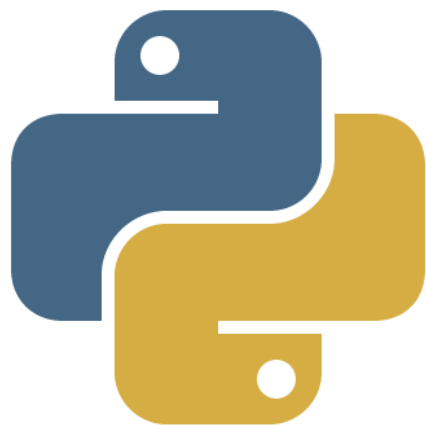




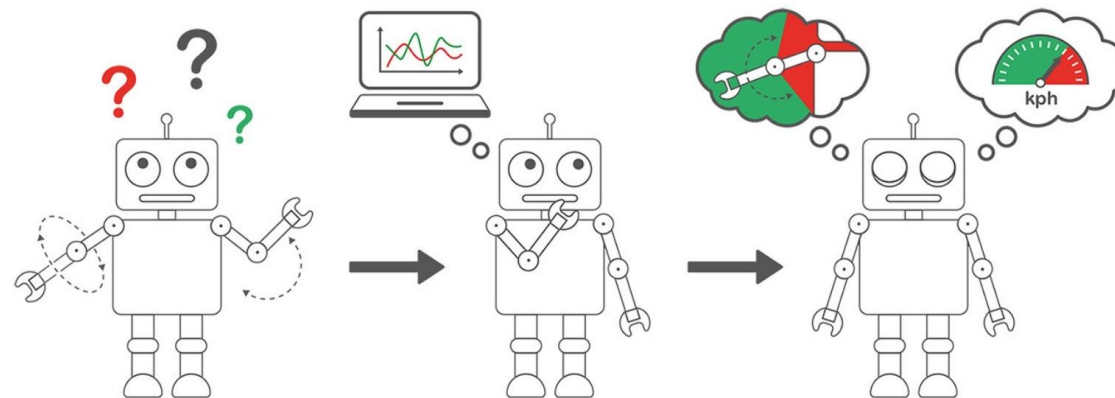
MAGIC CODE INSTITUTE



python

Phần 1

GIỚI THIỆU VỀ MACHINE LEARNING



ML ALGORITHMS

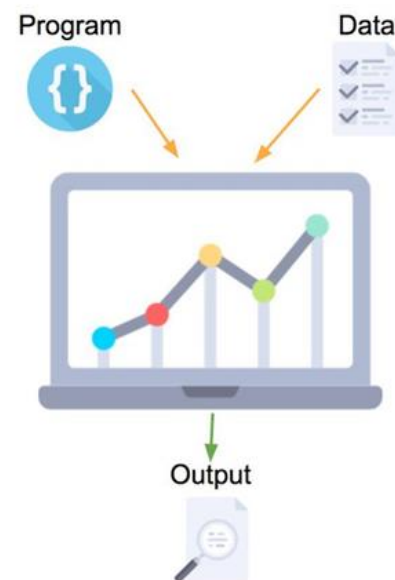
TECHGRABYTE

MACHINE LEARNING LÀ GÌ?

Machine Learning là ngành khoa học nghiên cứu cách để máy tính học mà không cần phải lập trình tường minh.

- Nhận diện mặt người trong ảnh (Auto-Tagging, Face ID).
- Công cụ tìm kiếm (Google, Bing)
- Trợ lý ảo (Siri, Cortana, Alexa)
- Lọc các email spam (Gmail, Outlook)
- Gợi ý phim (Netflix)

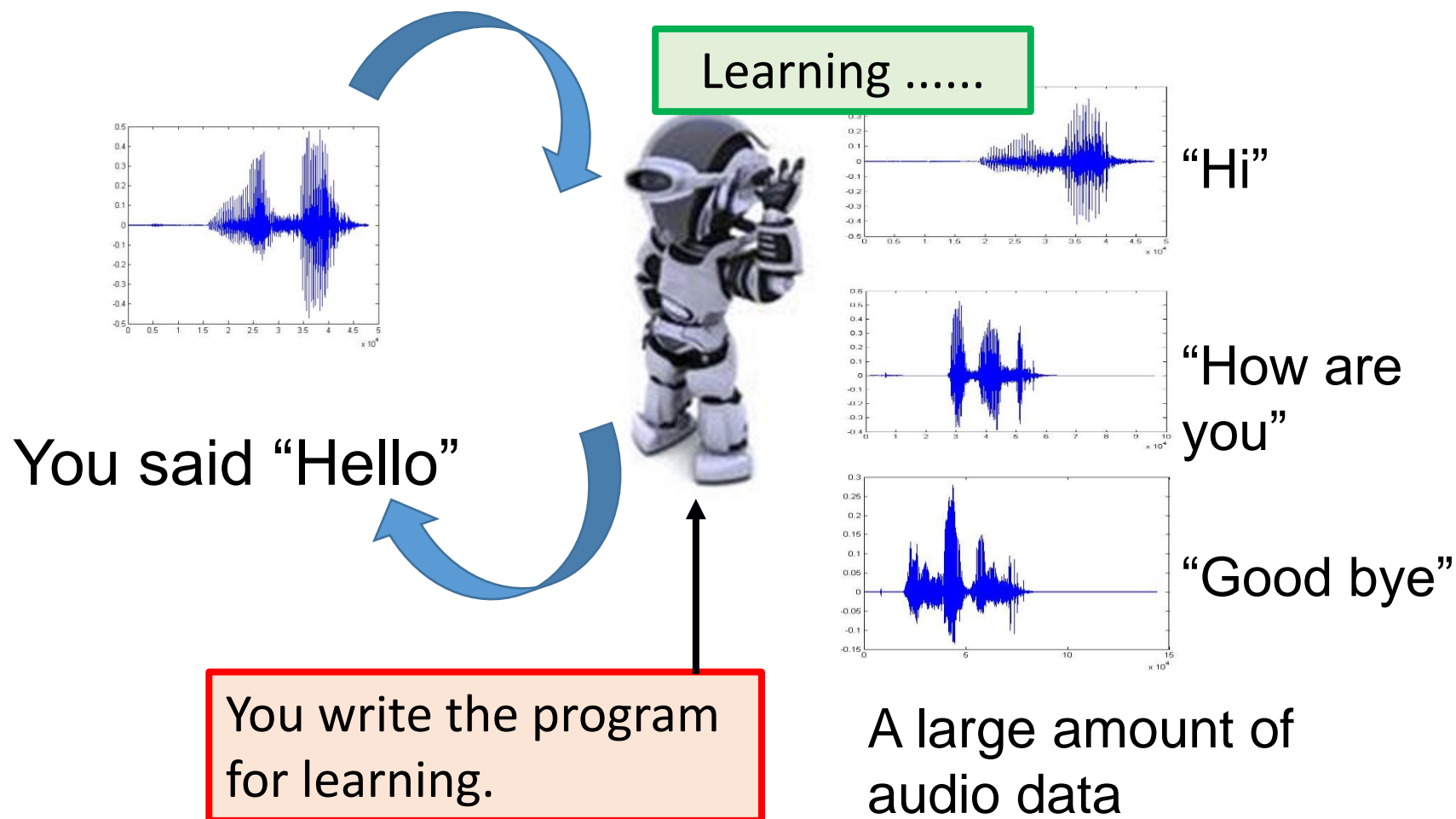
Traditional Programming



Machine Learning

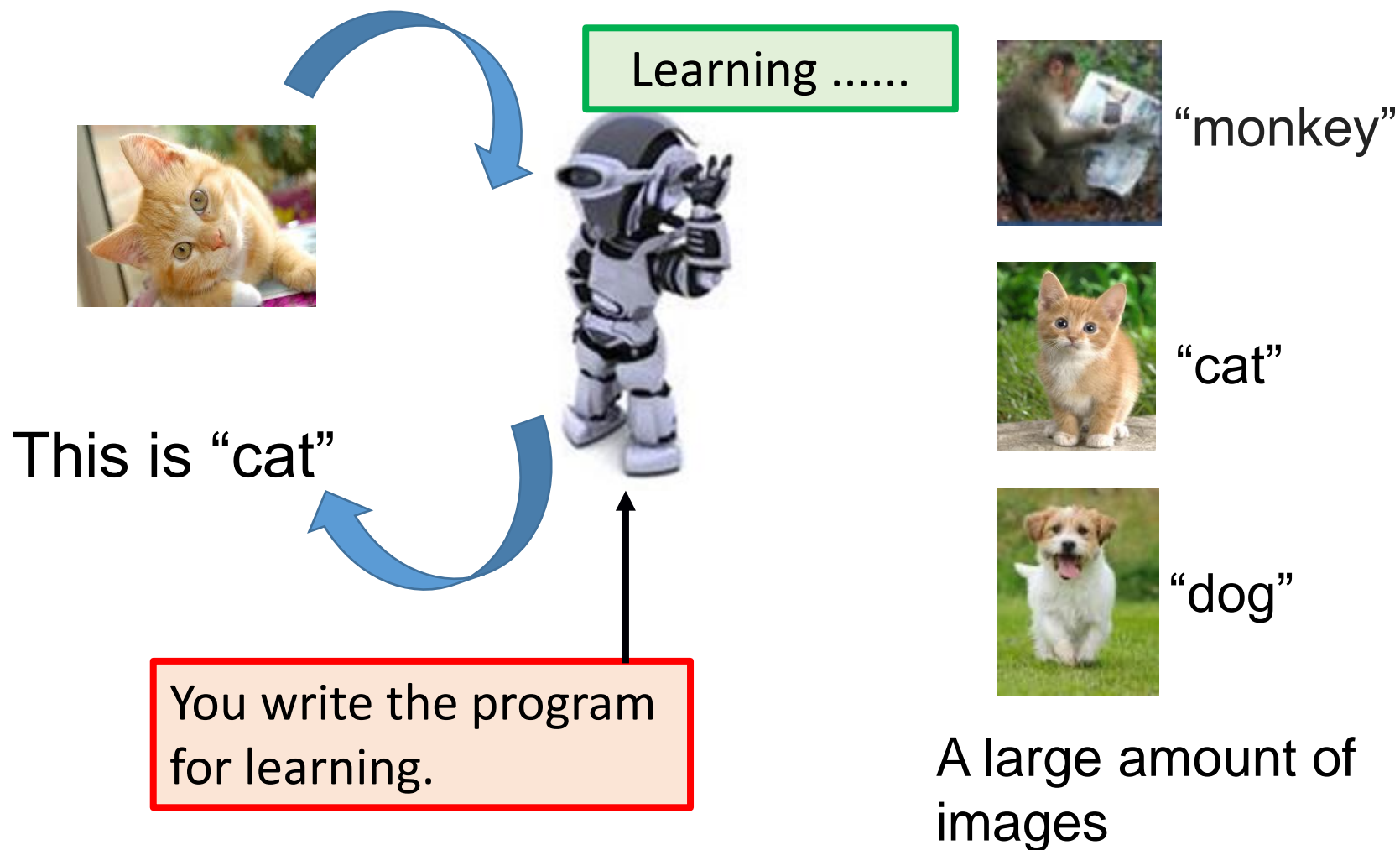


MACHINE LEARNING LÀ GÌ?

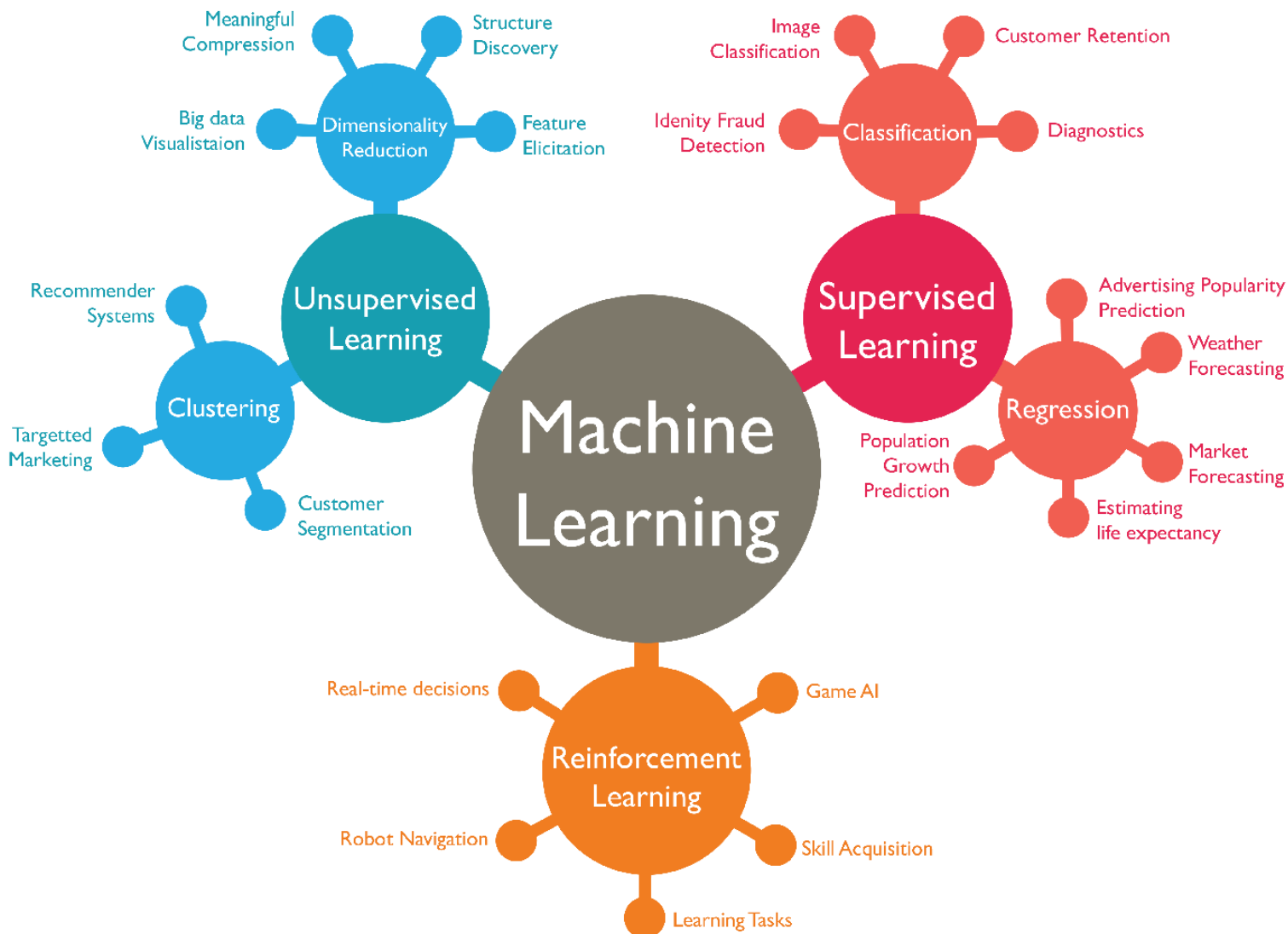


1

MACHINE LEARNING LÀ GÌ?



ỨNG DỤNG CỦA MACHINE LEARNING

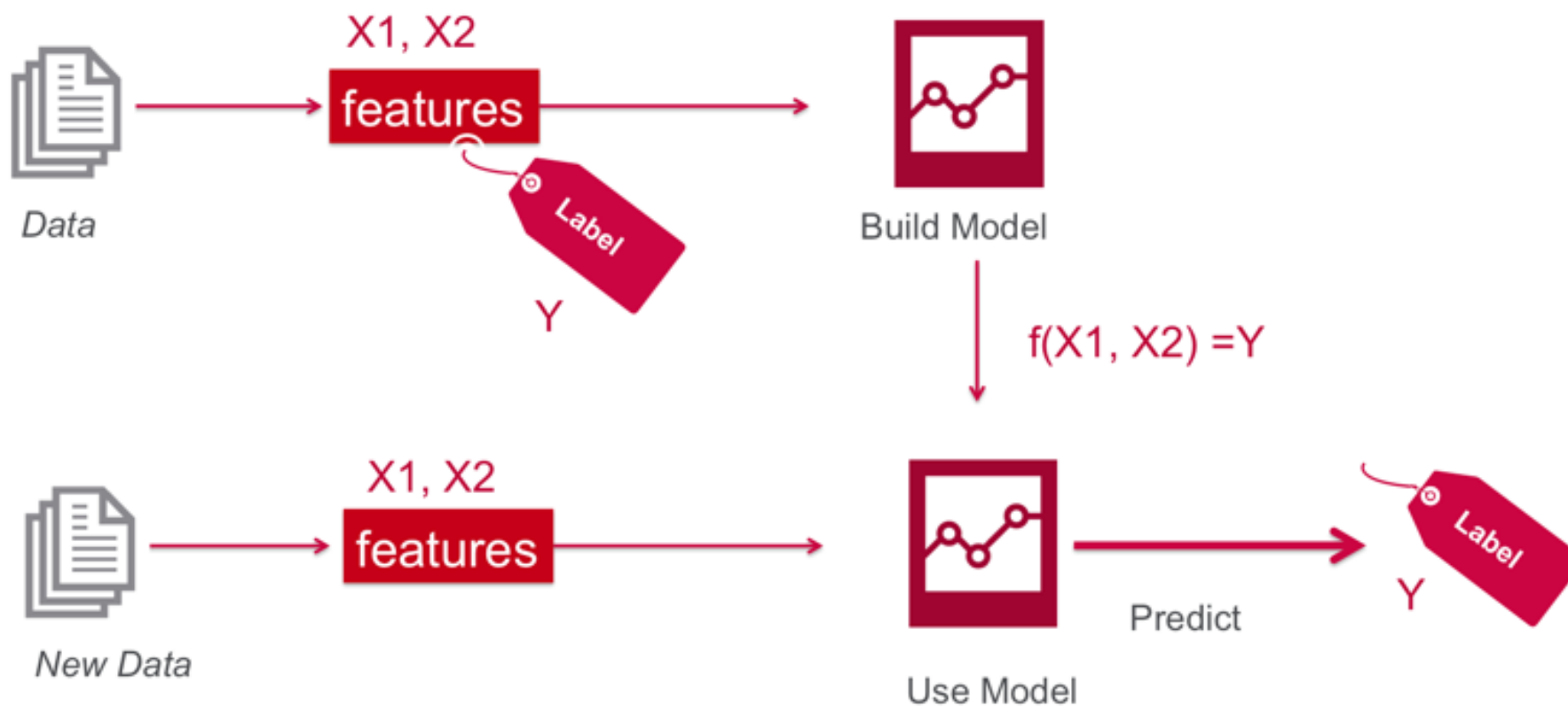


CÁC LOẠI HỌC MÁY

Học có giám sát (supervised learning)

Dữ liệu học là các cặp quan sát (sample) và nhãn (label), thuật toán máy xây dựng ra một mô hình (model) bằng dữ liệu học, mô hình đó sẽ giúp xác định nhãn của các quan sát mới.

(Việc xây dựng mô hình ở đây là tìm trọng số của một kiến trúc cho trước)

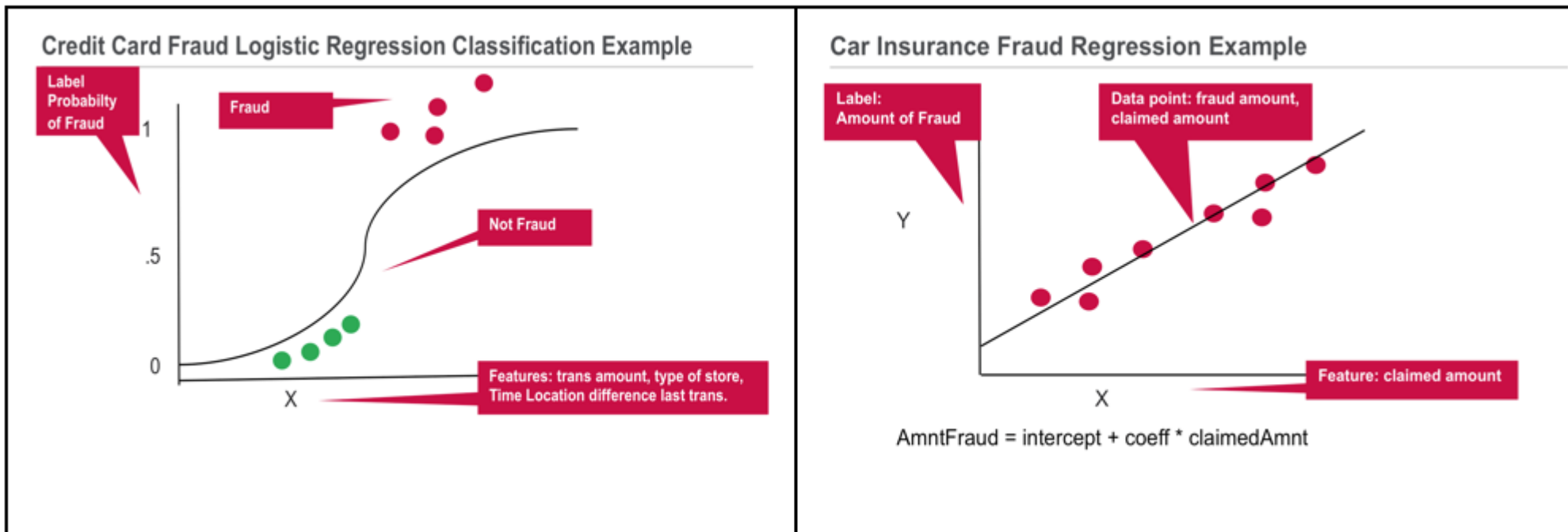


CÁC LOẠI HỌC MÁY

Học có giám sát (supervised learning)

Phân lớp: từ mỗi đầu vào (sample), xác định một lớp (class) trong một tập rời rạc cho trước

Hồi quy: từ mỗi đầu vào (sample), xác định một giá trị trong miền giá trị liên tục









CÁC LOẠI HỌC MÁY

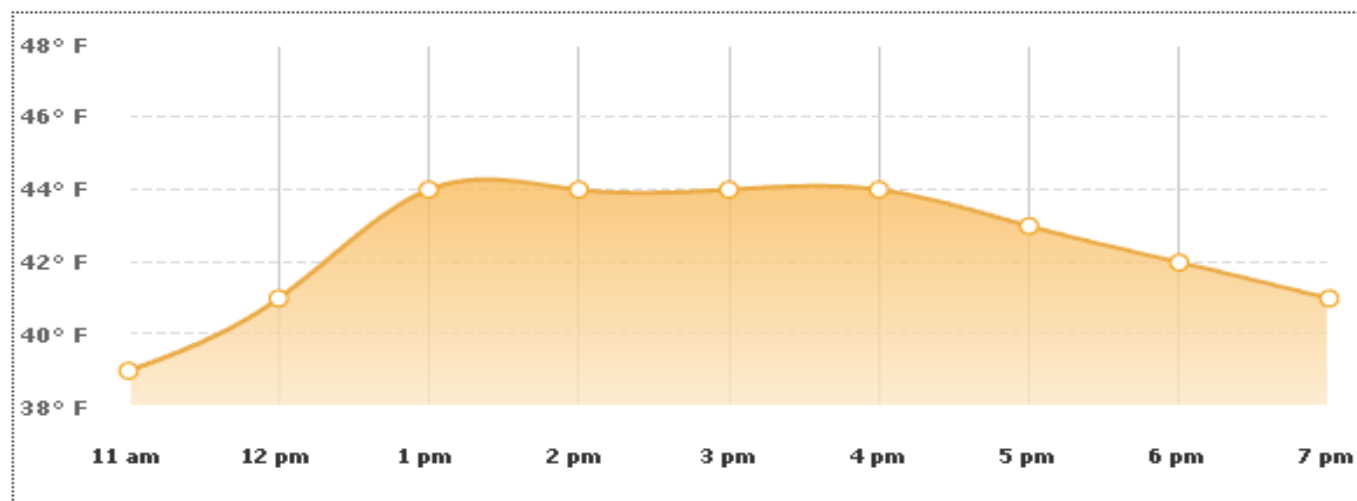
Học có giám sát (supervised learning)

Features?

Labels?

Classification/Regression?

11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm
							
39° F	41° F	44° F	44° F	44° F	44° F	43° F	42° F
Precip: 10%	Precip: 10%	Precip: 10%	Precip: 10%	Precip: 10%	Precip: 10%	Precip: 10%	Precip: 0%



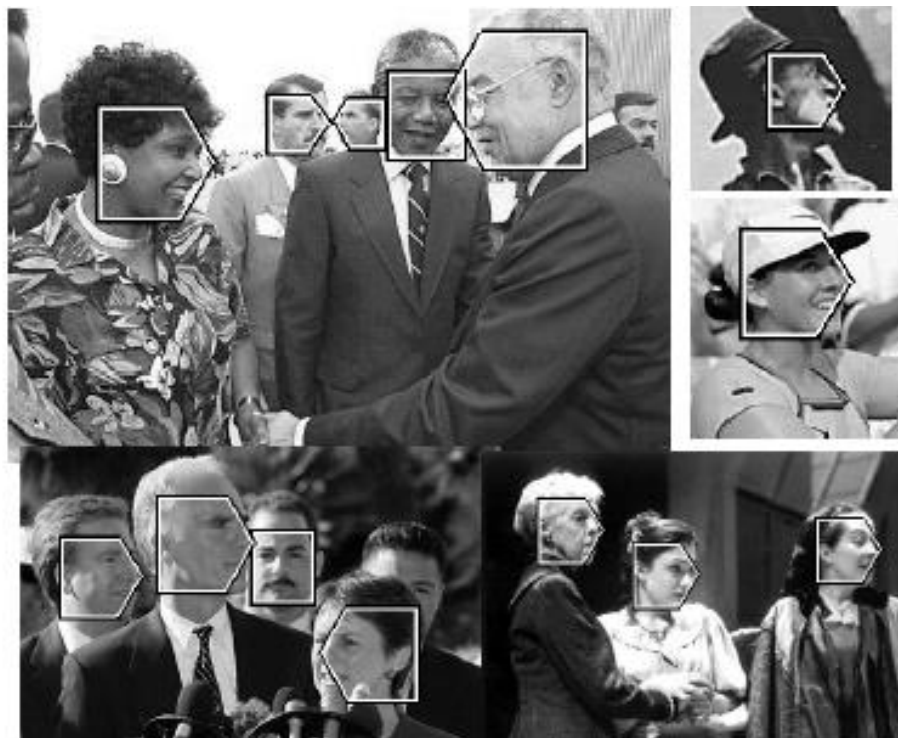
Temperature/Weather prediction



CÁC LOẠI HỌC MÁY

Học có giám sát (supervised learning)

Features? Labels? Classification/Regression?



Face Detection



CÁC LOẠI HỌC MÁY

Học có giám sát (supervised learning)

Features?

Labels?

Classification/Regression?

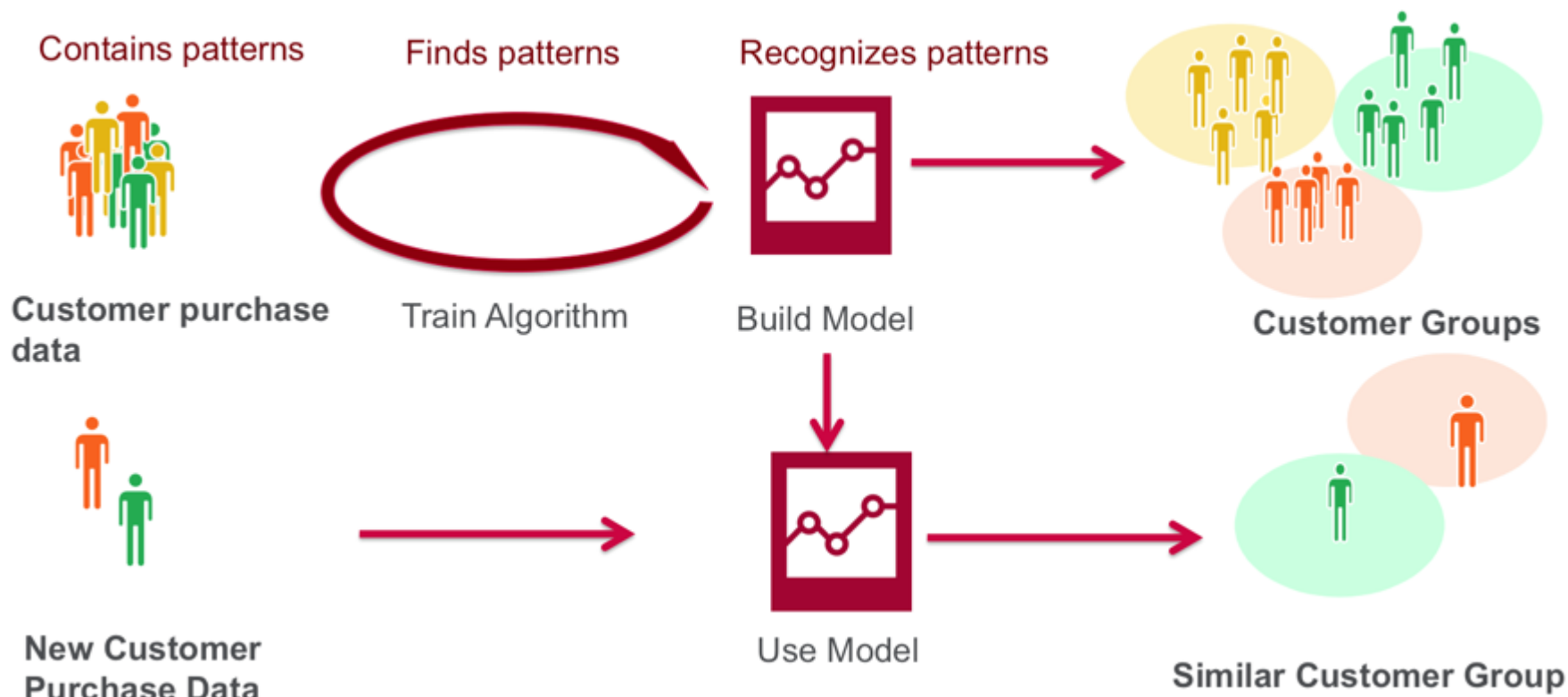


Robotic Control

CÁC LOẠI HỌC MÁY

Học không giám sát (supervised learning)

Đầu vào là tập các quan sát (sample), thuật toán tìm ra các cấu trúc ẩn hay đặc trưng quan trọng của tập các quan sát đó.



CÁC LOẠI HỌC MÁY

Học không giám sát (supervised learning)

Group similar things e.g. images



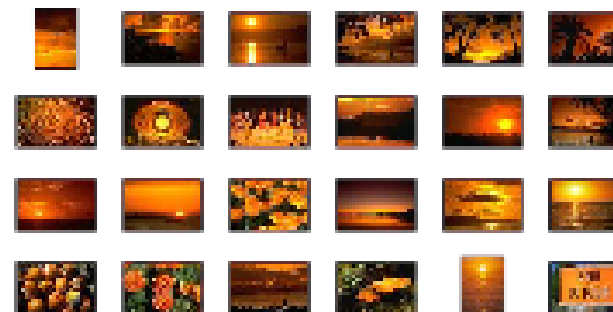
C_1



C_2



C_3



C_4



C_5

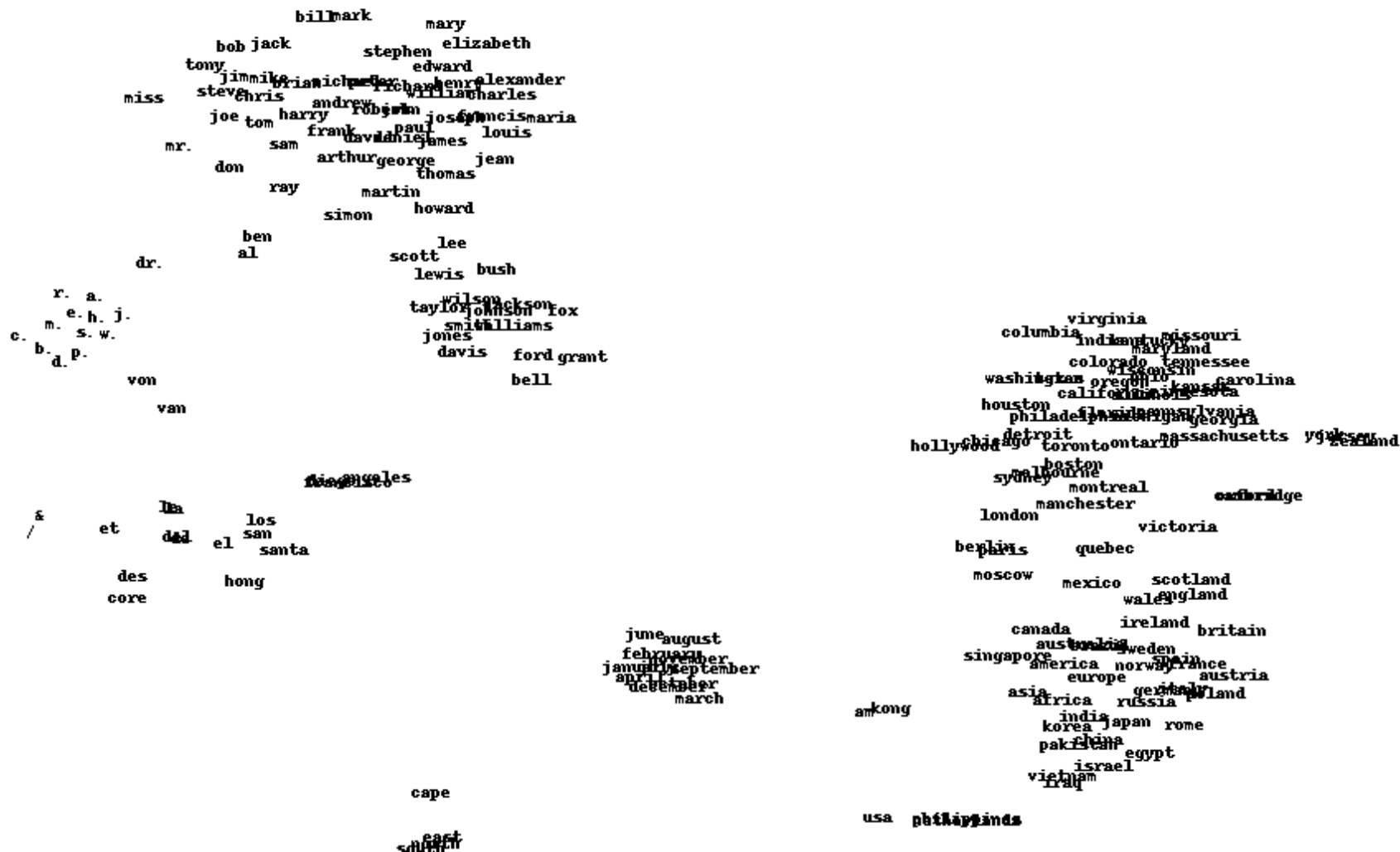
CÁC LOẠI HỌC MÁY

Học không giám sát (supervised learning)

The screenshot shows the Clusty search engine interface. The search bar contains the word "race". The results are organized into clusters. The "Human" cluster is selected, showing 8 documents. The documents listed are:

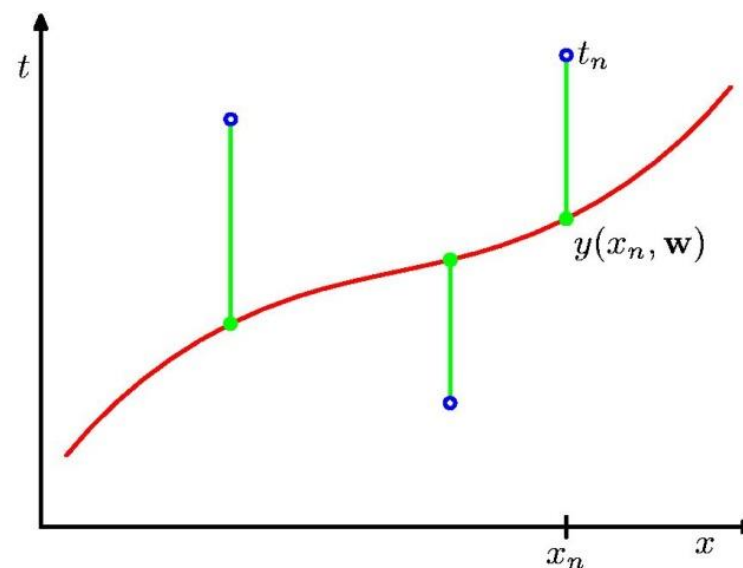
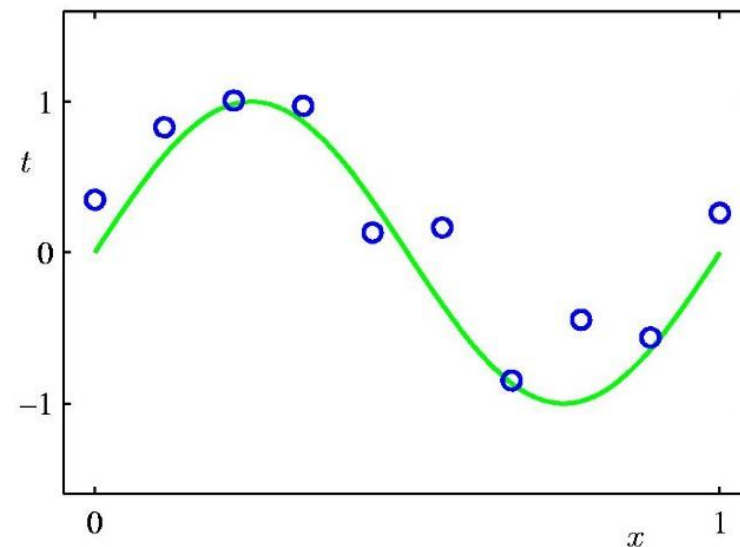
- Race (classification of human beings) - Wikipedia, the free ...**
The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of categories based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identity by culture and over time, and are often controversial for scientific as well as social and political reasons. [en.wikipedia.org/wiki/Race_\(classification_of_human_beings\)](https://en.wikipedia.org/wiki/Race_(classification_of_human_beings)) - [cache] - Live, Ask
- Race - Wikipedia, the free encyclopedia**
General. **Racing** competitions The **Race** (yachting **race**), or La course du millénaire, a no-rules round-the-world sail of **human beings** **Race** and ethnicity in the United States Census, official definitions of "**race**" used by the US Census. Historical definitions of **race**; **Race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** Literature · Video games
en.wikipedia.org/wiki/Race - [cache] - Live, Ask
- Publications | Human Rights Watch**
The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers ...
www.hrw.org/backgroundunder/usa/race - [cache] - Ask
- Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...**
Amazon.com: **Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books ...** From Publishers Weekly
www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340861 - [cache] - Live
- AAPA Statement on Biological Aspects of Race**
AAPA Statement on Biological Aspects of **Race** ... Published in the American Journal of Physical Anthropology, vol. 100, 1999, pp. 1-10.
www.physanth.org/positions/race.html - [cache] - Ask
- race: Definition from Answers.com**
race n. A local geographic or global **human** population distinguished as a more or less distinct group by genetically determined characteristics.
www.answers.com/topic/race-1 - [cache] - Live

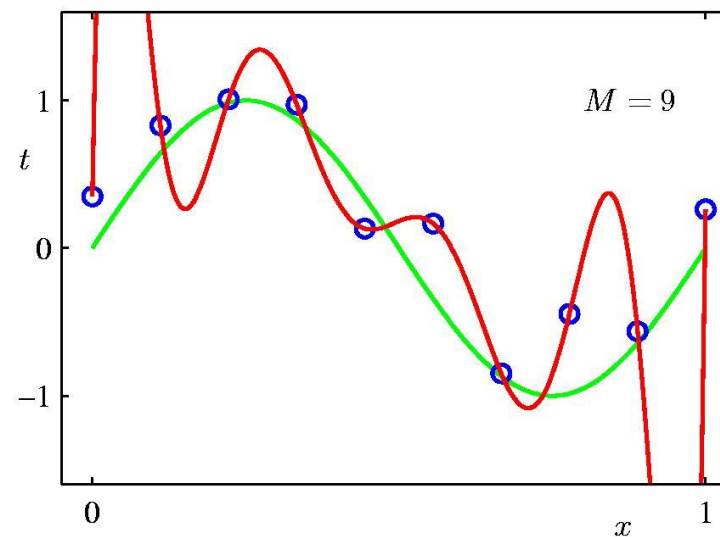
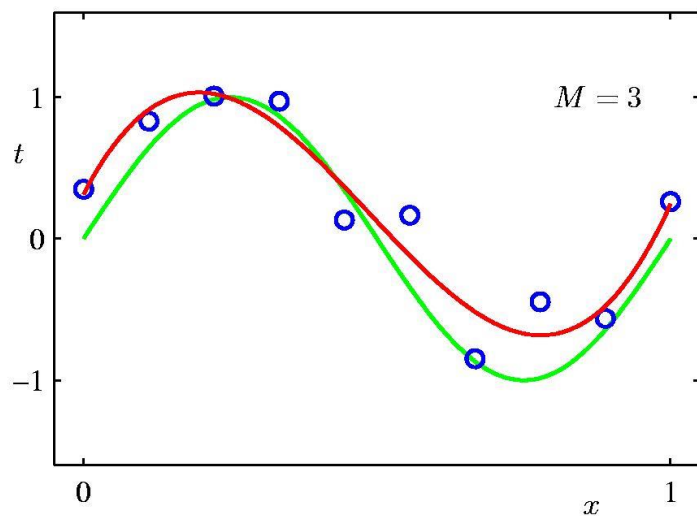
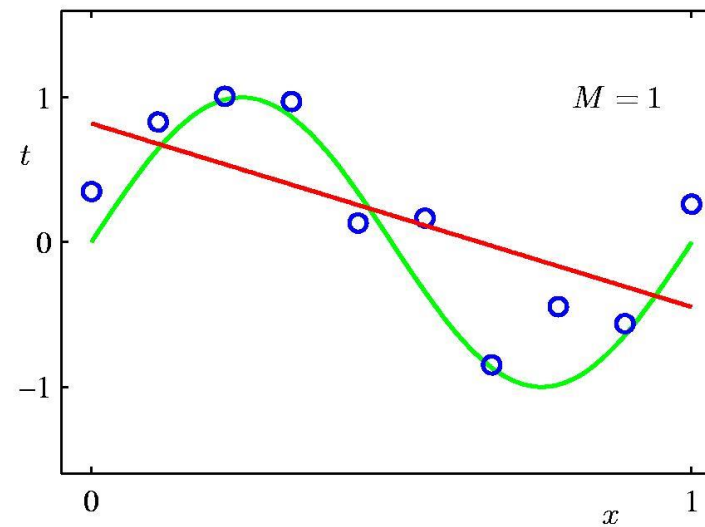
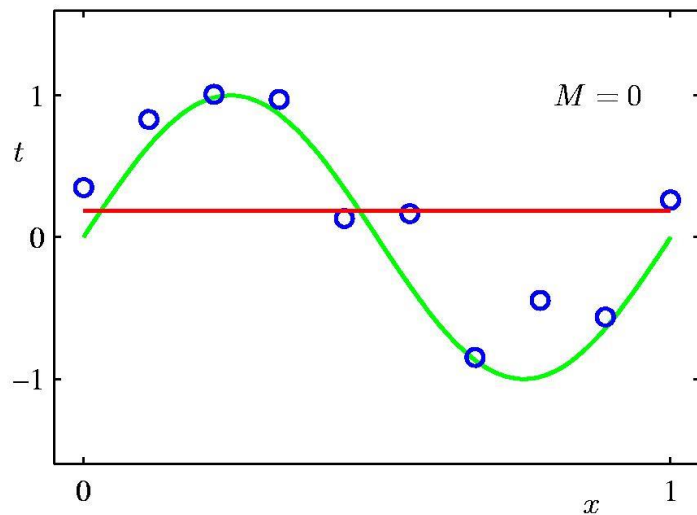
Học không giám sát (unsupervised learning)

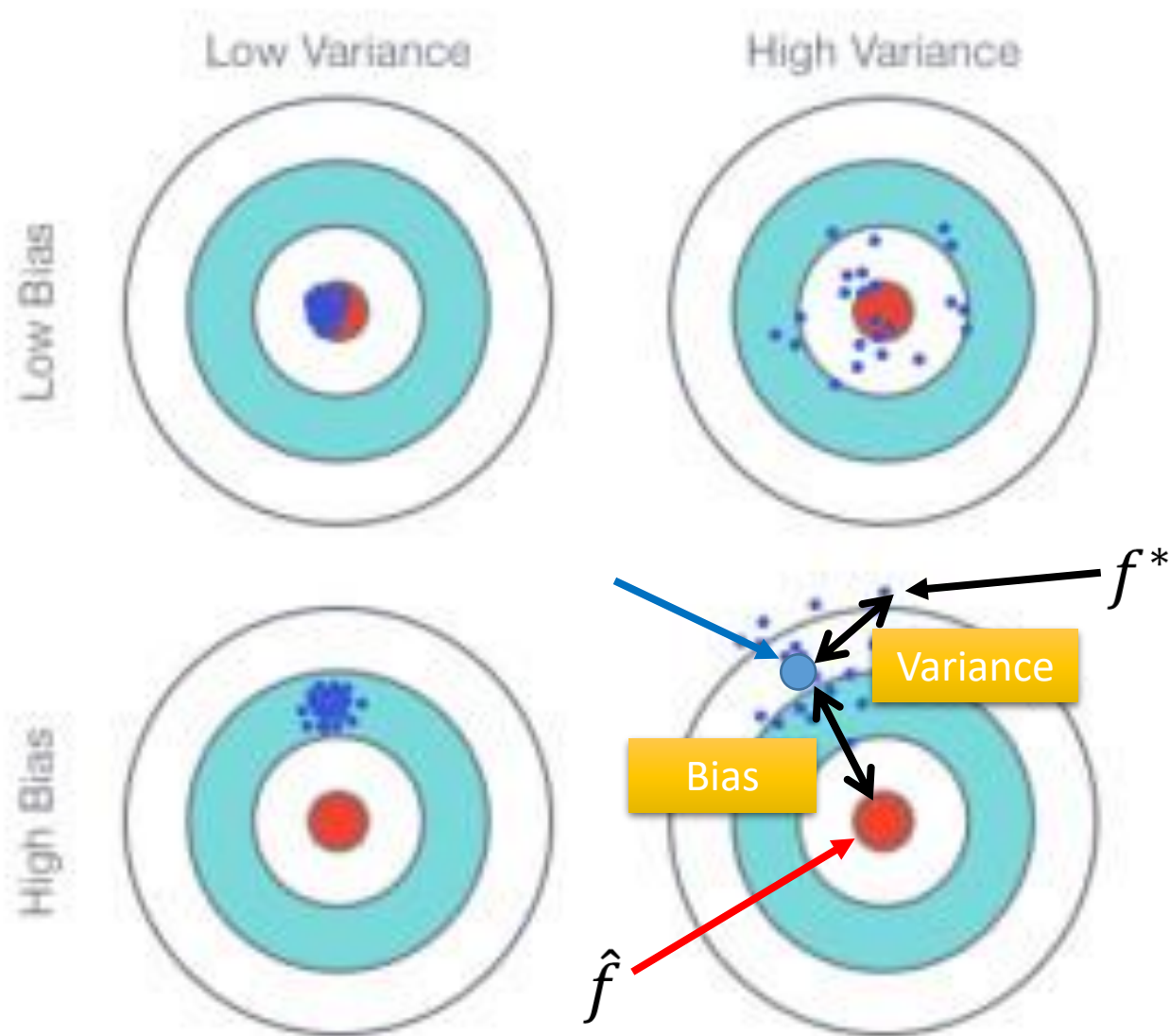


VÍ DỤ

- Đường màu xanh lá cây là hàm đúng
- Sử dụng một hàm mất mát để tính tổng khoảng cách từ các điểm đến đường dự đoán được (đường màu đỏ).



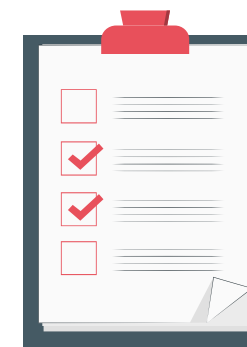
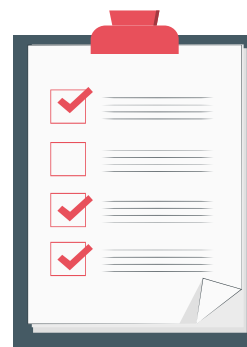
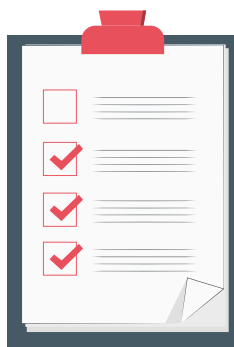




Phần 2 | CREDIT SCORING








VẤN ĐỀ: Làm thế nào để đánh giá năng lực trả nợ của một người muốn vay tín dụng?



CÁC PHƯƠNG PHÁP ĐÁNH GIÁ RỦI RO TÍN DỤNG

Phương pháp chuyên gia

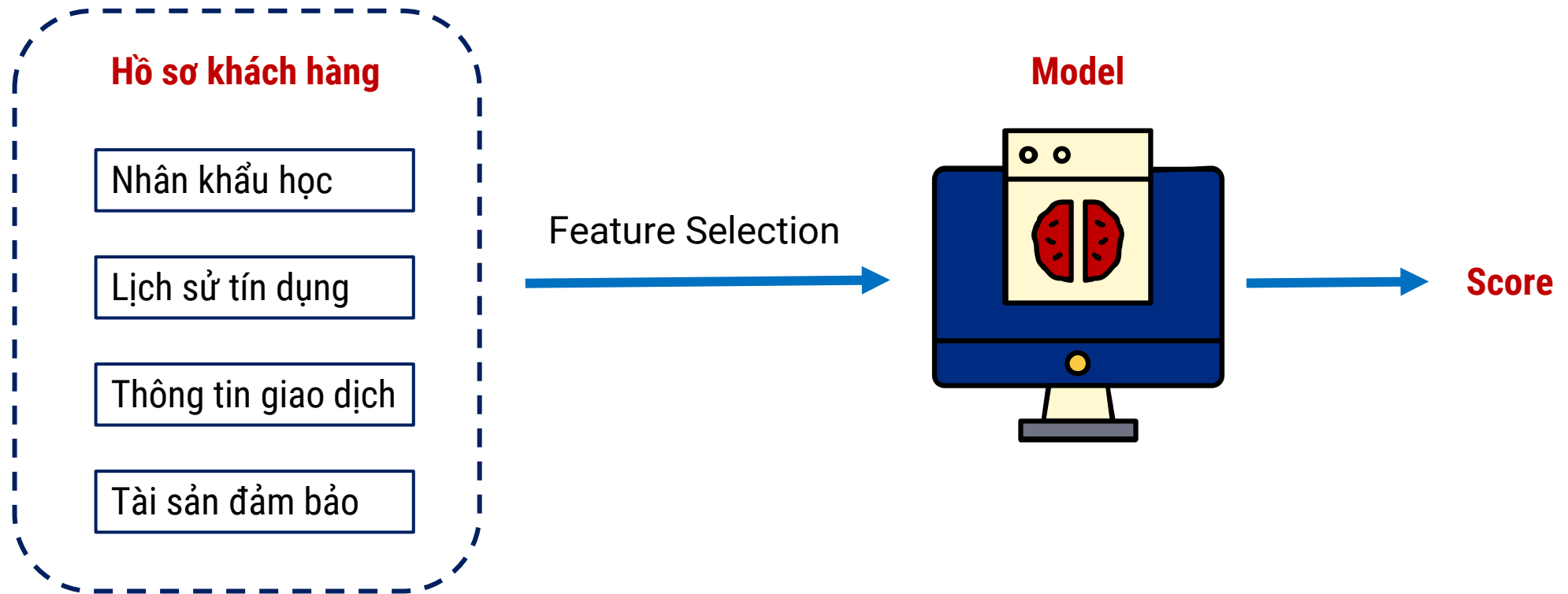
Dựa trên ý kiến thẩm định của các chuyên gia dựa trên các thông tin của khách hàng.

-  **Đặc điểm của chủ thể vay (Character)**
Thẩm định danh tiếng, tính trung thực của người vay vốn.
-  **Vốn (capital)**
Thẩm định sự chênh lệch giữa tài sản và nguồn chi của người vay
-  **Tài sản đảm bảo (Collateral)**
Tài sản thế chấp có thể dùng để trả nợ của người vay
-  **Khả năng trả nợ**
Liên quan đến khả năng tài chính: nghề nghiệp, mức thu nhập, số người phụ thuộc, ...
-  **Điều kiện**
Đánh giá điều kiện người vay với điều kiện thị trường

CÁC PHƯƠNG PHÁP ĐÁNH GIÁ RỦI RO TÍN DỤNG

Phương pháp mô hình

Xây dựng mô hình Credit ScoreCard nhằm mục đích ước lượng giá trị xác suất xảy ra một sự kiện đối với một khách hàng như vỡ nợ, phá sản hoặc trả nợ chậm.



FEATURE SELECTION

Weight of Evidence - WoE

Biểu diễn mối quan hệ giữa các biến đầu vào và biến kết quả, thường được thể hiện thông qua sự chênh lệch về giá trị tốt và xấu của biến kết quả trong biến đầu vào đó.

Cách tính

- Biến liên tục: chia thành các bins, mỗi bin có số lượng quan sát gần bằng nhau.
- Biến phân loại: mỗi class là một bin hoặc một số class có số lượng ít vào một bin.

Công thức

$$WoE_{ij} = \ln \frac{P(X_j \in B_i | Y = 1)}{P(X_j \in B_i | Y = 0)}$$

trong đó X_j là tiêu chí thứ j , B_i là bin thứ i , Y là kết quả của hồ sơ.



FEATURE SELECTION

Information Value - IV

Chỉ số giá trị thông tin, nhằm đánh giá sức ảnh hưởng của tiêu chí đó đến kết quả, kiểm tra xem tiêu chí có đủ tốt để phân biệt hồ sơ hay không.

Công thức

$$IV_j = \sum_{i=1}^k (P(X_j \in B_i | Y = 1) - P(X_j \in B_i | Y = 0)) \times W_o E_{ij}$$

trong đó X_j là tiêu chí thứ j , B_i là bin thứ i , Y là kết quả của hồ sơ.

FEATURE SELECTION

Weight of Evidence - WoE

Độ tuổi khách hàng: **20 - 60** tuổi

Bins	Observation	Good	Bad	Good/Bad	%Good	%Bad	WoE	IV
20 - 30	1000	105	895	0.117	0.313	0.192	0.491	0.060
30 - 35	1000	90	910	0.099	0.269	0.195	0.320	0.024
35 - 40	1000	80	920	0.087	0.239	0.197	0.191	0.008
40 - 50	1000	50	950	0.053	0.149	0.204	- 0.311	0.017
50 +	1000	10	990	0.010	0.030	0.212	- 1.961	0.358
								0.466

$$WoE_{20-30} = \ln \frac{P(X_j \in B_i | Y = 1)}{P(X_j \in B_i | Y = 0)} = \ln \frac{\%Good}{\%Bad} = \ln \frac{0.313}{0.192} = 0.491$$

$$IV = \sum_{i=1}^k (\%Good - \%Bad) \times WoE_i = 0.466$$

FEATURE SELECTION

Information Value - IV

Chỉ số giá trị thông tin, nhằm đánh giá sức ảnh hưởng của tiêu chí đó đến kết quả, kiểm tra xem tiêu chí có đủ tốt để phân biệt hồ sơ hay không.

Công thức

$$IV_j = \sum_{i=1}^k (P(X_j \in B_i | Y = 1) - P(X_j \in B_i | Y = 0)) \times W_o E_{ij}$$

trong đó X_j là tiêu chí thứ j , B_i là bin thứ i , Y là kết quả của hồ sơ.

Xếp hạng

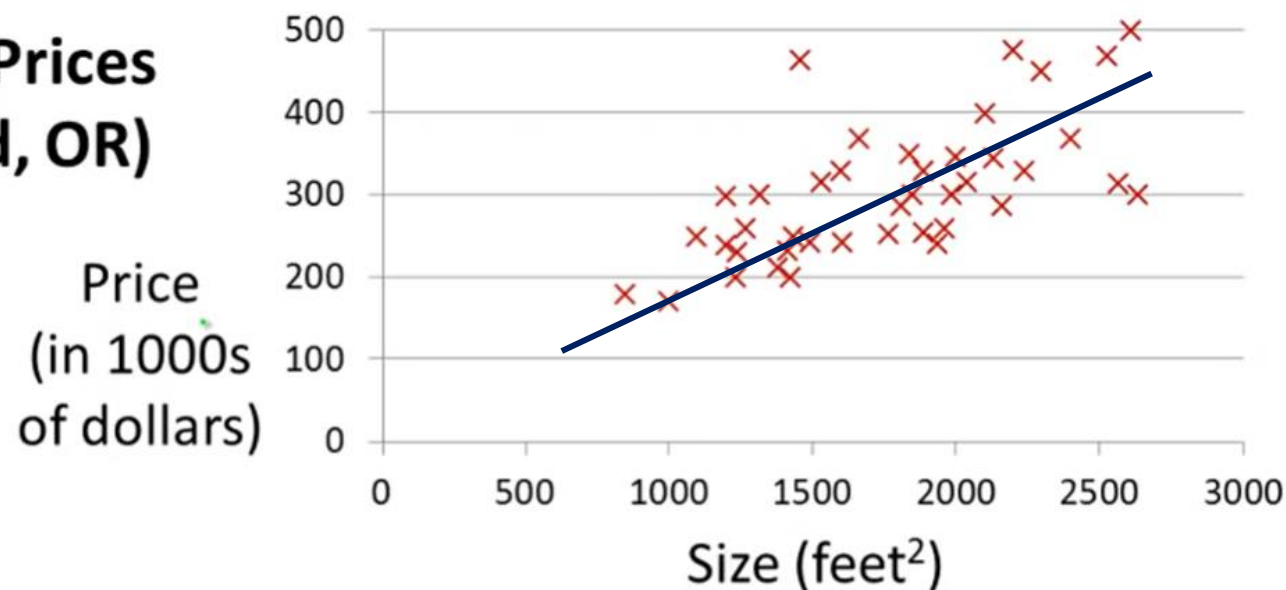
- ≤ 0.02 : Biến không có tác dụng trong việc phân loại hồ sơ
- $0.02 - 0.1$: Yếu
- $0.1 - 0.3$: Trung bình
- $0.3 - 0.5$: Mạnh
- ≥ 0.5 : Biến rất mạnh (cẩn thận tránh biến liên hệ trực tiếp)

MÔ HÌNH REGRESSION

Linear Regression

Dự đoán giá nhà dựa trên diện tích của căn nhà

Housing Prices (Portland, OR)



Supervised Learning

Đã cho kết quả ứng với mỗi dữ liệu đầu vào.

Hồi quy

Dự đoán giá trị thực



MÔ HÌNH REGRESSION

Linear Regression

Dự đoán giá nhà dựa trên diện tích của căn nhà

**Training set of
housing prices
(Portland, OR)**

Size in feet ² (x)	Price (\$) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

Biến Input

Biến kết quả

m: Số lượng bản ghi dữ liệu

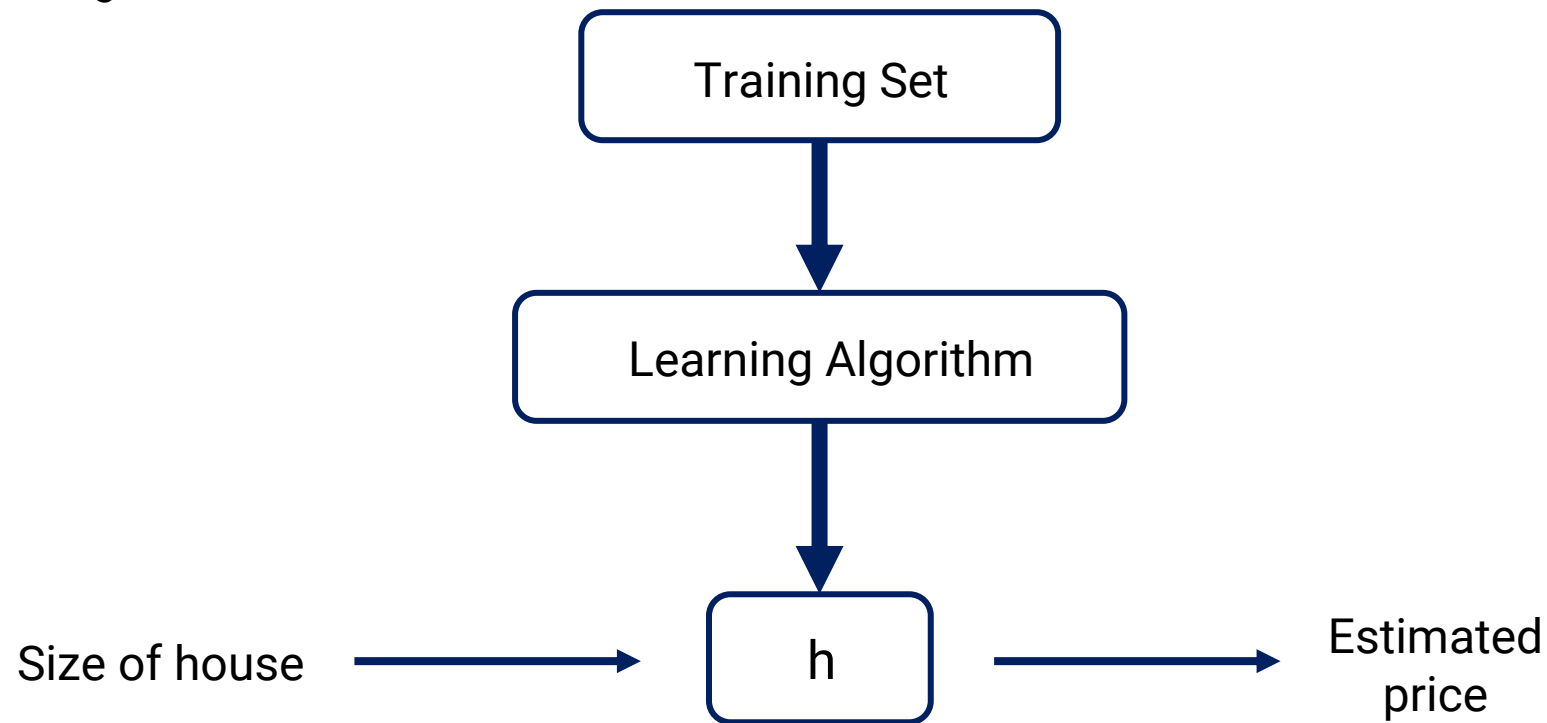
$(x^{(i)}, y^{(i)})$: cặp input và kết quả thứ i
(cặp dữ liệu train)



MÔ HÌNH REGRESSION

Linear Regression

Mô hình chung



$$h(x) = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n \quad \text{parameters}$$

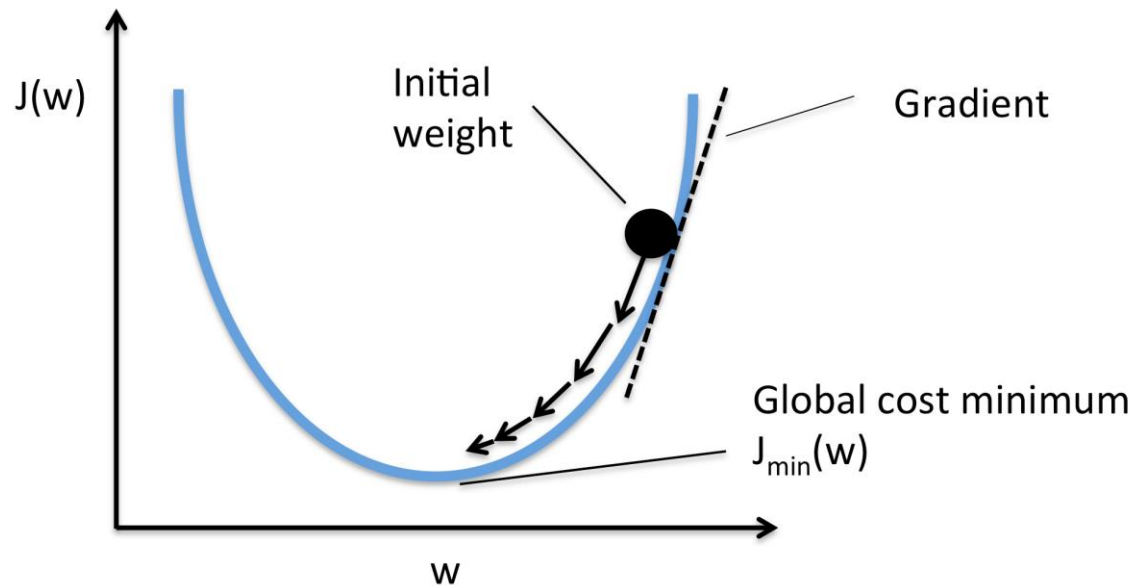
MÔ HÌNH REGRESSION

Linear Regression

Cost function (Loss function)

$$J(w) = \frac{1}{2m} \sum_{i=1}^m (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^m (h(x_i) - y_i)^2$$

Gradient Descent



$$w_0 := w_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)$$

$$w_1 := w_1 - \alpha \frac{1}{m} \sum_{i=1}^m ((h(x_i) - y_i)x_i)$$

MÔ HÌNH REGRESSION

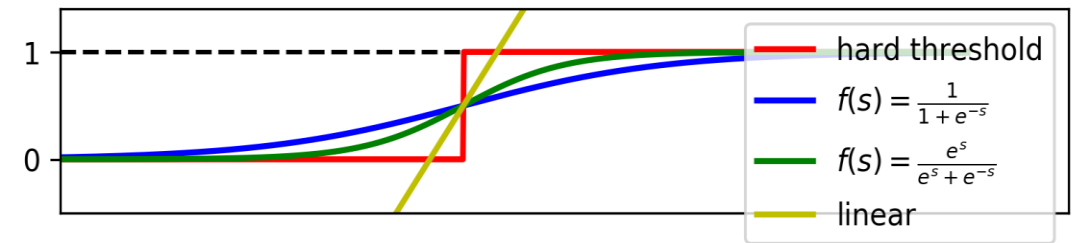
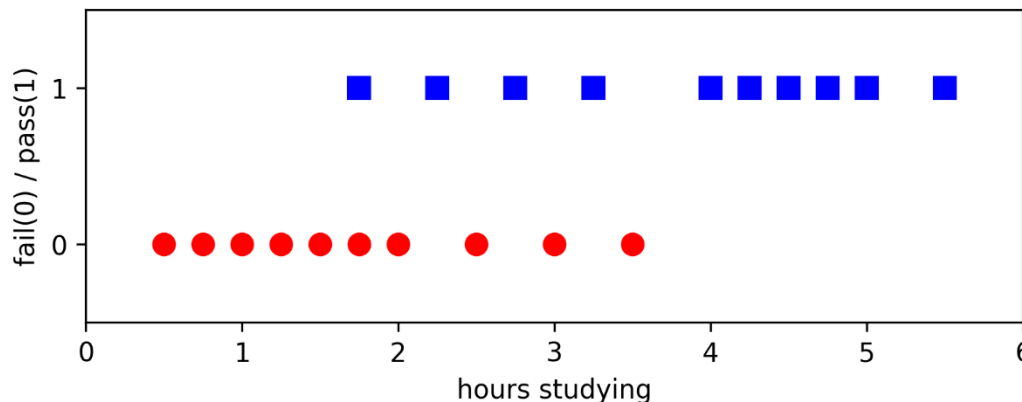
Logistic Regression

Hồi quy logistic là một thuật toán được áp dụng cho các bài toán phân loại.

Ví dụ: Thời gian học và kết quả thi của 20 học sinh được ghi lại như sau. Đưa ra một mô hình dự đoán kết quả thi



Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1



Một số loại hàm logistic có vẻ đủ linh hoạt cho vấn đề

MÔ HÌNH REGRESSION

Logistic Regression

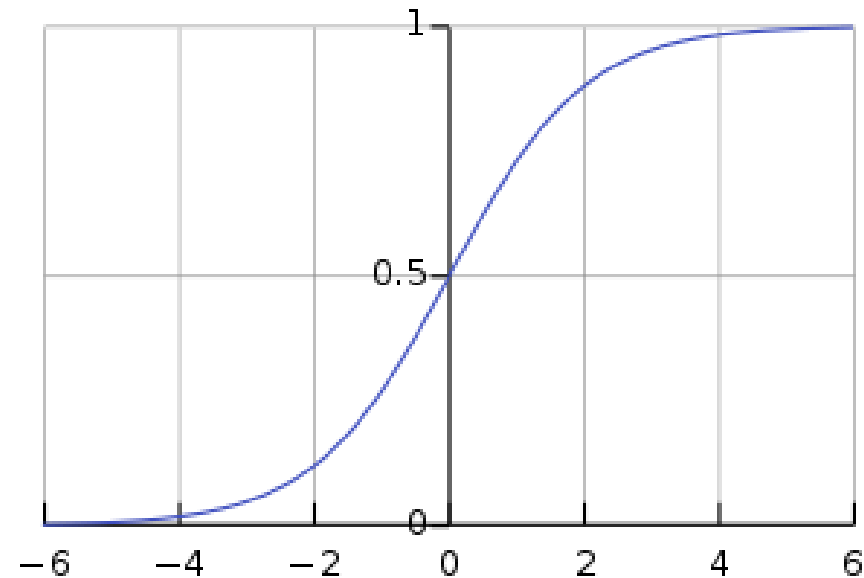
$$h(x) = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n$$



$$h(x) = f(w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n)$$

Hàm sigmoid

$$f(s) = \frac{1}{1 + e^{-s}} \triangleq \sigma(s)$$

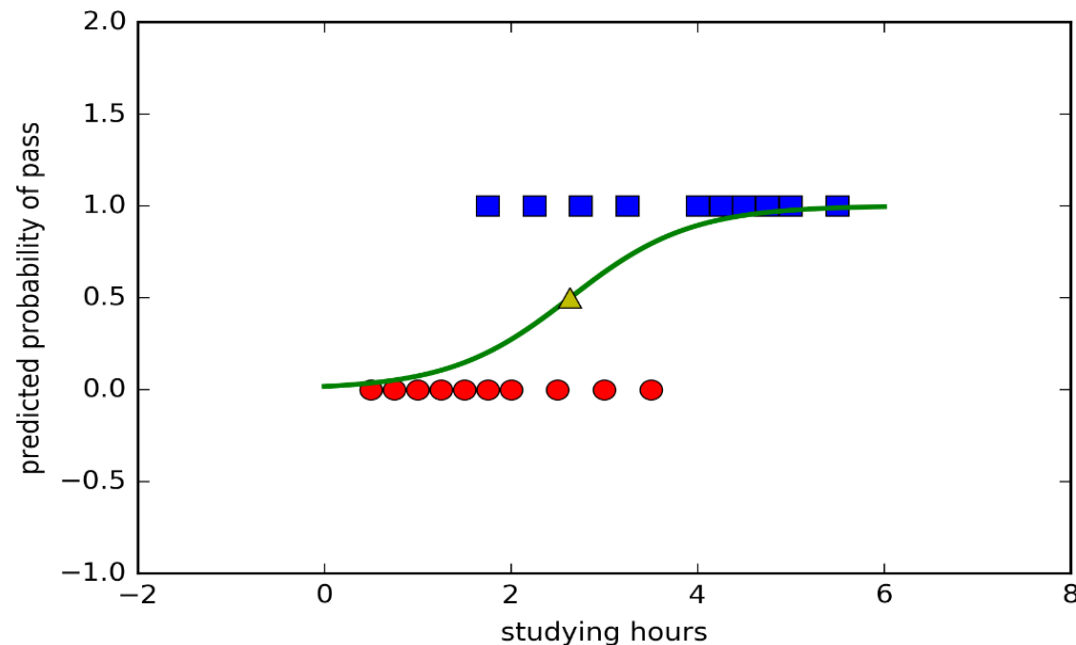


MÔ HÌNH REGRESSION

Logistic Regression

Trong mô hình hồi quy logistic, hàm logistic sẽ đưa ra xác suất một biến x thuộc lớp nào hay không. Nếu xác suất lớn hơn 0.5 thì được coi là thuộc.

Mô hình dự đoán thời gian học và kết quả thi có thể xác định như hình bên. Mặc dù vẫn có trường hợp ngoại lệ, kết quả vẫn tốt hơn so với hàm tuyến tính hay xét ngưỡng.



MÔ HÌNH REGRESSION

Logistic Regression

Hàm mất mát: Cross Entropy

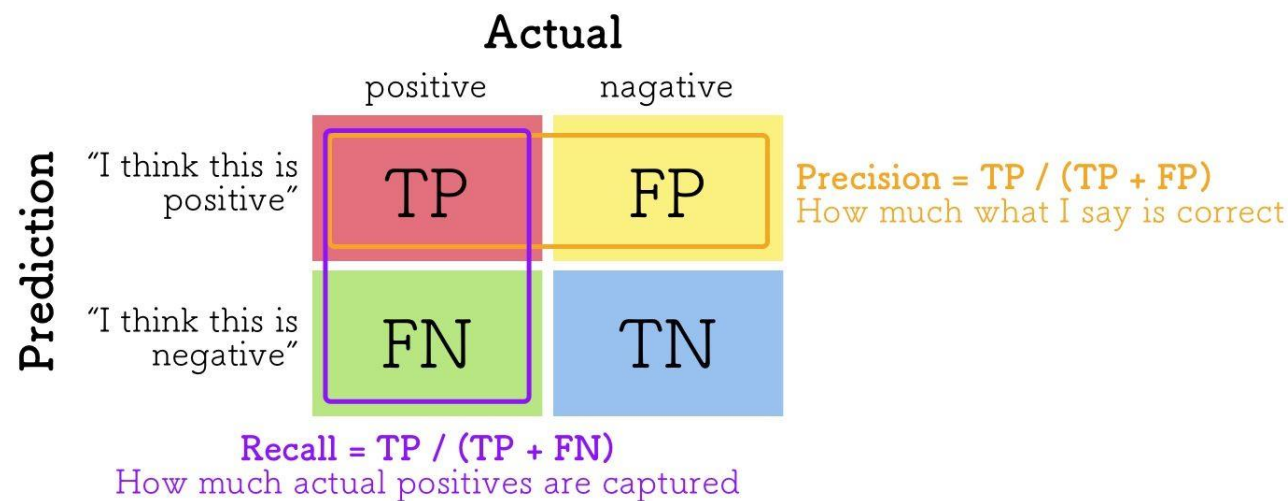
$$\begin{aligned} J(\mathbf{w}) &= -\log P(\mathbf{y}|\mathbf{X}; \mathbf{w}) \\ &= -\sum_{i=1}^N (y_i \log z_i + (1 - y_i) \log(1 - z_i)) \end{aligned}$$

ĐÁNH GIÁ ĐỘ CHÍNH XÁC CỦA MÔ HÌNH

Accuracy

Tỉ lệ số hồ sơ dự đoán đúng trên tổng số hồ sơ.

Recall & Precision



Đường cong precision, recall

Đường cong precision, recall giúp lựa chọn ngưỡng xác suất phù hợp để mang lại độ chính xác cao hơn trong precision hoặc recall. Precision cho ta biết tỷ lệ dự báo chính xác trong số các hồ sơ được dự báo là GOOD (tức nhãn là 1). Recall đo lường tỷ lệ dự báo chính xác các hồ sơ GOOD trên thực tế. Luôn có sự đánh đổi giữa 2 tỷ lệ này, nên ta cần phải dựa vào biểu đồ của 2 đường precision vs recall để tìm ra ngưỡng tối ưu

TÍNH ĐIỂM CREDIT SCORE

Tính điểm số cho mỗi feature

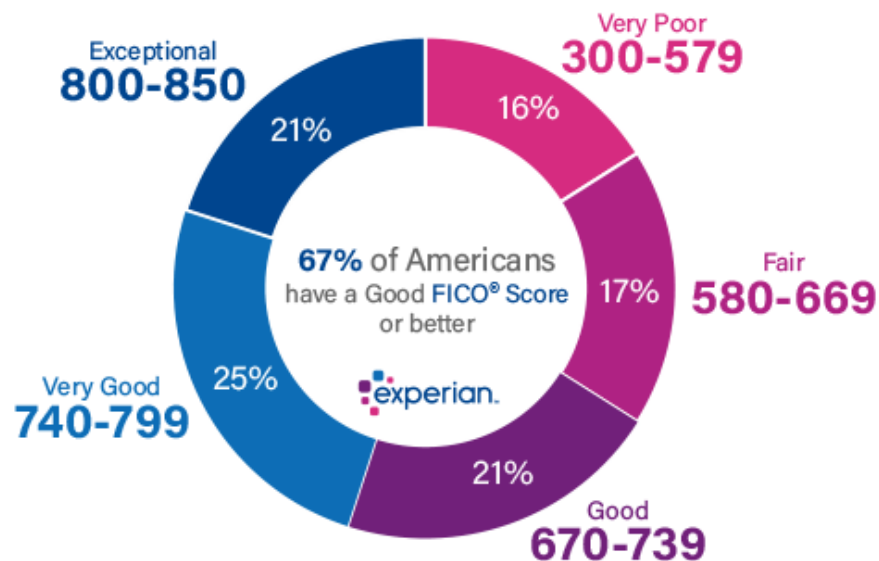
$$\text{Score} = (\beta \cdot \text{WOE} + \frac{\alpha}{n}) \cdot \text{Factor} + \frac{\text{Offset}}{n}$$

Trong đó:

- β : Hệ số của biến trong phương trình hồi quy logistic
- α : Hệ số chặn của phương trình hồi quy logistic
- WOE: Hệ số trọng số bằng chứng của mỗi feature
- n : Số lượng các biến của mô hình
- $\text{Factor} = \frac{\text{pdo}}{\ln(2)}$
- $\text{Offset} = \text{Base_Score} - (\text{Factor} \cdot \ln(\text{Odds}))$
- odds là tỷ lệ xác suất GOOD/ BAD, pdo (point double odds ratio) điểm thay đổi để gấp đôi odds ratio, Base_Score bằng 600



KẾT QUẢ ĐÁNH GIÁ



Tỉ lệ phân chia các nhóm khách hàng theo ngưỡng điểm tín nhiệm

Credit Score	Rating	% of People	Impact
300-579	Very Poor	16%	Credit applicants may be required to pay a fee or deposit, and applicants with this rating may not be approved for credit at all.
580-669	Fair	17%	Applicants with scores in this range are considered to be subprime borrowers.
670-739	Good	21%	Only 8% of applicants in this score range are likely to become seriously delinquent in the future.
740-799	Very Good	25%	Applicants with scores here are likely to receive better than average rates from lenders.
800-850	Exceptional	21%	Applicants with scores in this range are at the top of the list for the best rates from lenders.

Bảng ảnh hưởng của các nhóm khách hàng theo điểm tín nhiệm



MAGIC CODE INSTITUTE

THANKS FOR LISTENING!!!