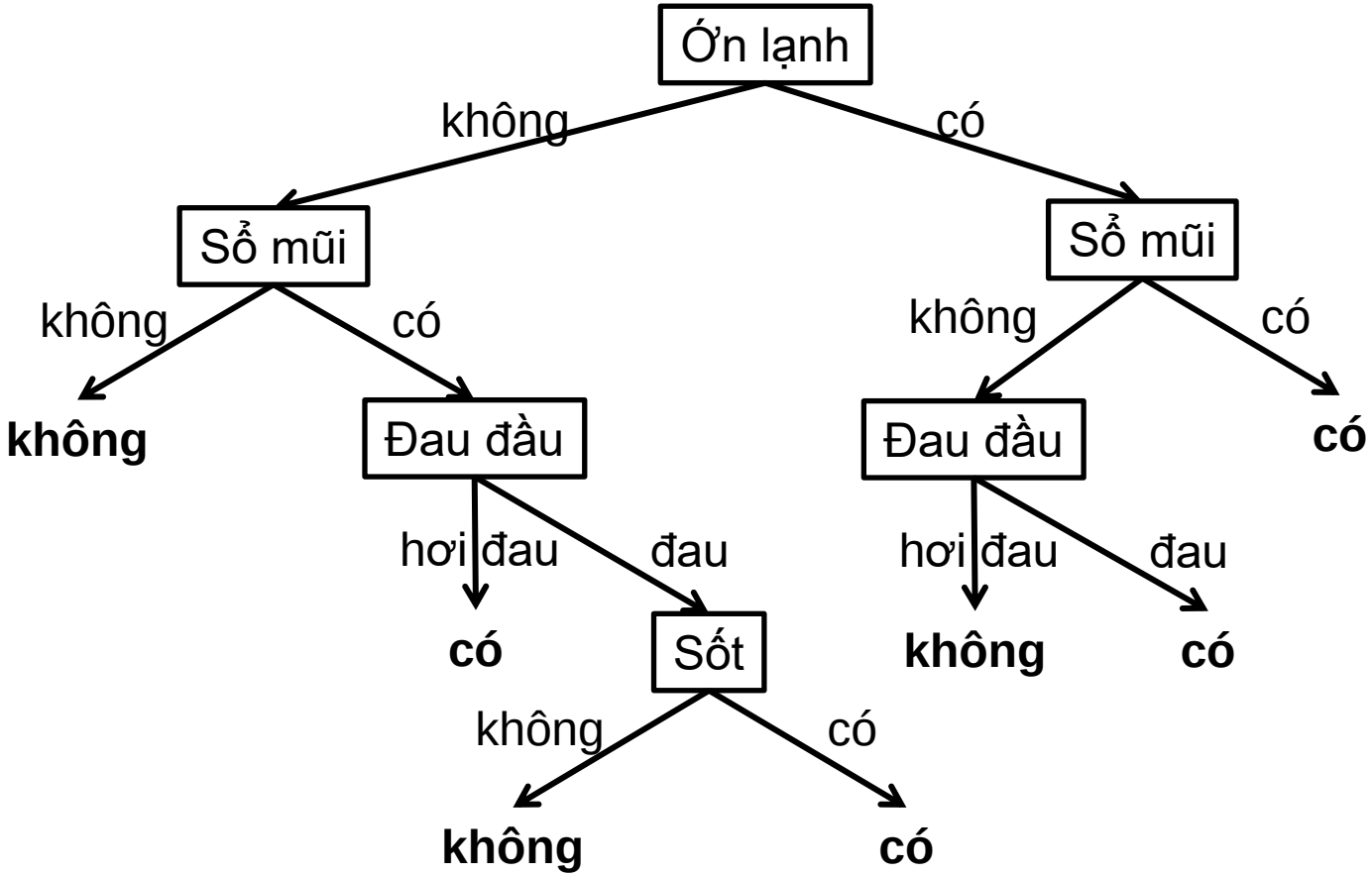
(nên chọn thuộc tính nào?)

- Giả sử chọn đại "Đau đầu"



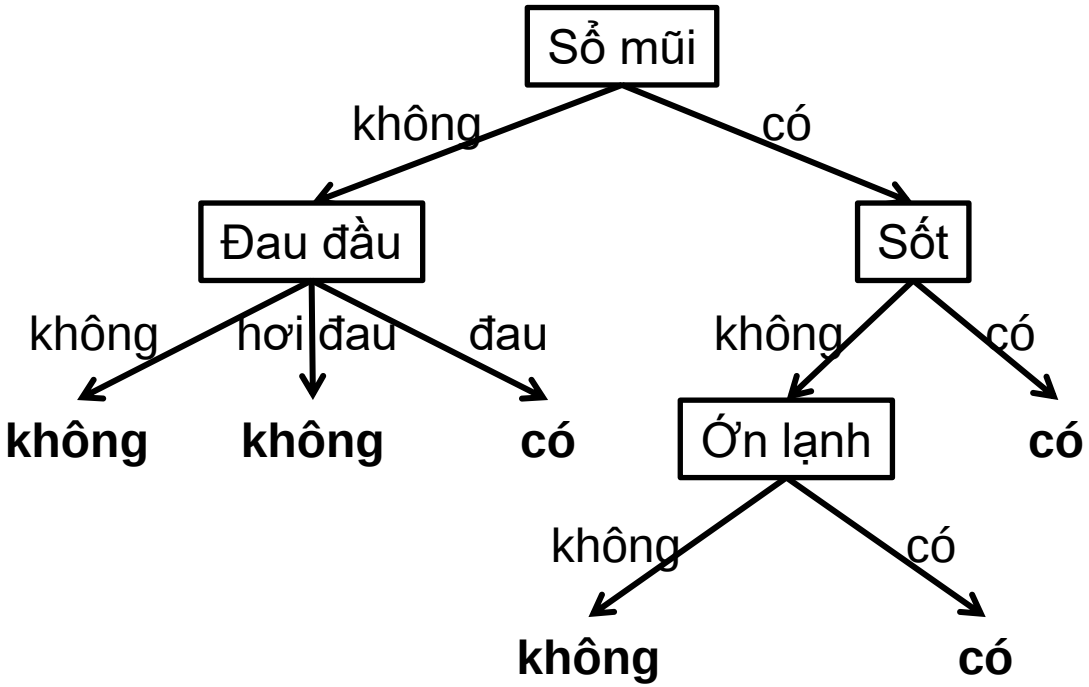
Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Cây cuối cùng



Ở mỗi bước trong quá trình xây cây, nên chọn thuộc tính nào?

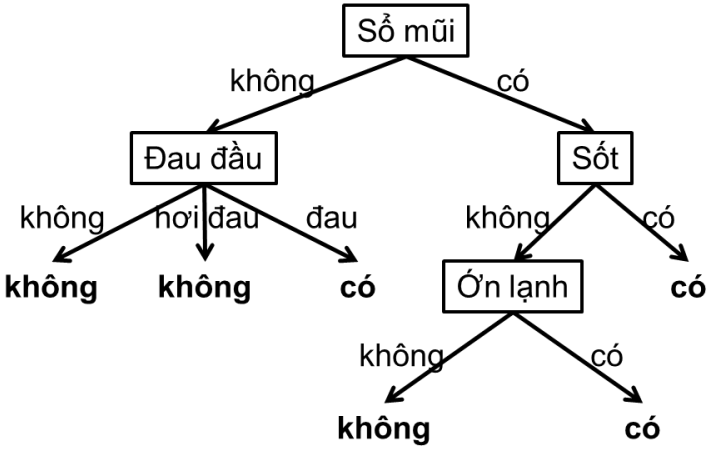
- Các cách chọn thuộc tính khác nhau sẽ đưa tới các cây khác nhau
- Dưới đây là một cây có được với một cách chọn thuộc tính khác với cách vừa làm:

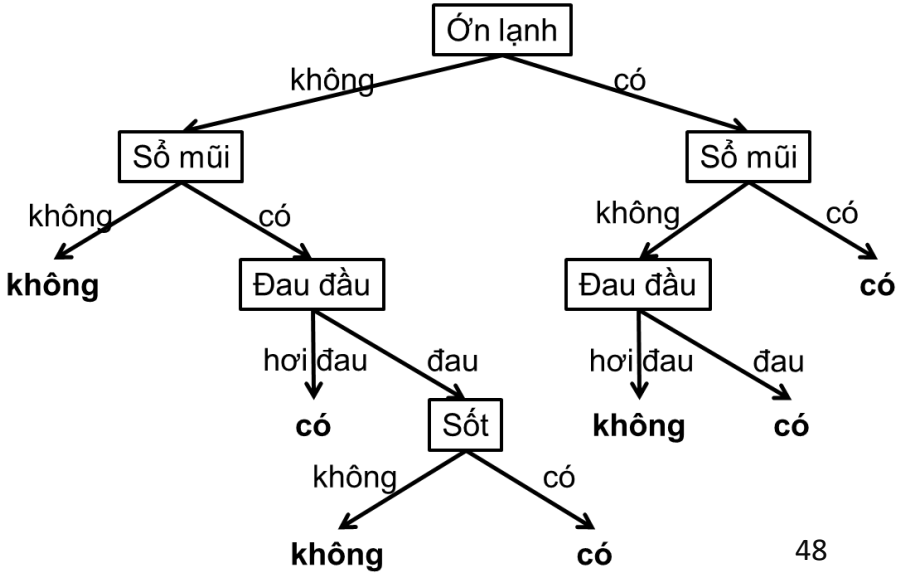


Ở mỗi bước trong quá trình xây cây, nên chọn thuộc tính nào?

Cả 2 cây đều phân lớp đúng tất cả các mẫu trong tập dữ liệu, nhưng bạn nghĩ cây nào tốt hơn?

- Cây đơn giản hơn (cây bên trái) tốt hơn, vì khả năng tổng quát hóa (khả năng dự đoán với các mẫu mới) tốt hơn
- Vậy: ở mỗi bước trong quá trình xây cây, ta nên chọn thuộc tính mà sẽ làm cho cây đơn giản nhất





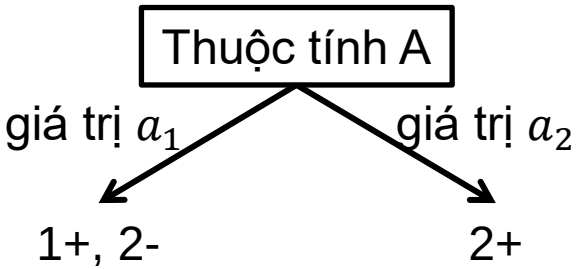
Ở mỗi bước trong quá trình xây cây, nên chọn thuộc tính *làm cho cây đơn giản nhất*

Thuộc tính nào làm cho cây đơn giản nhất?

- Thuộc tính có giá trị AE **(Average Entropy - Entropy trung bình)** nhỏ nhất
- Cách tính giá trị entropy trung bình AE của một thuộc tính A:
  - Bước 1: tính giá trị entropy E_a của từng giá trị a của thuộc tính A

p_{av}: tỉ lệ các mẫu thuộc về lớp v trong các mẫu có $A = a$

$$E_a = -\sum_{v \in Tập-các-lớp} p_{av} \times \log_2 p_{av}$$



$$E_{a_1} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} \approx 0.918$$

$$E_{a_2} = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$$

Giá trị a mà càng gần với việc phân lớp hoàn toàn các mẫu sẽ có entropy càng nhỏ!

Ở mỗi bước trong quá trình xây cây, nên chọn thuộc tính *làm cho cây đơn giản nhất*

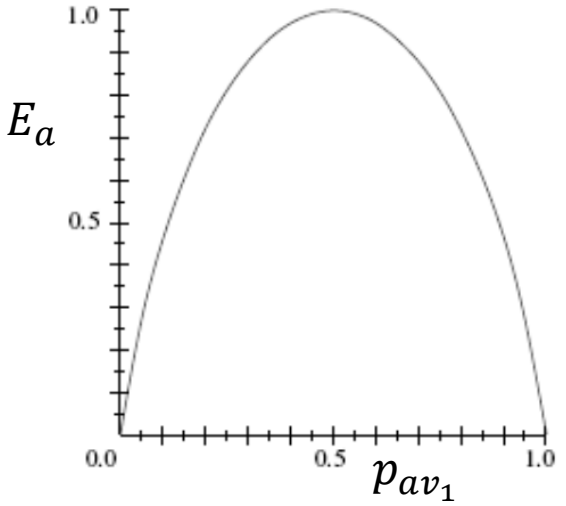
Thuộc tính nào làm cho cây đơn giản nhất?

- Thuộc tính có giá trị AE **(Average Entropy - Entropy trung bình)** nhỏ nhất
- Cách tính giá trị entropy trung bình AE của một thuộc tính A:
  - Bước 1: tính giá trị entropy E_a của từng giá trị a của thuộc tính A

p_{av}: tỉ lệ các mẫu thuộc về lớp v trong các mẫu có $A = a$

$$E_a = -\sum_{v \in Tập-các-lớp} p_{av} \times \log_2 p_{av}$$

Đồ thị thể hiện mối quan hệ giữa E_a và p_{av_1} trong trường hợp chỉ có 2 lớp là v_1 và v_2
(lúc này $p_{av_2} = 1 - p_{av_1}$)



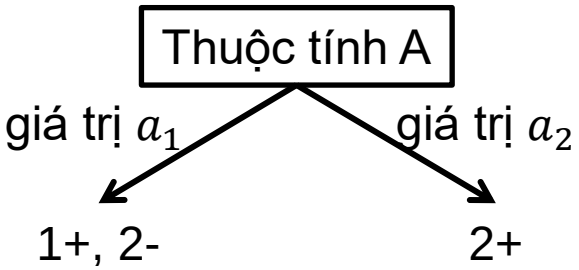
Ở mỗi bước trong quá trình xây cây, nên chọn thuộc tính *làm cho cây đơn giản nhất*

Thuộc tính nào làm cho cây đơn giản nhất?

- Thuộc tính có giá trị AE **(Average Entropy - Entropy trung bình)** nhỏ nhất
- Cách tính giá trị entropy trung bình AE của một thuộc tính A:
  - Bước 2:

p_a: tỉ lệ các mẫu có $A = a$

$$AE_A = \sum_{a \in Tập-giá-trị-của-A} p_a \times E_a$$



$$E_{a_1} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} \approx 0.918$$

$$E_{a_2} = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$$

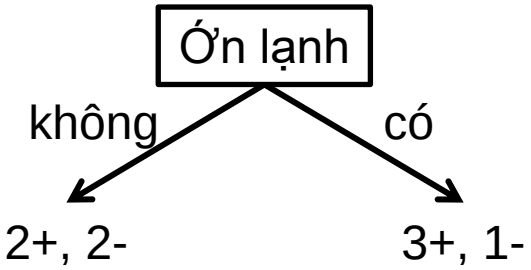
$$AE_A = \frac{3}{5} \times 0.918 + \frac{2}{5} \times 0$$

Chạy lại thuật toán ID3

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Chọn thuộc tính gốc của cây

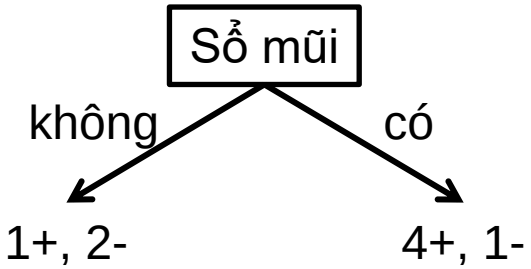
- $E_{không} = -\frac{2}{4} \times \log_2 \frac{2}{4} - \frac{2}{4} \times \log_2 \frac{2}{4} = 1$
- $E_{có} = -\frac{3}{4} \times \log_2 \frac{3}{4} - \frac{1}{4} \times \log_2 \frac{1}{4} \approx 0.811$
- $AE = \frac{4}{8} \times 1 + \frac{4}{8} \times 0.811 \approx 0.906$



Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Chọn thuộc tính gốc của cây

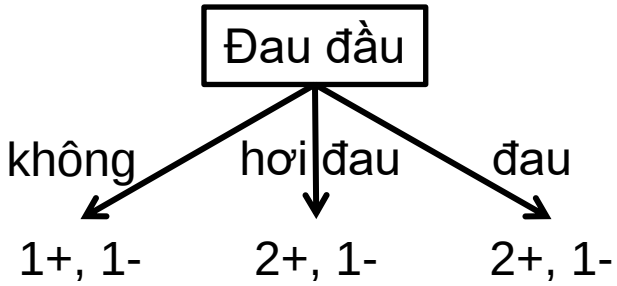
- $E_{không} = -\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3} \approx 0.918$
- $E_{có} = -\frac{4}{5} \times \log_2 \frac{4}{5} - \frac{1}{5} \times \log_2 \frac{1}{5} \approx 0.722$
- $AE = \frac{3}{8} \times 0.918 + \frac{5}{8} \times 0.722 \approx 0.796$



Input				Output
Ớn lạnh	Sổ mũi	Đau đầu	Sốt	Cúm
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Chọn thuộc tính gốc của cây

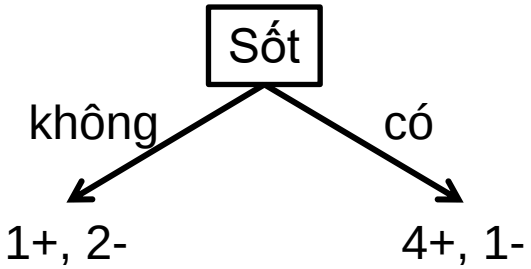
- $E_{không} = -\frac{1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \times \log_2 \frac{1}{2} = 1$
- $E_{hơi\ đau} = -\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3} \approx 0.918$
- $E_{đau} = -\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3} \approx 0.918$
- $AE = \frac{2}{8} \times 1 + \frac{3}{8} \times 0.918 + \frac{3}{8} \times 0.918 \approx 0.939$



				Output
Ớn lạnh	Sổ mũi	Đau đầu	Sốt	Cúm
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Chọn thuộc tính gốc của cây

- $E_{không} = -\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3} \approx 0.918$
- $E_{có} = -\frac{4}{5} \times \log_2 \frac{4}{5} - \frac{1}{5} \times \log_2 \frac{1}{5} \approx 0.722$
- $AE = \frac{3}{8} \times 0.918 + \frac{5}{8} \times 0.722 \approx 0.796$

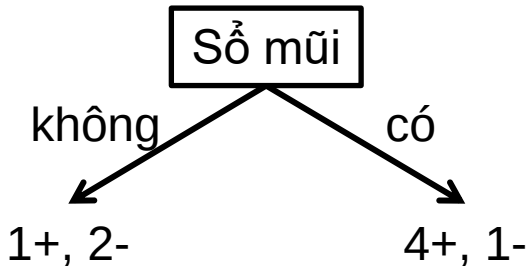


Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Chọn thuộc tính gốc của cây

- Ớn lạnh: $AE = 0.906$
- Sổ mũi: $AE = 0.796$
- Đau đầu: $AE = 0.939$
- Sốt: $AE = 0.796$

→ Chọn “Sổ mũi” hoặc “Đau đầu” đều được; giả sử chọn “Sổ mũi”

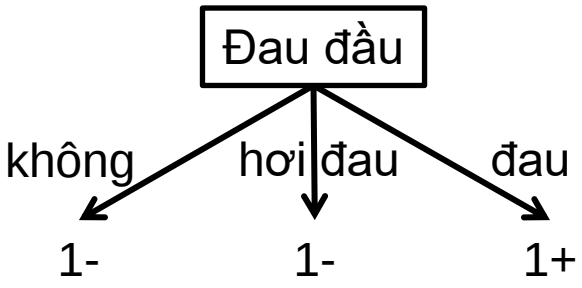


Chọn thuộc tính tiếp theo của nhánh "Sổ mũi = không"

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

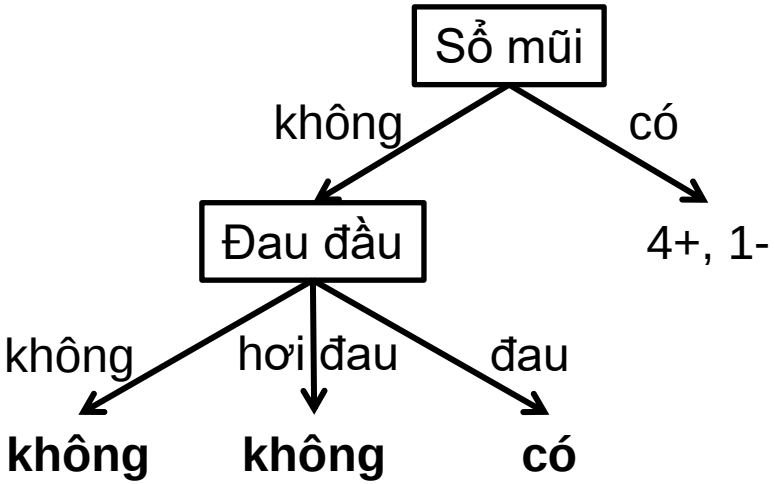
Chọn thuộc tính tiếp theo của nhánh "Sổ mũi = không"

Làm nhanh: nhìn sơ bộ các thuộc tính thì thấy thuộc tính "Đau đầu" sẽ phân lớp hoàn toàn các mẫu ($AE = 0$) → chọn luôn "Đau đầu" là thuộc tính tiếp theo của nhánh "Sổ mũi = không"



Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Cây cho tới thời điểm hiện tại

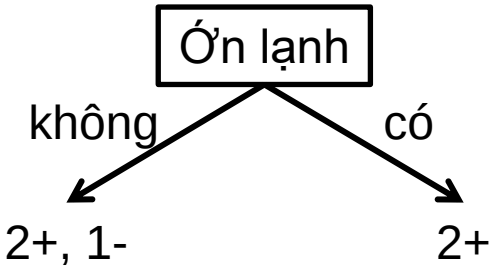


Chọn thuộc tính tiếp theo của nhánh "Sổ mũi = có"

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Chọn thuộc tính tiếp theo của nhánh "Sổ mũi = có"

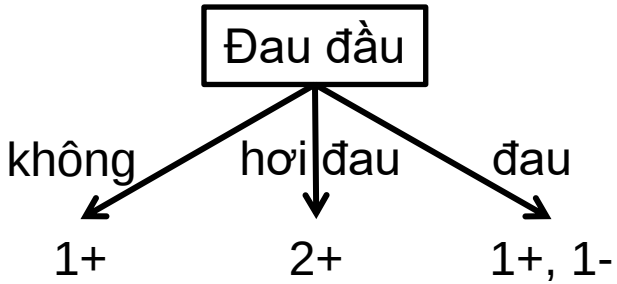
- $E_{không} = -\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3} \approx 0.918$
- $E_{có} = 0$
- $AE = \frac{3}{5} \times 0.918 + \frac{2}{5} \times 0 \approx 0.551$



Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Chọn thuộc tính tiếp theo của nhánh "Sổ mũi = có"

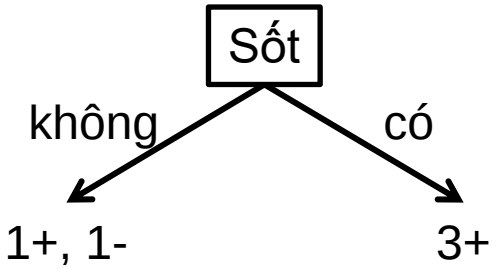
- $E_{không} = 0$
- $E_{hơi\ đau} = 0$
- $E_{đau} = 1$
- $AE = \frac{1}{5} \times 0 + \frac{2}{5} \times 0 + \frac{2}{5} \times 1 = 0.4$



Input				Output
Ớn lạnh	Sổ mũi	Đau đầu	Sốt	Cúm
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Chọn thuộc tính tiếp theo của nhánh "Sổ mũi = có"

- $E_{không} = 1$
- $E_{có} = 0$
- $AE = \frac{2}{5} \times 1 + \frac{3}{5} \times 0 = 0.4$

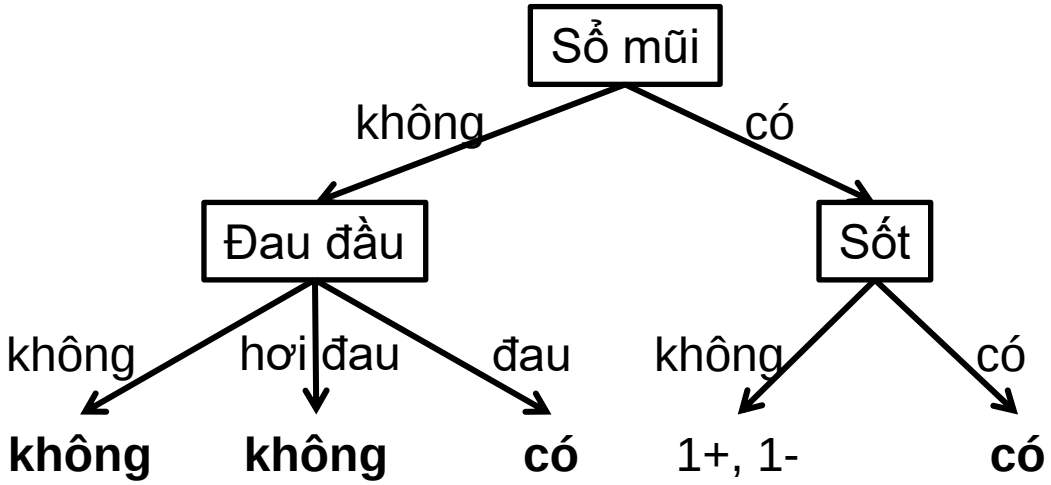


Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Chọn thuộc tính tiếp theo của nhánh "Sổ mũi = có"

- Ớn lạnh: $AE = 0.551$
- Đau đầu: $AE = 0.4$
- Sốt: $AE = 0.4$

→ Chọn "Đau đầu" hoặc "Sốt" đều được; giả sử chọn "Sốt"

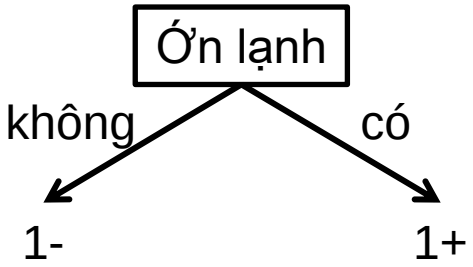


Chọn thuộc tính tiếp theo của nhánh "Sổ mũi = có" và "Sốt = không"

Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

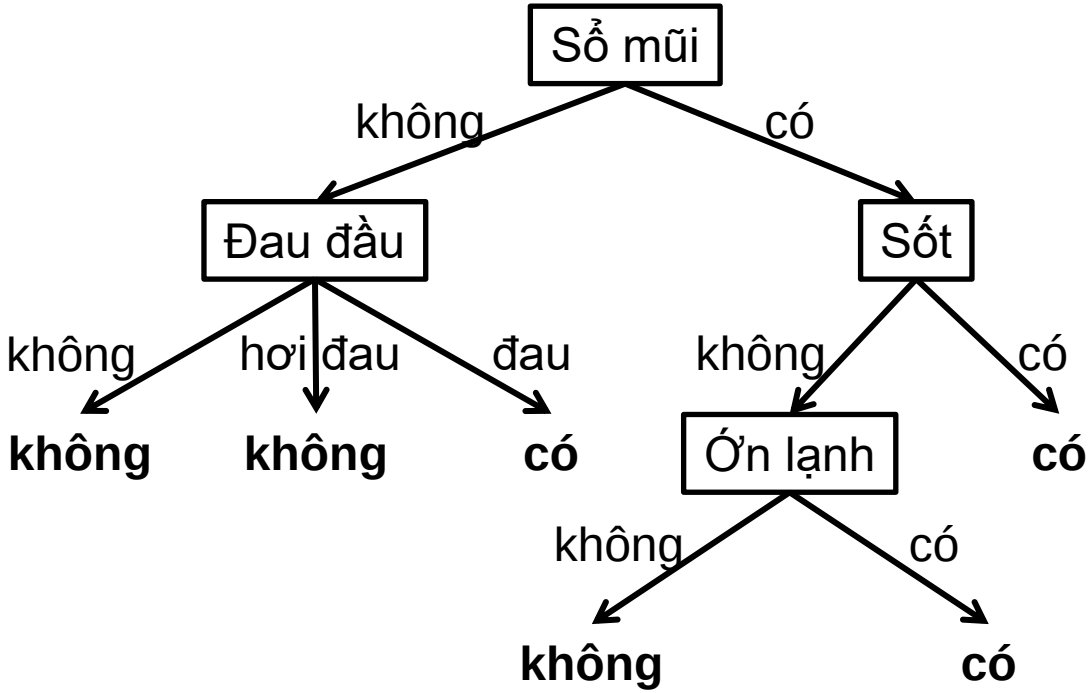
Chọn thuộc tính tiếp theo của nhánh "Sổ mũi = có" và "Sốt = không"

Ta thấy thuộc tính "Ớn lạnh" phân lớp hoàn toàn các mẫu ($AE = 0$) → chọn luôn "Ớn lạnh" là thuộc tính tiếp theo của nhánh "Sổ mũi = có" và "Sốt = không"



Input				Output
Ớn lạnh	**Sổ mũi**	**Đau đầu**	**Sốt**	**Cúm**
có	không	hơi đau	có	không
có	có	không	không	có
có	không	đau	có	có
không	có	hơi đau	có	có
không	không	không	không	không
không	có	đau	có	có
không	có	đau	không	không
có	có	hơi đau	có	có

Cây cuối cùng



ID3: Bài tập ví dụ 1

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

ID3: Bài tập ví dụ 2

- Hãy xây dựng mô hình cây quyết định ID3 từ tập dữ liệu huấn luyện như bên dưới. Chọn thuộc tính theo độ đo Entropy. Biết rằng thuộc tính phân lớp là Lớp

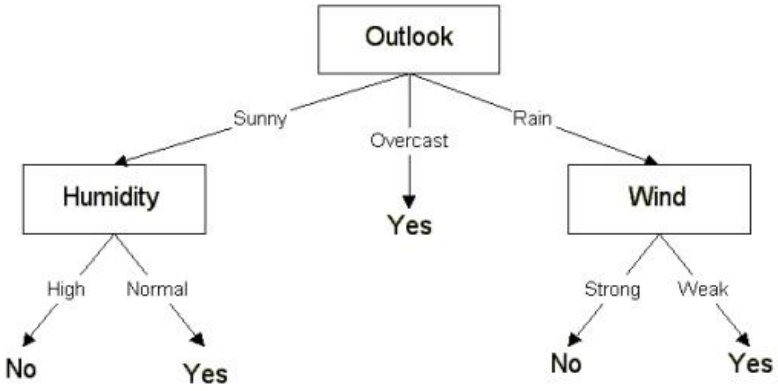
Mẫu	Độ tuổi	Hệ số	Khu vực	Lớp
1	15	1	A	C1
2	20	3	B	C2
3	25	2	A	C1
4	30	4	A	C1
5	35	2	B	C2
6	25	4	A	C1
7	15	2	B	C2
8	20	3	B	C2

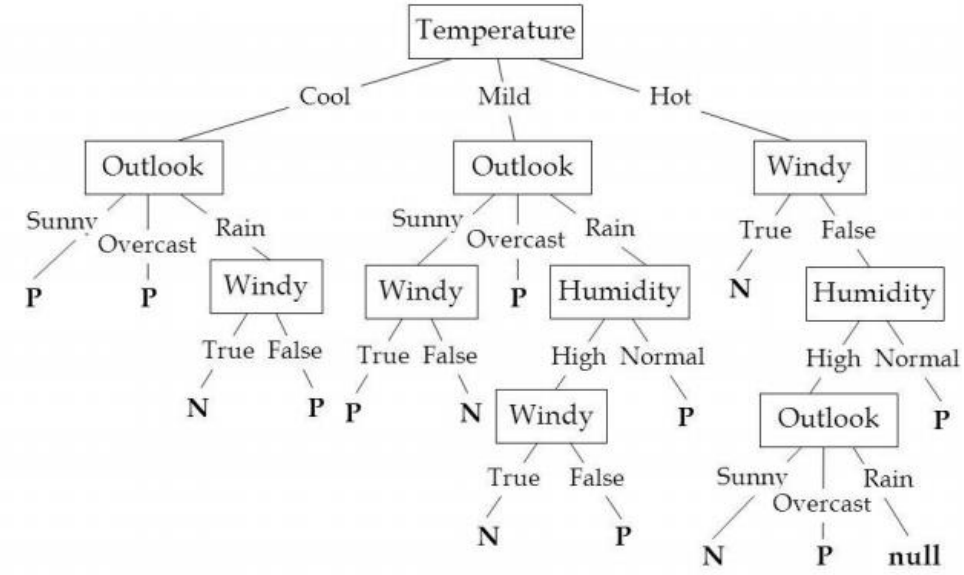
outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

outlook	temperature	humidity	windy	play
overcast	mild	normal	TRUE	?

Cây quyết định ID3

- Nguyên lý Occam Razor: những cây đơn giản là những cây quyết định tốt hơn





Tri thức dạng luật

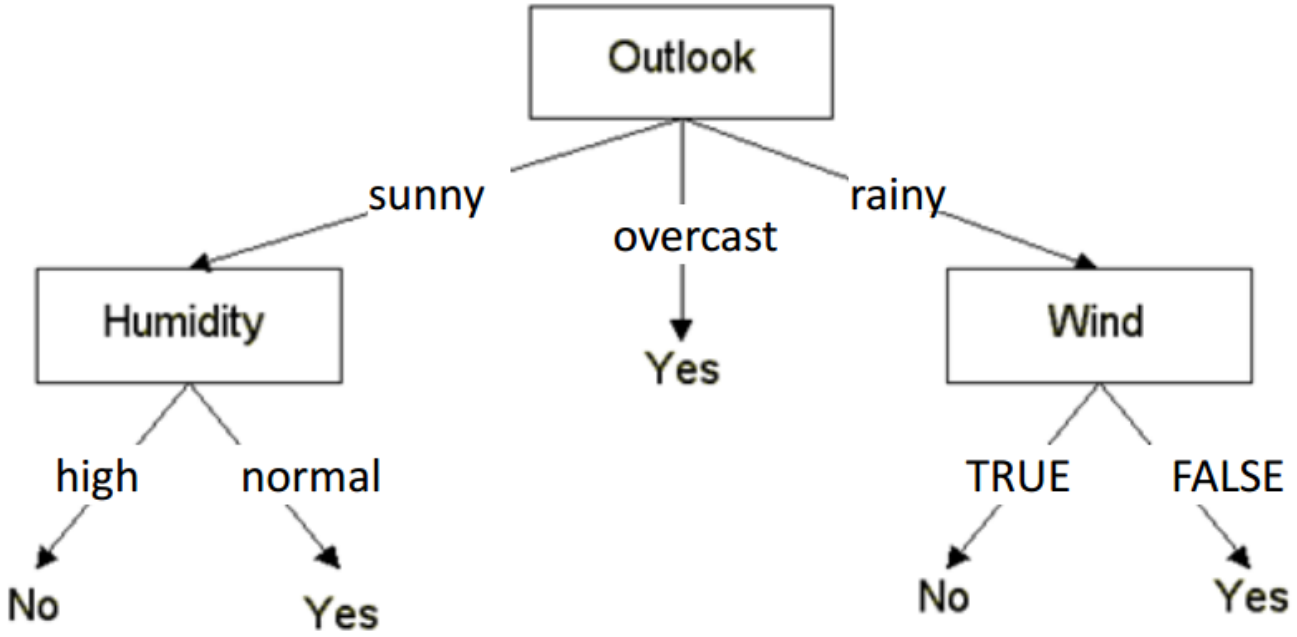
- Tri thức được biểu diễn dưới dạng luật:

  IF Điều kiện 1 $\wedge$ Điều kiện 2…

  THEN Kết luận

- Dễ hiểu với con người, được sử dụng chủ yếu trong các hệ chuyên gia
- Rút luật từ cây quyết định: đi từ nút gốc đến nút lá, lấy các phép kiểm tra làm tiền đề và phân loại của nút lá làm kết quả

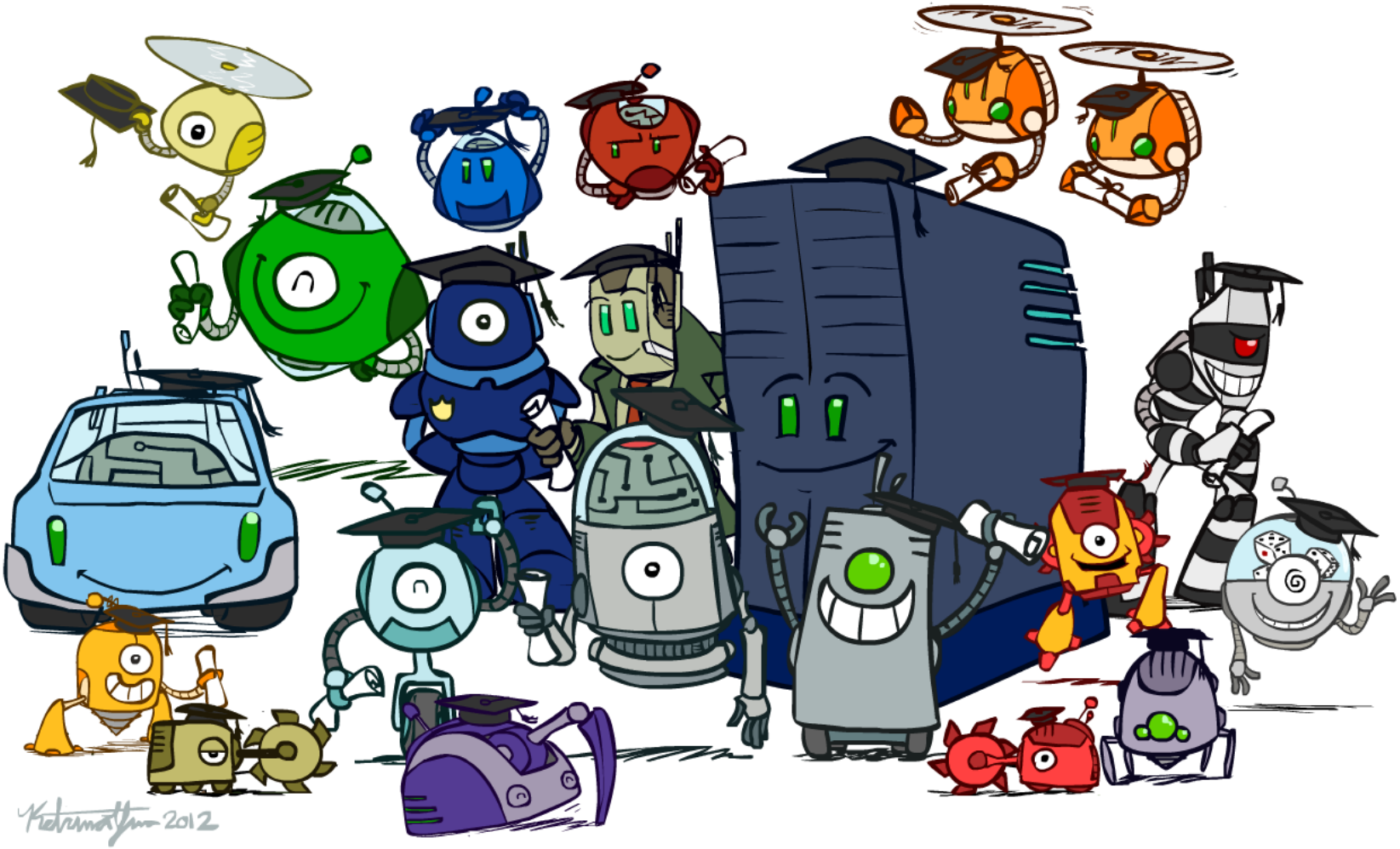
Tri thức dạng luật: ví dụ



- IF outlook = sunny AND humidity = high THEN play = no
- IF outlook = sunny AND humidity = normal THEN play = yes
- IF outlook = overcast THEN play = yes
- IF outlook = rainy AND windy = TRUE THEN play = no
- IF outlook = rainy AND windy = TRUE THEN play = yes

Tổng kết

- **Tìm kiếm**
  - DFS, BFS
  - UCS
  - A*
- **Tìm kiếm đối kháng**
  - Minimax + tỉa alpha-beta
  - Expectimax
- **Logic**
  - Kiểm tra suy dẫn bằng thuật toán hợp giải Robinson
- **Máy học**
  - Naïve Bayes
  - Cây quyết định ID3



Thank you

Nguồn ảnh: UC Berkely. CS188 Intro to AI