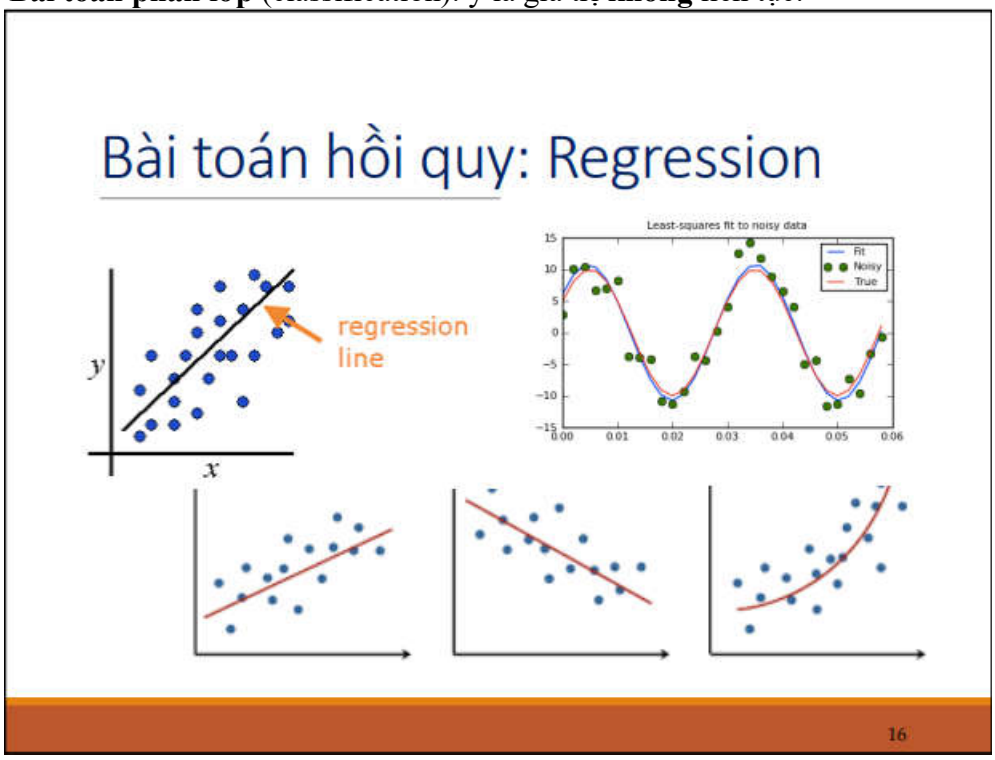
MÁY HỌC ỨNG DỤNG

A. Lý thuyết

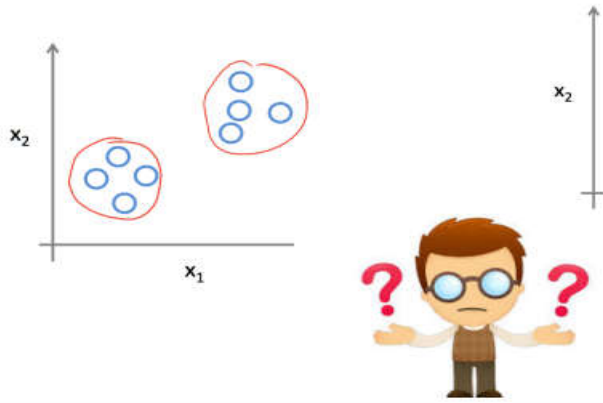
I. Các phương pháp học từ dữ liệu (C4 - Nhập môn AI):

- Học không có giám sát (Unsupervised Learning):
  + Kmeans, Hierachical Clustering.
  + Thực hiện mô hình hóa một tập data input, **không được gán nhãn**(lớp, giá trị cần predict).
  + Gom cụm, gom nhóm (Clustering): Xây dựng mô hình gom cụm data tập học (không có nhãn) sao cho các data cùng nhóm có các tính chất tương tự nhau và data của 2 nhóm khác nhau sẽ có các tính chất khác nhau.

- Học có giám sát (Supervised Learning):
  + Là thuật toán học tạo ra một hàm ánh xạ data input tới KQ đích mong muốn (nhãn, lớp, giá trị cần dự báo). Tập dữ liệu dùng để huấn luyện phải **được gán nhãn, lớp hay giá trị cần dự báo**.
  + **Bài toán hồi quy** (regression): y (nhãn) là giá trị liên tục.
  + **Bài toán phân lớp** (classification): y là giá trị **không** liên tục.



Bài toán phân lớp



17

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]

Chỉ ra thuộc tính? Nhãn/lớp của tập dữ liệu thời tiết trong bảng trên

\+ Thuộc tính: Outlook, Temperature, Humidity, Wind.
\+ Nhãn: PlayTennis.
=> Đây là **bài toán phân lớp.**

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

+ Thuộc tính: Outlook, Temp., Humidity, Wind.
+ Nhãn: Golf Players.
=> Đây là **bài toán hồi quy.**

+ Đối với **bài toán phân lớp**, xây dựng mô hình phân loại dựa trên data tập học đã có **nhãn** (lớp) **là kiểu liệt kê.**

VD: Có sẵn tập data thư điện thử, *mỗi thư có 1 nhãn là thư rác hay thư bình thường*, mục tiêu là *build mô hình phân lớp tập data thư điện tử thành thư rác hay thư bình thường* để khi có một *thư điện tử mới đến thì mô hình dự báo được thư này có phải là thư rác hay không*.

+ Đối với **bài toán hồi quy**, xây dựng mô hình phân loại dựa trên data tập học đã có nhãn (lớp) là **giá trị liên tục**.

VD: Build mô hình dự báo nước sông Mekong từ các yếu tố thời tiết, mùa, …….

- Học bán giám sát: Kết hợp không giám sát và giám sát.

- Học tăng cường (Reinforcement Learning):

+ Là cách tiếp cận tập trung vào việc học để hoàn thành được mục tiêu bằng việc tương tác trực tiếp với MT.

+ Là các bài toán giúp cho một hệ thống tự động xác định hành động dựa vào môi trường cụ thể để đạt được hiệu quả cao nhất.

+ Bản chất là **trial-and-eror**, nghĩa là thử đi thử lại và rút ra kinh nghiệm sau mỗi lần thử.

***Tóm gọn lại:**

Từ tập dữ liệu đầu vào, xác định các thuộc tính (X) và nhãn (Y):

- Nếu có Y, xét Y:
  + Y rời rạc => Phân lớp.
  + Y liên tục => Hồi quy.
- Nếu không có Y => Gom cụm, gom nhóm.

II. Giải thuật K láng giềng (KNN):

- Là thuật toán phân lớp các trường hợp mới đến dựa trên một số lượng thông tin "phổ biến" nhất của **k** láng giềng gần nhất với nó.
- Để xác định được lớp của phần tử mới đến:
  + Tính toán khoảng cách từ phần tử mới đến các phần tử còn lại trong tập huấn luyện.
  + Cho **k** phần tử gần nhất với phần tử mới trong tập huấn luyện.
  + Gán nhãn cho phần tử mới bằng nhãn "phổ biến" nhất của **k** láng giềng gần nhất.
- Cách chọn **k**: Nên chọn **k** là giá trị lẻ và đủ lớn.
- Các độ đo khoảng cách: Khoảng cách được tính theo từng kiểu của data:
  + Kiểu số.
  + Kiểu rời rạc (nominal type).
  + Nhị phân.
- Đo khoảng cách **kiểu số**:

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q)}$$

$i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ và $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ là 2 phần tử data trong p-dimensional, q là số nguyên dương.

Nếu $q = 1$, d là khoảng cách Manhattan:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

Nếu $q = 2$, d là khoảng cách Euclid:

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)}$$

- Đo khoảng cách **kiểu rời rạc**:

Trong đó:

+ m: là số lượng matches.
+ p: là tổng số biến (thuộc tính).
+ Khoảng cách được định nghĩa:

$$d(i,j) = \frac{p - m}{p}$$

VD1:

	Màu tóc	Màu mắt	Chiều cao	Cân nặng	Trình độ
Nam	Đen	Đen	Cao	Trung bình	Cao đẳng
Lan	Nâu	Đen	Thấp	Trung bình	Đại học
Điệp	Nâu	Đen	Cao	Trung bình	Cao đẳng

d(Nam, Lan) =? **d(Nam, Điệp) =?**

d(Nam, Lan) = 2 / 5	d(Nam, Điệp) = 1 / 5

VD2:

Cách tính khoảng cách giữa các phần tử

- **Mỗi phần tử được biểu diễn bởi tập hợp các thuộc tính**

John:
Tuổi = 35
Thu nhập = 95K
Số thẻ tín dụng = 3

Mary:
Tuổi = 41
Thu nhập = 215K
Số thẻ tín dụng = 2

- **Sử dụng khoảng cách Euclide giữa 2 phần tử.**

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$$

Khoảng cách (John, Mary) =
sqrt [(35-41)²+(95K-215K)² +(3-2)²]

- Phương pháp KNN:
  + Đơn giản, **không có quá trình học**. Không có mô hình được xây dựng.
  + Mất nhiều thời gian so với một mô hình.
  + KQ phụ thuộc vào việc chọn **khoảng cách**.
  + Có thể làm việc trên nhiều loại data.
  + Giải quyết các vấn đề **về phân loại, hồi quy.**
  + Độ phức tạp khá lớn.
  + Ứng dụng thành công trong lĩnh vực **tìm kiếm thông tin, nhận dạng, phân tích dữ liệu**.
  + Kết quả cho ra thường chính xác, nhưng chậm do phải duyệt qua data để tìm phần tử gần.
- Cách chuẩn hóa dữ liệu: (max, min thuộc [0;100], đưa các cột dữ liệu về một kiểu số)

$$new\_value = \frac{value - min}{max - min}$$

***Để xác định nhãn** của phần tử mới đến, xác định **khoảng cách** giữa phần tử mới đến tất cả các phần tử có trong dữ liệu và chọn ra **k** phần tử gần nhất.

III. Phương pháp đánh giá mô hình:

- Nếu dữ liệu có **1 tập học** và **1 tập kiểm tra** sẵn dùng:
  + Dùng data học để build mô hình.
  + Dùng tập kiểm tra để đánh giá hiệu quả của giải thuật.

(*) Dữ liệu kiểm tra và dữ liệu huấn luyện mô hình không được giao nhau.

- Nếu dữ liệu **không có 1 tập kiểm tra** sẵn:
  + Nghi thức **k-fold**: Chia tập data thành **k** phần (fold) bằng nhau, lặp lại **k** lần, mỗi lần sửa dụng **k - 1** folds để học và 1 fold để kiểm tra, sau đó tính trung bình của **k** lần kiểm tra.
    [] Nếu data có số phần tử lớn hơn 300, sử dụng nghi thức k-fold với k = 10.
    [] Nếu data có số phần tử nhỏ hơn, sử dụng nghi thức **leave - 1 - out** (k-fold với k = số phần tử).
  + Nghi thức **hold-out**: Lấy ngẫu nhiên **2/3** tập data để học và **1/3** tập dữ liệu còn lại dùng cho kiểm tra, có thể lặp lại quá bước này **k** lần rồi tính giá trị trung bình.

- Confusion matrix (C) cho k lớp:

Dự đoán =>	**1**	**…**	**k**
1			
…			
k			

C[i, j]: Số phần tử lớp **i (dòng)** được giải thuật dự đoán là lớp **j (cột).**
C[i, I]: Số phần tử phân lớp đúng.
Độ chính xác lớp **I**: C[i, i] / C[i,]
Độ chính xác tổng thể: $\sum C[i,i] \;/\; C$

- Nếu data không cân bằng, cần chọn chỉ số đánh giá phù hợp.
- Confusion matrix (C) cho 2 lớp (+/-)
  + Precision.
  + Recall.
  + Accuracy.
  + F1

Dự đoán =>	**dương**	**âm**
dương	TP	FN
âm	FP	TN

$$prec = \frac{tp}{tp + fp} \qquad acc = \frac{tp + tn}{tp + fn + tn + fp}$$

$$rec = \frac{tp}{tp + fn} \qquad F1 = \frac{2\,x\,prec\,x\,rec}{prec + rec}$$

  + **TP**: tổng số phần tử lớp **dương** được giải thuật dự đoán là lớp **dương**.
  + **FN**: tổng số phần tử lớp **dương** được giải thuật dự đoán là lớp **âm**.
  + **TN**: tổng số phần tử lớp **âm** được giải thuật dự đoán là lớp **âm**.
  + **FP**: tổng số phần tử lớp **âm** được giải thuật dự đoán là lớp **dương**.

Dự đoán =>	dương	âm
dương	10 (TP)	5 (FN)
âm	8 (FP)	22 (TN)

VD: Giả sử tập data có 40000 mẫu tin, trong đó có 8 mẫu tin thuộc lớp dương (+1) và 39992 mẫu tin thuộc lớp âm (-1), có 2 mô hình phân lớp M1 và M2 cho KQ tương ứng trong bảng 1, 2 như bên dưới. Tìm mô hình thích hợp để xử lý tập dữ liệu trên.

Ma trận confusion thu được từ mô hình M1 (trái) và M2 (phải):

Dự đoán =>	dương	âm
dương	1	7
âm	1	39991

Dự đoán =>	dương	âm
dương	8	0
âm	32	39960

- Sử dụng mô hình M2 vì giữa dự đoán đúng 1 / 7 và đúng cả 8 thì chọn đúng cả 8.

- Chỉ số đánh giá cho bài toán hồi quy: Đo lường mức độ sai số của các dự đoán. Càng thấp càng tốt. (MAE, MSE, RMSE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |p_i - r_i|$$

Trong đó: $\mathbf{p_i}$ là giá trị dự đoán đánh giá của item **i.**
$\mathbf{R_i}$ là giá trị thực tế của item **i**.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (p_i - r_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (p_i - r_i)^2}$$

IV. Giải thuật Bayes thơ ngây (Naive - Bayes)

- Phương pháp Bayes là hệ thống **ham học.**
- Là giải thuật học có giám sát (supervised learning) - xây dựng mô hình phân loại dựa trên dữ liệu tập học đã có nhãn (lớp).
- Dựa vào **các đặc trưng** đưa ra kết luận **nhãn** của đối tượng mới đến.
- Khi đưa ra một tập train, hệ thống **ngay lập tức** phân tích dữ liệu và **xây dựng mô hình.** Khi cần phân loại một đối tượng mới đến, hệ thống sử dụng mô hình đã xây dựng để xác định đối tượng mới.
- Có xu hướng phân loại nhanh hơn KNN (lười học).
- Cho kết quả tốt.
- Phân lớp không yêu cầu phải ước lượng một cách chính xác xác suất.
- Sử dụng trong phân loại text, spam, etc.
- Data có nhiều attribute dư thừa => Naive - Bayes k hiệu quả.
- Data liên tục có thể k theo phân phối chuẩn.
- Sử dụng kiến thức **Xác suất thống kê.**

Xác suất của A với điều kiện B xảy ra được định nghĩa như sau:

$$P(A/B) = \frac{P(AB)}{P(B)} = \frac{P(B/A)P(A)}{P(B)}$$

- Triển khai giải thuật:

B1. Huấn luyện mô hình: Tính xác suất của tất cả các trường hợp.

B2. Dự đoán: Xác định nhãn của đối tượng mới đến thông qua giá trị xác suất lớn nhất tính được.

VD1: Cho tập dữ liệu weather như sau:

Outlook	Temp	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

Bước 1: Huấn luyện mô hình / Học (Learning Phase)

Outlook			Temperature			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Bước 2: Dự báo / Dự đoán

Yêu cầu: Dự báo cho mẫu tin X = [Sunny, Cool, High, True] => Xác định nhãn.

=> Để xác định nhãn cho X (Yes / No) cần tính xác suất của lớp "yes" và xác suất của lớp "no". (Sử dụng CT xác suất có điều kiện)

=> Nhãn của X dựa trên xác suất lớn hơn giữa 2 lớp "yes" và "no".

$$P(Yes|X) = [\,P(Outlook = Sunny \mid Yes).\; P(Temperature = Cool \mid Yes).\; P(Humidity = High \mid Yes).\; P(Windy = True \mid Yes).\; P(Yes)\,] \,/\, P(X)$$

$$= \frac{\frac{2}{9} x \frac{3}{9} x \frac{3}{9} x \frac{3}{9} x \frac{9}{14}}{P(X)} = 0.0053 \,/\, P(X)$$

$$P(No|X) = [\,P(Outlook = Sunny \mid No).\; P(Temperature = Cool \mid No).\; P(Humidity = High \mid No).\; P(Windy = True \mid No).\; P(No)\,] \,/\, P(X)$$

$$= \frac{\frac{3}{5} x \frac{1}{5} x \frac{4}{5} x \frac{3}{5} x \frac{5}{14}}{P(X)} = 0.0206 \,/\, P(X)$$

Ta có: $P(Yes|X) + P(No|X) = 1 \Rightarrow P(X) = 0.0259.$

$\Rightarrow P(Yes|X) = 0.205 < P(No|X) = 0.795$

Vậy, nhãn cho mẫu tin X[Sunny, Cool, High, True] là "No".

VD2: Cho tập dữ liệu weather như sau:

Outlook	Temp	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rainy	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes
Rainy	71	91	True	No

Bước 1: Huấn luyện mô hình

Outlook			Temperature			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3		83	85		86	85	False	6	2	9	5
Overcast	4	0		70	80		96	90	True	3	3		
Rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
Sunny	2/9	3/5	μ	73	74.6	μ	79.1	86.2	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ	6.2	7.9	σ	10.2	9.7	True	3/9	3/5		
Rainy	3/9	2/5	σ^2	38	62.3	σ^2	104.4	94.7					

Bước 2: Dự đoán / Dự báo:

Yêu cầu: Dự báo cho mẫu tin X = [Sunny, 66, 90, True] => Xác định nhãn.

=> Để xác định nhãn cho X (Yes / No) cần tính xác suất của lớp "yes" và xác suất của lớp "no" dựa vào công thức phân phối xác suất.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

=> Nhãn của X dựa trên xác suất lớn hơn giữa 2 lớp "yes" và "no".

$$P(Yes|X) = [\,P(Outlook = Sunny \mid Yes).\; P(Temperature = Cool \mid Yes).\; ?\; P(Humidity = High \mid Yes).\; ?\; P(Windy = True \mid Yes).\; P(Yes)\,] \,/\, P(X)$$

$$= \frac{\frac{2}{9}\, x\, ?\, x\, ?\, x\, \frac{3}{9}\, x\, \frac{9}{14}}{P(X)} = ? \,/\, P(X)$$

$$P(No|X) = [\,P(Outlook = Sunny \mid No).\; P(Temperature = Cool \mid No).\; ?\; P(Humidity = High \mid No)\; ?.\; P(Windy = True \mid No).\; P(No)\,] \,/\, P(X)$$

$$= \frac{\frac{3}{5}\, x\, ?\, x\, ?\, x\, \frac{3}{5}\, x\, \frac{5}{14}}{P(X)} = ? \,/\, P(X)$$

Ở cột Temperature:

- Đối với nhãn "yes":

$$\mu = \frac{1}{9}(83 + 70 + 68 + 64 + 69 + 75 + 75 + 72 + 81) = 73$$

$$\sigma^2 = \frac{1}{9-1}[(83-73)^2 + (70-73)^2 + (68-73)^2 + (64-73)^2 + (69-73)^2 + (75-73)^2 + (75-73)^2 + (72-73)^2 + (81-73)^2] = 38$$

$$f(Temperature = 66 \mid Yes) = \frac{1}{\sqrt{2\pi}.\sqrt{38}} e^{\frac{-(66-73)^2}{2.38}} = 0.034$$

- Đối với nhãn "no":

$$\mu = \frac{1}{5}(85 + 90 + 70 + 95 + 91) = 74.6$$

$$\sigma^2 = \frac{1}{5-1}[(85-74.6)^2 + (90-74.6)^2 + (70-74.6)^2 + (95-74.6)^2 + (91-74.6)^2] = 62.3$$

$$f(Temperature = 66 \mid No) = \frac{1}{\sqrt{2\pi}.\sqrt{62{,}3}} e^{\frac{-(66-74.6)^2}{2.62{,}3}} = 0.028$$

Ở cột Humidity:

- Đối với nhãn "yes":

$$\mu = \frac{1}{9}(86 + 96 + 80 + 65 + 70 + 80 + 70 + 90 + 75) = 79.1$$

$$\sigma^2 = \frac{1}{9-1}[(86-79.1)^2 + (96-79.1)^2 + (80-79.1)^2 + (65-79.1)^2 + (70-79.1)^2 + (80-79.1)^2 + (70-79.1)^2 + (90-79.1)^2] + (75-79.1)^2 = 104.4$$

$$f(Humidity = 90 \mid Yes) = \frac{1}{\sqrt{2\pi}.\sqrt{104{,}4}} e^{\frac{-(90-79.1)^2}{2.104{,}4}} = 0.022$$

- Đối với nhãn "no":

$$f(Humidity = 90 \mid No) = \frac{1}{\sqrt{2\pi}.\sqrt{94{,}7}} e^{\frac{-(90-86.2)^2}{2.94{,}7}} = 0.038$$

Thế vào 2 công thức $P(Yes|X)$ và $P(No|X)$ ở trên:

=> $P(Yes|X) = [2/9 * 0{,}0340 * 0{,}0221 * 3/9 * 9/14] / P(X) = 0{,}000036 / P(X)$

$P(No|X) = [3/5 * 0{,}0291 * 0{,}0380 * 3/5 * 5/14] / P(X) = 0{,}000136 / P(X)$

Ta có, $P(Yes|X) + P(No|X) = 1$ => $P(X) = 0{,}000172$

=> $P(Yes|X) = 0{,}000036 / 0{,}000172 = 20{,}9\%$

$P(No|X) = 0{,}000136 / 0{,}000172 = 79.1\%$

Vậy, nhãn cho mẫu tin X = [Sunny, 66, 90, True] là "yes".

- Khi xác suất bằng 0, sử dụng *ước lượng Laplace.* (Laplace estimator)
  + Xác suất không bao giờ có giá trị 0.
  + Cộng thêm cho tử một giá trị là $\boldsymbol{p_i\mu}$ **và mẫu số giá trị** $\boldsymbol{\mu}$ **để tính xác suất.** $\boldsymbol{\mu}$ **hằng số dương và** $\boldsymbol{p_i}$ **là hệ số dương sao cho tổng các** $\boldsymbol{p_i = 1}$ $\boldsymbol{(i = 1 \dots n)}$**.**

VD: Cho tập dữ liệu sau:

Outlook	Temp	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

Dự đoán nhãn của X = [Overcast, Cool, High, True].

Từ bảng huấn luyện mô hình đã có, ta được kết quả:

$$P(Yes|X) = [\, P(Outlook = Overcast \mid Yes).\ P(Temperature = Cool \mid Yes).\ P(Humidity = High \mid Yes).\ P(Windy = True \mid Yes).\ P(Yes)\,] / P(X)$$

$$= \frac{\frac{4}{9} x \frac{3}{9} x \frac{3}{9} x \frac{3}{9} x \frac{9}{14}}{P(X)} = 0.0105 / P(X)$$

$$P(No|X) = [\,P(Outlook = Overcast|\,No).$$
$$P(Temperature = Cool\,|\,No).$$
$$P(Humidity = High\,|\,No).$$
$$P(Windy = True\,|\,No).$$
$$P(No)\,]\,/\,P(X)$$
$$= \frac{\frac{0}{5}\,x\,\frac{1}{5}\,x\,\frac{4}{5}\,x\,\frac{3}{5}\,x\,\frac{5}{14}}{P(X)} = 0$$

=> Xảy ra vấn đề => Sử dụng *Laplace estimator*.

Thuộc tính *outlook* cho lớp "no" => $\mathbf{p_1 = p_2 = p_3 = 1/3;\ \mu = 1}$.

Sunny:

$$\frac{3 + \frac{\mu}{3}}{5 + \mu} = \frac{3 + \frac{1}{3}}{5 + 1} = \frac{10}{18}$$

Overcast:

$$\frac{0 + \frac{\mu}{3}}{5 + \mu} = \frac{0 + \frac{1}{3}}{5 + 1} = \frac{1}{18}$$

Rainy:

$$\frac{2 + \frac{\mu}{3}}{5 + \mu} = \frac{2 + \frac{1}{3}}{5 + 1} = \frac{7}{18}$$

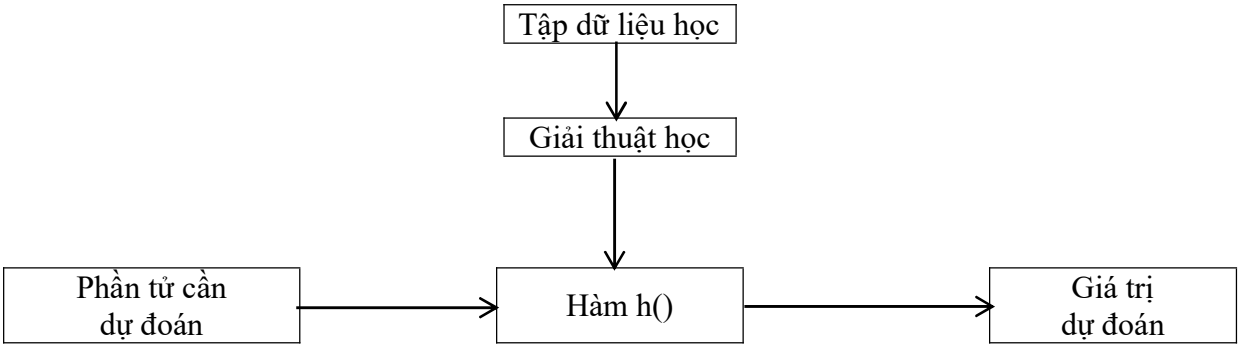
$$=> P(No|X) = [\,P(Outlook = Overcast|\,No).$$
$$P(Temperature = Cool\,|\,No).$$
$$P(Humidity = High\,|\,No).$$
$$P(Windy = True\,|\,No).$$
$$P(No)\,]\,/\,P(X)$$
$$= \frac{\frac{1}{18}\,x\,\frac{1}{5}\,x\,\frac{4}{5}\,x\,\frac{3}{5}\,x\,\frac{5}{14}}{P(X)} = 0{,}0019$$

Vậy, nhãn của **X = [Overcast, Cool, High, True]** là "yes".

- Giá trị thuộc tính bị nhiễu:
  + Học: Bỏ qua dữ liệu nhiễu.
  + Phân lớp: Bỏ qua các thuộc tính nhiễu.

(*) Khi thuộc tính của dữ liệu mới đến bị thiếu, bỏ qua luôn thuộc tính đó. Nói cách khác, xác suất chỗ đó được xem như bằng 1.

V. Phương pháp học cây quyết định: (Decision Tree)



- Cây quyết định là giải thuật học:
  + Kết quả sinh ra dễ diễn dịch **(if … then …)**.
  + Khá đơn giản, nhanh, hiệu quả được sử dụng nhiều.
  + Giải quyết các vấn đề của phân loại, hồi quy.
  + Làm việc cho **dữ liệu số và kiểu liệt kê.**
  + Thành công trong hầu hết các lĩnh vực về phân tích dữ liệu, phân loại text, spam, phân loại gien, etc.
- 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 data 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.
- Xây dựng **cây Top-down**:
  + Bắt đầu **từ nút gốc**, **tất cả các data học ở nút gốc**.
  + Nếu data tại 1 nút có cùng lớp => Nút lá. Nếu dữ liệu ở nút chứa **các phần tử có lớp rất khác nhau** thì **phân hoạch dữ liệu một cách đệ quy** bằng việc chọn 1 thuộc tính để thực hiện phân hoạch tốt nhất có thể.
- Chọn thuộc tính phân hoạch:
  + Tại mỗi nút, các thuộc tính được đánh giá dựa trên phân tách dữ liệu học **tốt nhất** có thể.
  + **Độ lợi thông tin** (chọn thuộc tính có chỉ số lớn) - information gain (ID3/C4.5 Quinlan).
  + **Chỉ số gini** (chọn thuộc tính có chỉ số nhỏ) - gini index (CART - Breiman).

- Nếu dữ liệu T có n lớp, chỉ số gini(T) được định nghĩa như sau:

$$gini(T) = 1 - \sum_{j=1}^{n} (p_j)^2$$

p_j là xác suất của lớp j trong T.

- gini(T) là nhỏ nhất nếu những lớp trong T bị lệch.
- Sau khi phân hoạch T thành 2 tập con T1 & T2 với kích thước N1 & N2, chỉ số gini:

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

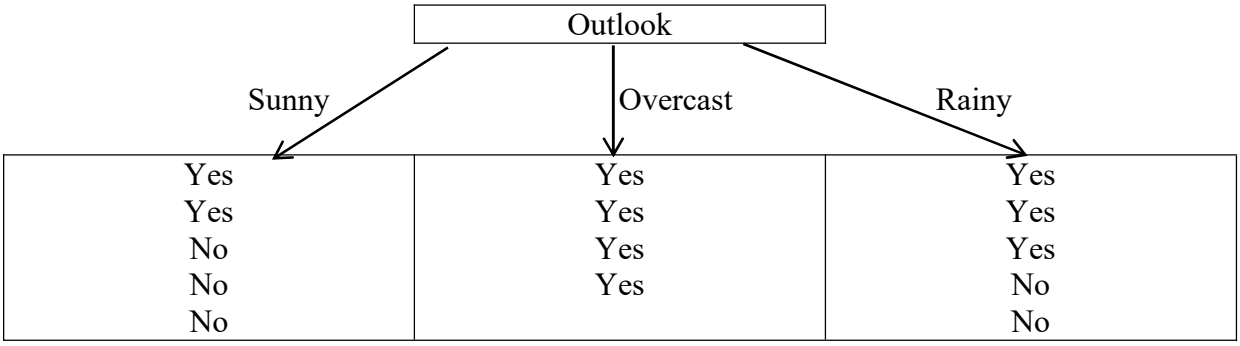
- Thuộc tính có **$gini_{split}$(T) nhỏ nhất** được chọn để phân hoạch.

VD: Cho tập dữ liệu (kinh điển) sau:

Outlook	Temp	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

Xây dựng cây quyết định với chỉ số gini từ tập dữ liệu trên.

*Tính gini cho thuộc tính Outlook:



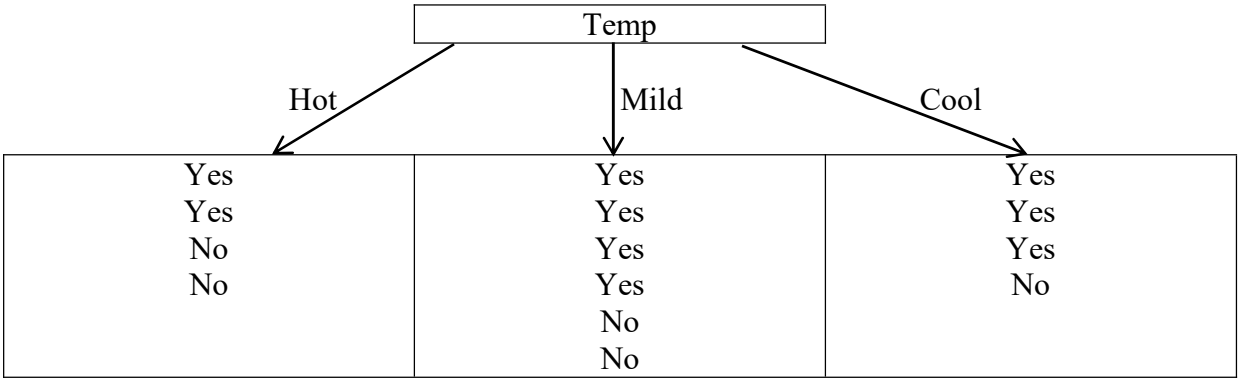
Gini(Outlook=Sunny) = $1 - [(\frac{2}{5})^2 + (\frac{3}{5})^2] = 1 - 0.16 - 0.36 = 0.48.$

Gini(Outlook=Overcast) = $1 - [(\frac{4}{4})^2 + (\frac{0}{4})^2] = 0$

Gini(Outlook=Rain) = $1 - [(\frac{3}{5})^2 + (\frac{2}{5})^2] = 1 - 0.36 - 0.16 = 0.48$

=> Gini(Outlook) = $\frac{5}{14} * 0.48 + \frac{4}{14} * 0 + \frac{5}{14} * 0.48 = 0.342$

*Tính gini cho thuộc tính Temperature:



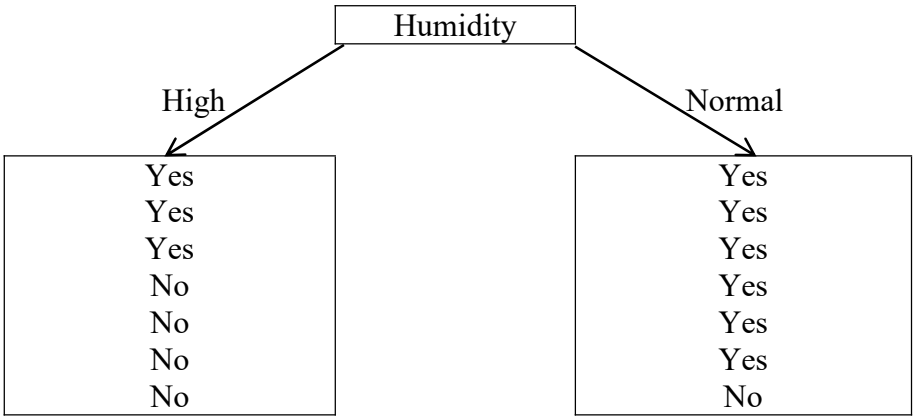
Gini(Temp=Hot) $= 1 - [(\frac{2}{4})^2 + (\frac{2}{4})^2] = 1 - 0.25 - 0.25 = 0.5$

Gini(Temp=Mild) $= 1 - [(\frac{4}{6})^2 + (\frac{2}{6})^2] = 1 - 0.444 - 0.111 = 0.445$

Gini(Temp=Cool) $= 1 - [(\frac{3}{4})^2 + (\frac{1}{4})^2] = 1 - 0.5625 - 0.0625 = 0.375$

=>Gini(Temp) $= 0.5 * \frac{4}{14} + 0.445 * \frac{6}{14} + 0.375 * \frac{4}{14} = 0.439$

*Tính gini cho thuộc tính Humidity:



Gini(Humidity=

B. Bài tập:

BT1: Cho tập dữ liệu như sau:

age	income	student	credit_rating	Class: buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

- Dự đoán nhãn của phần tử X1 = (youth, medium, yes, fair) bằng KNN với k = 3.
- Dự đoán nhãn của phần tử X2 = (senior, high, no, excellent) bằng Naive_Bayes.
- Dự đoán nhãn của phần tử X3 = (middle_aged, ?, yes, fair) bằng Naive_Bayes.

BT2: Cho tập dữ liệu như sau:

Outlook	Temp	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
Sunny	High	Cool	True	No

Dự đoán nhãn của phần tử X = {Rain, Hot, High, False} bằng Navie_Bayes và KNN với k = 3.

BT3: Cho tập dữ liệu như sau:
(Tập data được trích từ file "iris_data.csv")

sepalLength	sepalWidth	petalLength	petalWidth	nhan
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
6.3	3.3	6	2.5	Iris-virginica
4.7	3.2	1.3	0.2	Iris-setosa
6.2	2.2	4.5	1.5	Iris-versicolor
5.9	3	5.1	1.8	Iris-virginica
7	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
5	3	1.6	0.2	Iris-setosa
4.3	3	1.1	0.1	Iris-setosa
5.8	2.7	5.1	1.9	Iris-virginica
5.6	2.8	4.9	2	Iris-virginica
5.5	2.4	3.8	1.1	Iris-versicolor
5.1	3.8	1.9	0.4	Iris-setosa
6.5	3	5.2	2	Iris-virginica

- Dự đoán nhãn X1 = [6, 3, 4.8, 1.8] bằng KNN với k = 3.
- Dự đoán nhãn X2 = [5, 2.3, 3.3, 1] bằng Naive_Bayes.

BT4: Ước lượng Laplace cho trường hợp sau: $\boldsymbol{p_i, \mu} = ?$

	A	B	C
T1	2/8	2/10	5/13
T2	2/8	1/10	3/13
T3	1/8	2/10	0/13
T4	3/8	5/10	5/13

BT5: Cho tập dữ liệu sau:

age	income	student	credit_rating	Class: buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

Xây dựng cây quyết định từ tập dữ liệu trên.

BT6: