Bài giảng môn học:

Kỹ nghệ tri thức và học máy (4080540)

CHƯƠNG 3:
HỌC CÓ GIÁM SÁT - 02
(Supervised Learning)

Giảng viên: Đặng Văn Nam

Email: dangvannam@humg.edu.vn

09/2020

Nội dung chương 3

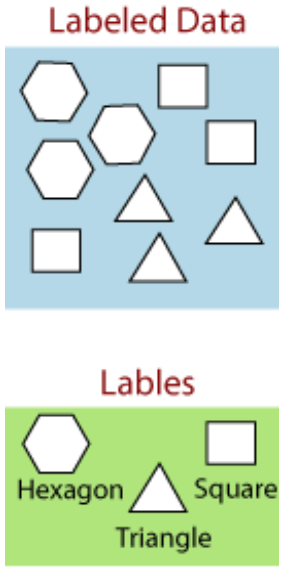
1. Các bước xây dựng một mô hình học máy
2. Datasets
3. Học có giám sát (Supervised Learning)
4. Phân loại học có giám sát (Classification, Regression)
5. Thuật toán phân loại (KNN, Decision Tree)
6. Thuật toán hồi quy (Linear Regression, KNN regression)
7. Đánh giá độ chính xác của mô hình phân lớp, hồi quy

3. Học có giám sát (supervised learning)

Học có giám sát là gì?

Một thuật toán học máy được gọi là học có giám sát (supervised learning) nếu việc xây dựng mô hình dự đoán mối quan hệ giữa đầu vào và đầu ra được thực hiện dựa trên các cặp (đầu vào - input, đầu ra – label) đã biết trong tập huấn luyện. Đây là nhóm thuật toán phổ biến nhất trong các thuật toán machine learning.

Tập dữ liệu học (Training data) bao gồm các quan sát (Examples, Observations), mà mỗi quan sát được gắn kèm với một giá trị đầu ra mong muốn (Label)



Học có giám sát là gì?

Dataset sẽ bao gồm:

- **Các thuộc tính đầu vào (Biến độc lập) – Features (input)**
- **Thuộc tính mục tiêu (Biến phụ thuộc) – Target (label)**

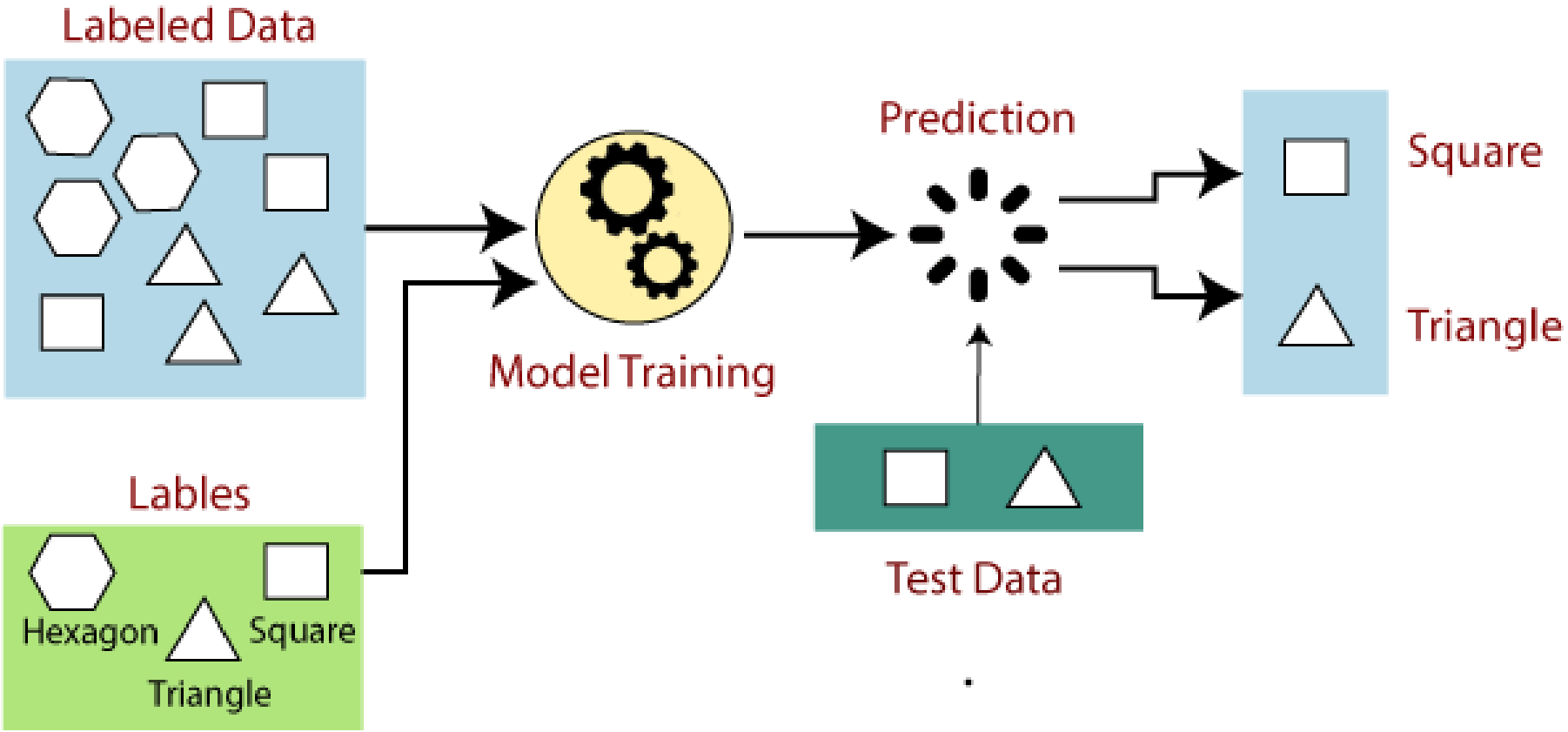
Biến độc lập (Features)

Biến phụ thuộc (Target) - Label

id	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Thalassemia	Result
Patient_01	63	Male	Typical angina	145	233	150	6	0
Patient_02	67	Male	Asymptomatic	160	286	108	3	1
Patient_03	67	Male	Asymptomatic	120	229	129	7	1
Patient_04	37	Male	Non-anginal pain	130	250	187	3	0
Patient_05	41	Female	Atypical angina	130	204	172		0
Patient_16	56	Male	Atypical angina	120	236	178	3	0
Patient_07	62	Female	Asymptomatic	140	268	160	3	1
Patient_08	57	Female	Asymptomatic	120	354	163	3	0
Patient_19	63	Male	Asymptomatic	130	254	147	7	1
Patient_10	53	Male	Asymptomatic	140	203	155	7	1
Patient_11	57	Male	Asymptomatic	140	192	148	6	0
Patient_12	56	Female	Atypical angina	140	294	153	3	0
Patient_13	56	Male	Non-anginal pain	130	256	142	6	1
Patient_14	44	Male	Atypical angina	120	263	173	7	0
Patient_15	52	Male	Non-anginal pain	172	199	162	7	0
Patient_16	57	Male	Non-anginal pain	150	168	174	3	0
Patient_17	48	Male	Atypical angina	110	229	168	7	1
Patient_18	54	Male	Asymptomatic	140	239	160	3	0
Patient_19	48	Female	Non-anginal pain	130	275	139	3	0
Patient_20	49	Male	Atypical angina	130	266	171	3	0
Patient_21	64	Male	Typical angina	110	211	144		0

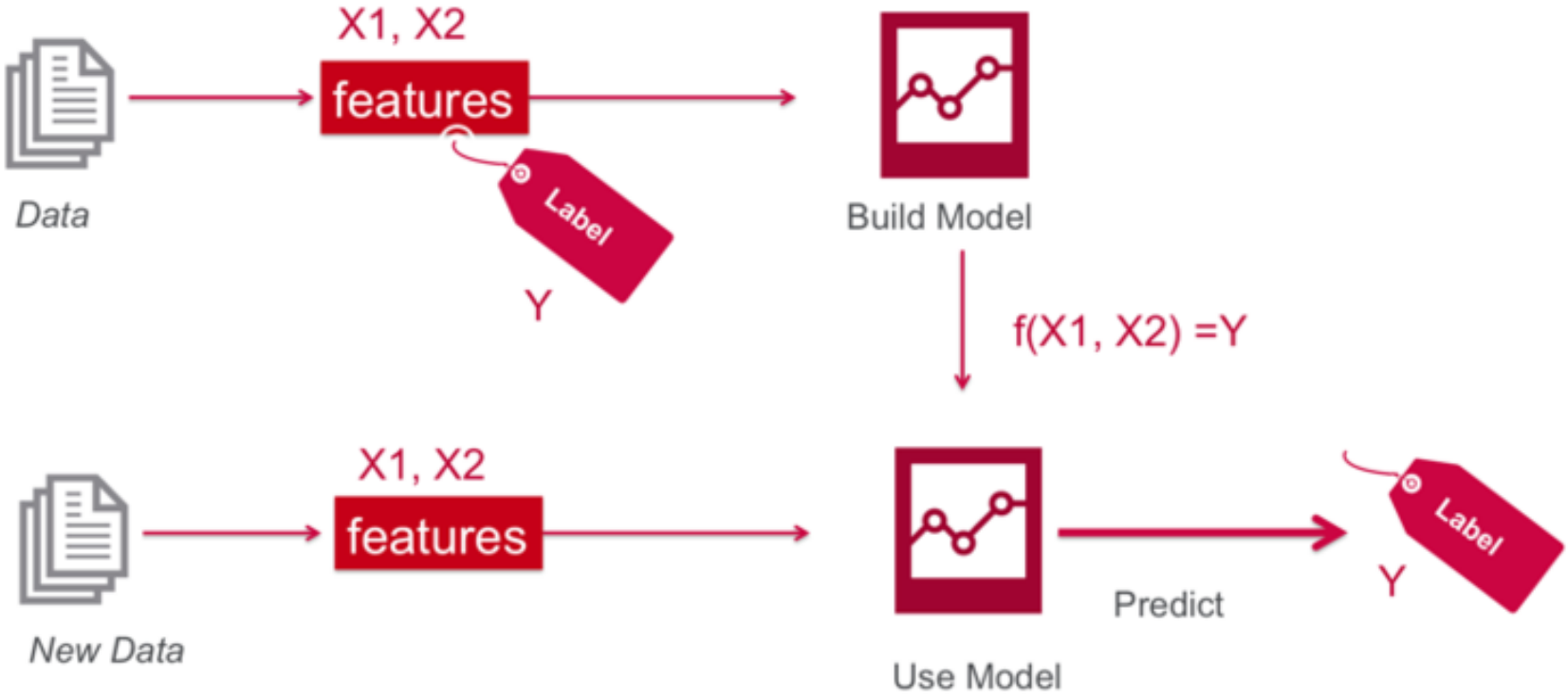
Học có giám sát là gì?

Quá trình huấn luyện và kiểm thử với Mô hình học máy có giám sát



Học có giám sát là gì?

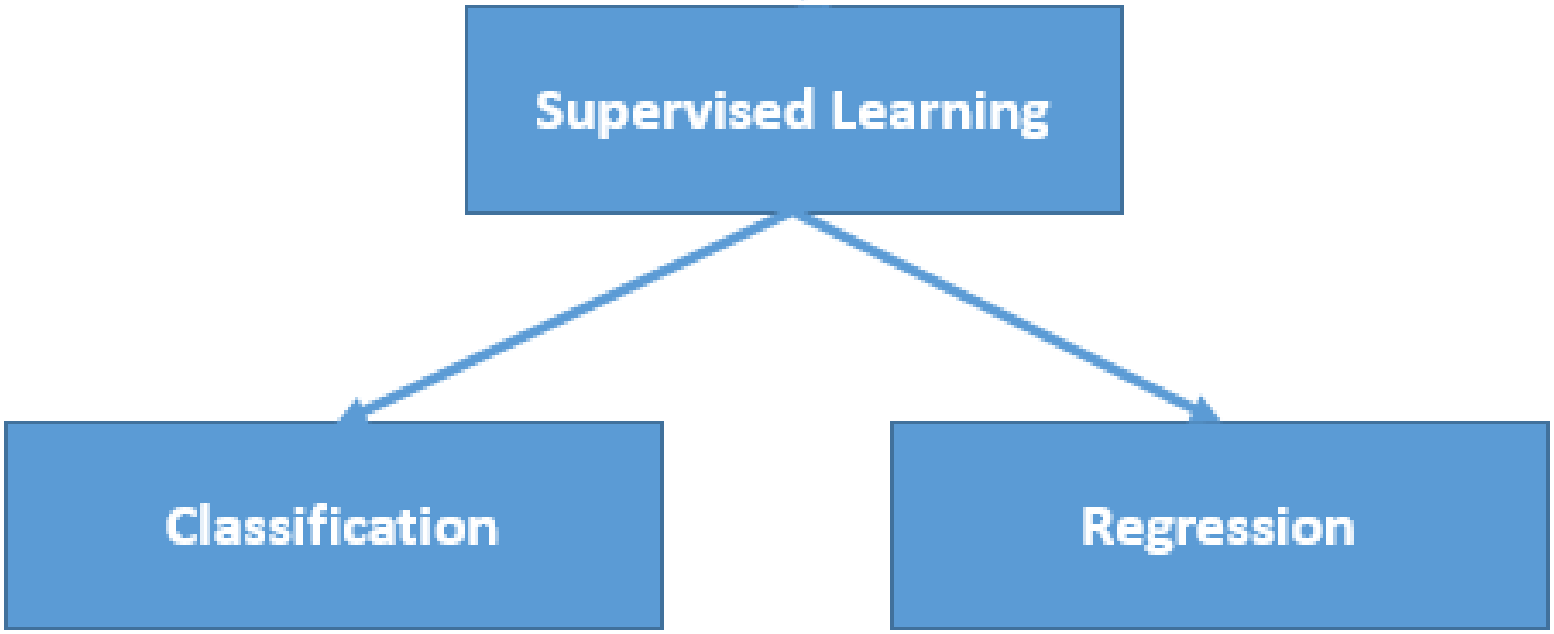
- Bản chất của Supervised learning là học một hàm f phù hợp với tập dữ liệu hiện có và có khả năng tổng quát hóa cao.
- Hàm học được sau đó sẽ dùng để dự đoán cho các quan sát mới.



4. Các loại supervised learning

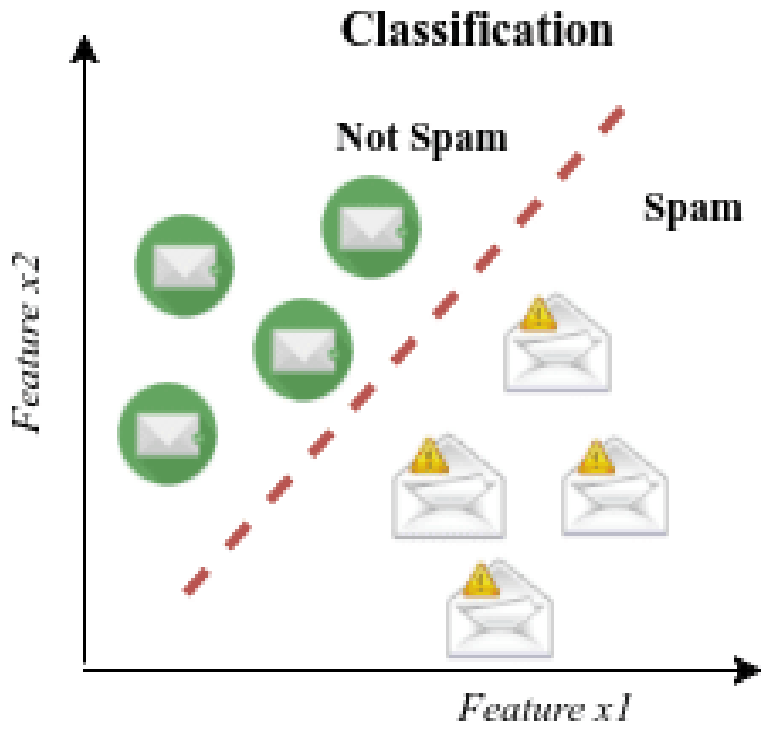
Các loại học có giám sát

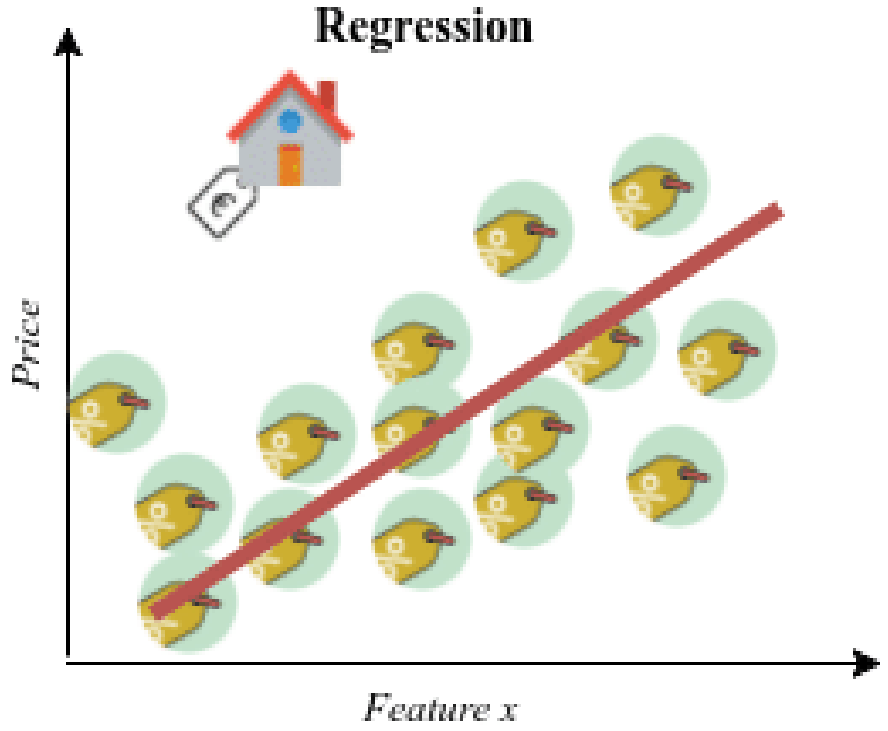
- Học có giám sát bao gồm 2 loại:
  - **Phân loại (Classification)**
  - **Hồi quy (Regression)**



Các loại học có giám sát

Phân loại (Classification) & Hồi quy (Regression)



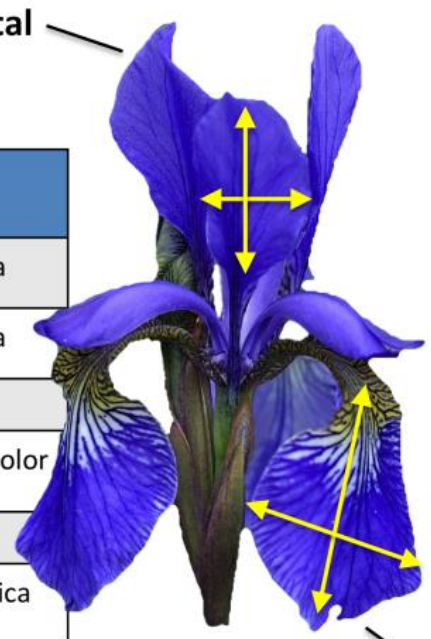


Phân loại (Classification)

Phân loại (Classification): Nếu nhãn (y – Target) thuộc tập rời rạc và hữu hạn

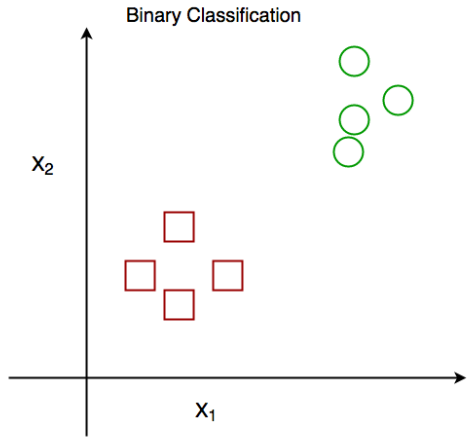
Samples
(instances, observations)

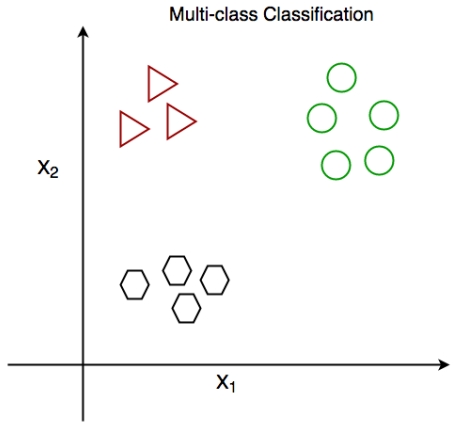
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

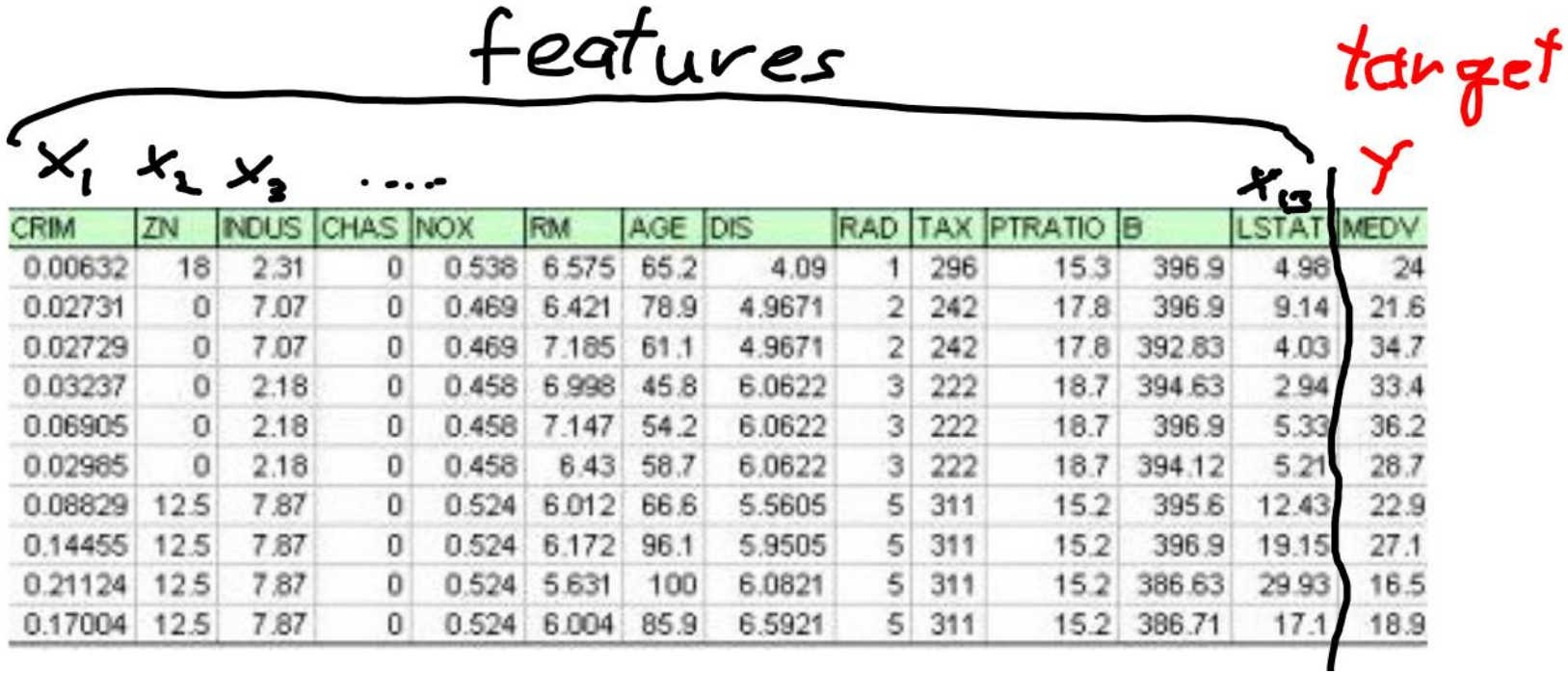
Class labels
(targets)





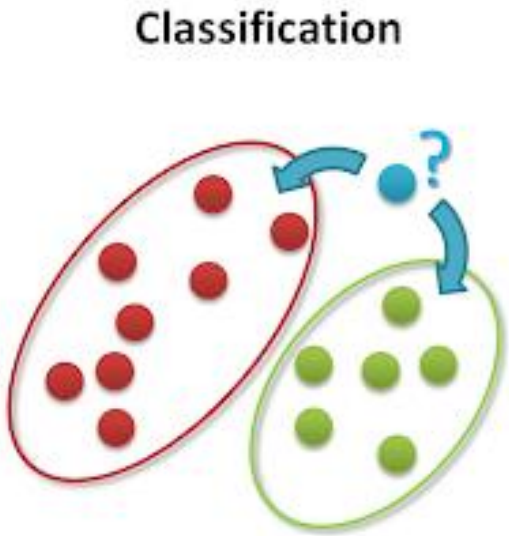
Hồi quy (Regression)

Hồi quy (Regression): Nếu nhãn (y – Target) là biến liên tục (các số thực) ví dụ như dự báo nhiệt độ, giá nhà, mức tiêu thụ điện năng…



CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9

5. Một số thuật toán phân loại (classification)



5.1 KNN
(K – Nearest Neighbors)

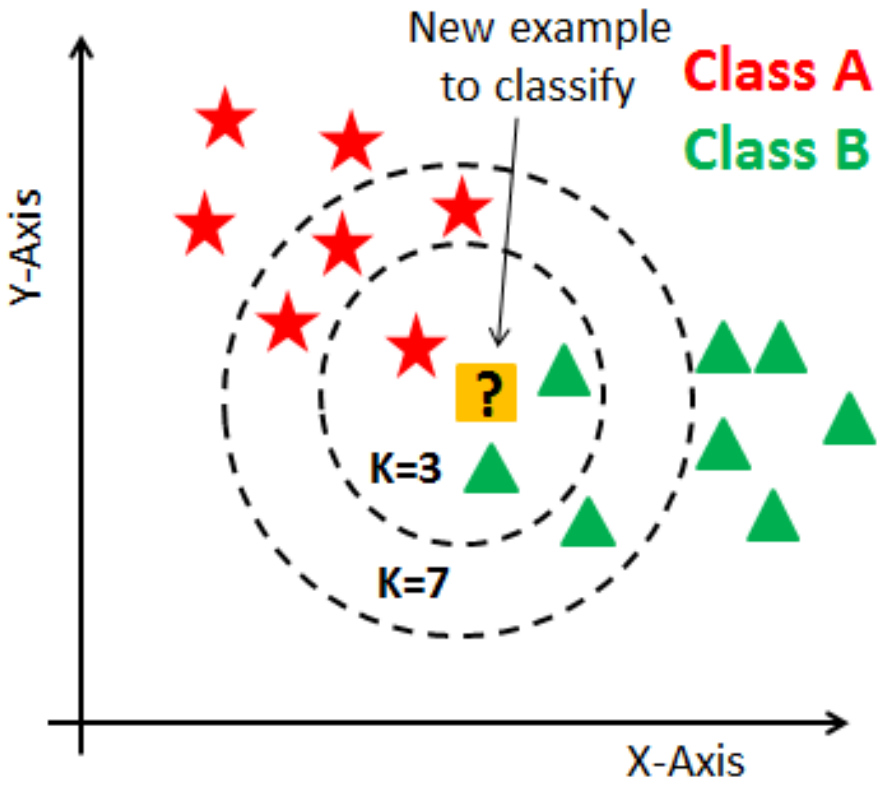
KNN

K-Nearest neighbors(k-NN) là một trong những thuật toán đơn giản nhất và phổ biến trong học máy. Một số tên gọi khác:

- Instance-based learning
- Lazy learning
- Memory-based learning

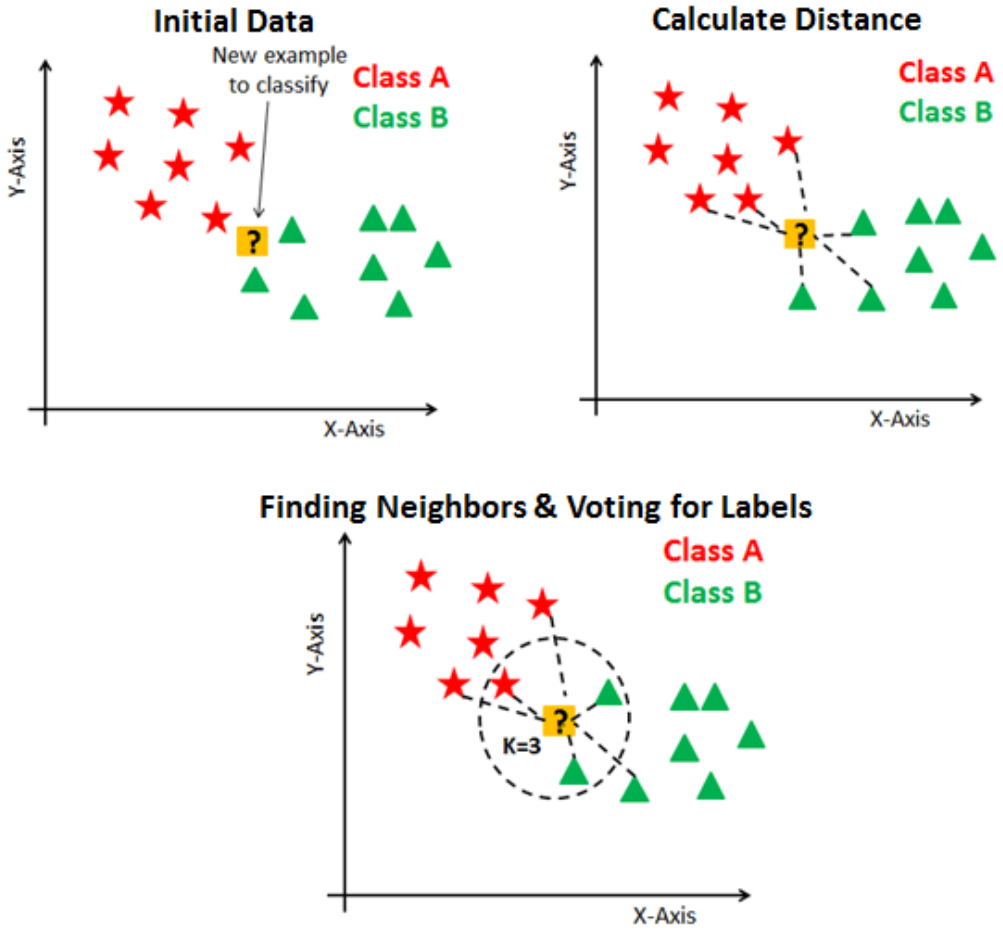
KNN

Bản chất, KNN là thuật toán đi tìm đầu ra cuả một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong tập huấn luyện gần nó nhất (K-lân cận)



KNN được sử dụng cho cả bài toán phân loại và hồi quy

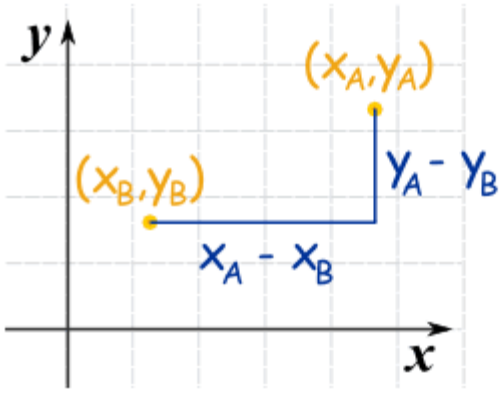
KNN



Những hàng xóm nào sẽ được sử dụng cho việc dự đoán?

KNN

Tính khoảng cách giữa hai điểm A - B



Now label the coordinates of points A and B.

x_A means the x-coordinate of point A

y_A means the y-coordinate of point A

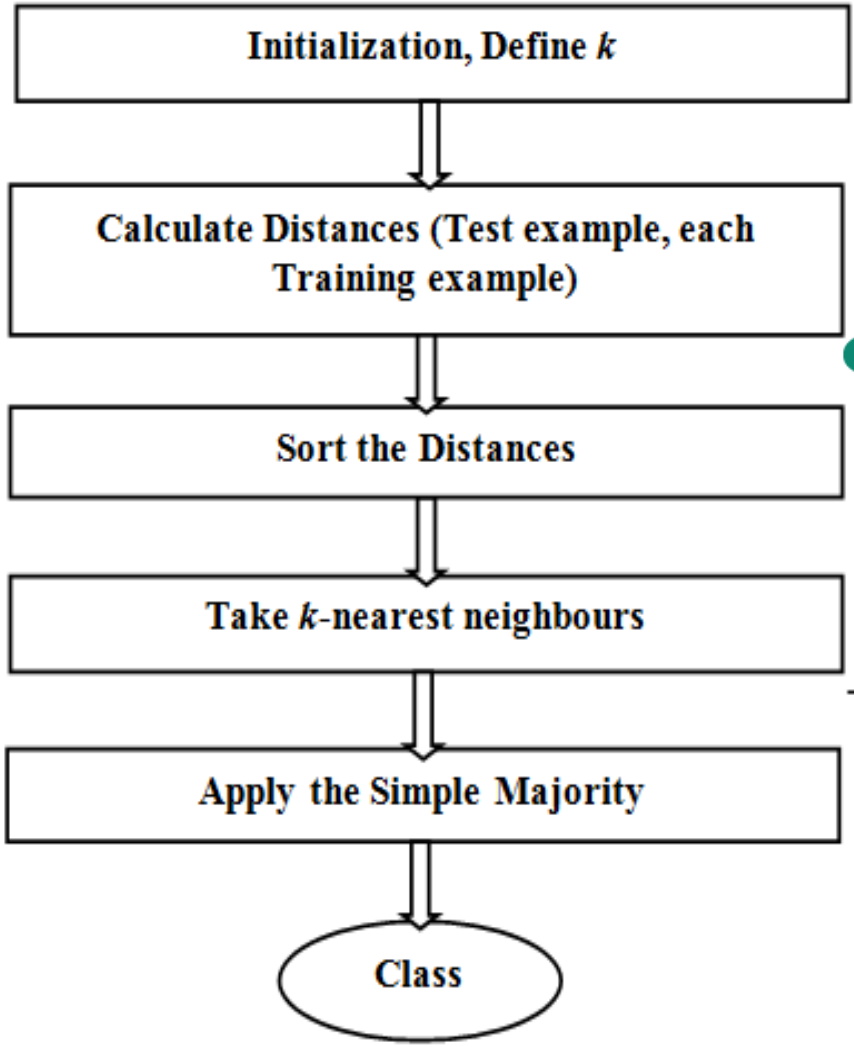
The horizontal distance **a** is $(x_A - x_B)$

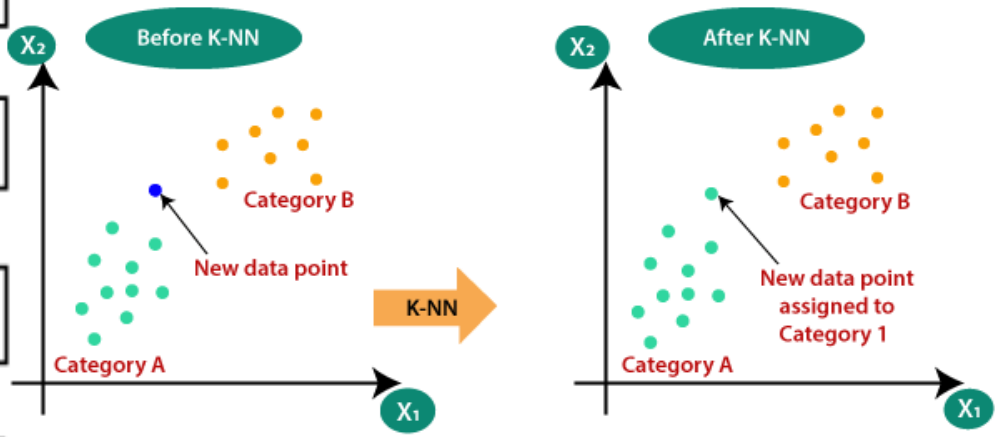
The vertical distance **b** is $(y_A - y_B)$

Euclidean Distance = $\sqrt{(x_A - x_B) + (y_A - y_B)}$

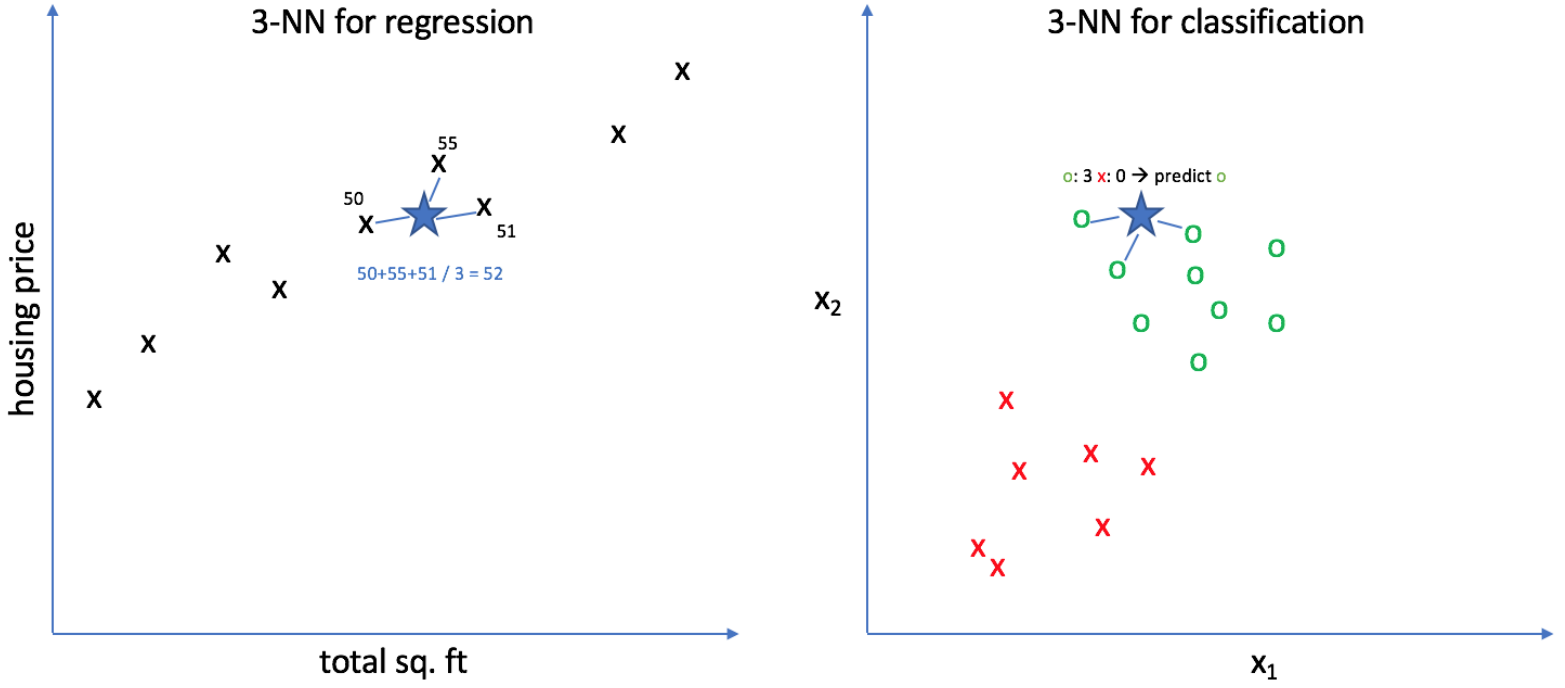
Manhattan Distance = $|(x_A - x_B)| + |(y_A - y_B)|$

Các bước thực hiện thuật toán KNN





KNN cho bài toán phân lớp và hồi quy



Hồi quy (regression): nhãn của điểm dữ liệu mới được là nhãn của điểm dữ liệu đã biết gần nhất (K=1) hoặc trung bình có trọng số của những điểm gần nhất.

Phân loại (classification): nhãn của điểm dữ liệu mới được suy ra trực tiếp từ K điểm dữ liệu gần nhất.

Ưu nhược điểm của KNN

Ưu điểm:

- Độ phức tạp tính toán trong quá trình huấn luyện bằng 0
- Việc dự đoán kết quả của dữ liệu mới rất đơn giản
- Không cần giả sử gì về phân phối của các class

Nhược điểm:

- KNN rất nhạy với nhiễu khi K nhỏ.
- Tính toán khoảng cách tới từng điểm dữ liệu trong trong tập huấn luyện tốn rất nhiều thời gian, đặc biệt với các CSDL có số chiều lớn và có nhiều điểm dữ liệu. K càng lớn thì độ phức tạp càng tăng.
- Lưu toàn bộ dữ liệu trong bộ nhớ ảnh hưởng tới hiệu năng của KNN

Ví dụ 1: Phân lớp hoa lan với KNN

Ví dụ: Phân lớp hoa lan với KNN

- Tập dữ liệu bao gồm 150 mẫu về thông số chiều rộng, chiều dài của lá hóa và cánh hoa của 3 loại hoa Lan

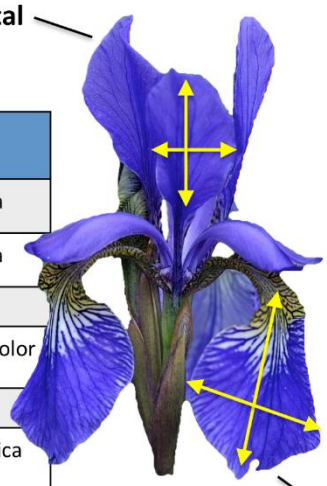
Iris Versicolor | Iris Setosa | Iris Virginica

Iris Dataset

Classes	3
Samples per class	50
Samples total	150
Dimensionality	4
Features	real, positive

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

- **Tham khảo tiến trình thực hiện trong file code trên Jupyter Notebook**

THỰC HÀNH 3.1

Yêu cầu 1:

- Sinh viên tìm hiểu về tập dữ liệu mẫu wine trong Dataset của Sklearn (xác định các features và label)

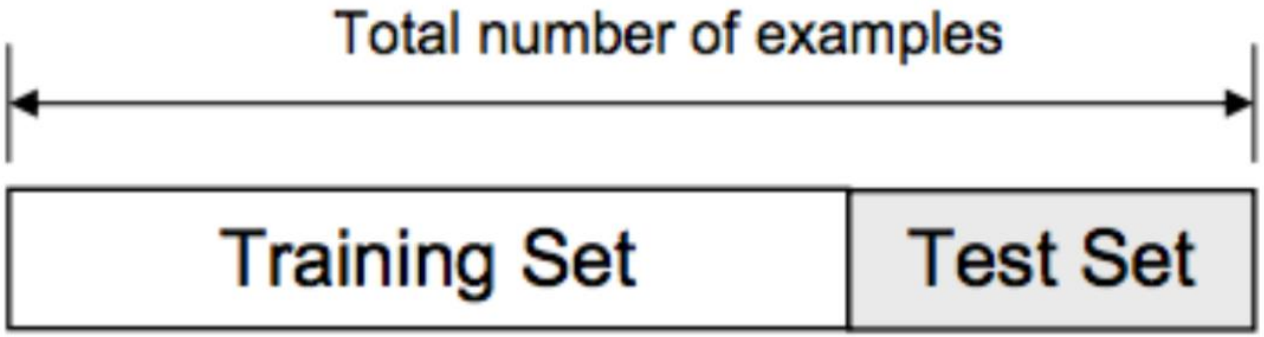
Number of Instances:	178 (50 in each of three classes)
Number of Attributes:	13 numeric, predictive attributes and the class
Attribute Information:	• Alcohol • Malic acid • Ash • Alcalinity of ash • Magnesium • Total phenols • Flavanoids • Nonflavanoid phenols • Proanthocyanins • Color intensity • Hue • OD280/OD315 of diluted wines • Proline

- **class:**
  - class_0
  - class_1
  - class_2

Classes	3
Samples per class	[59,71,48]
Samples total	178
Dimensionality	13
Features	real, positive

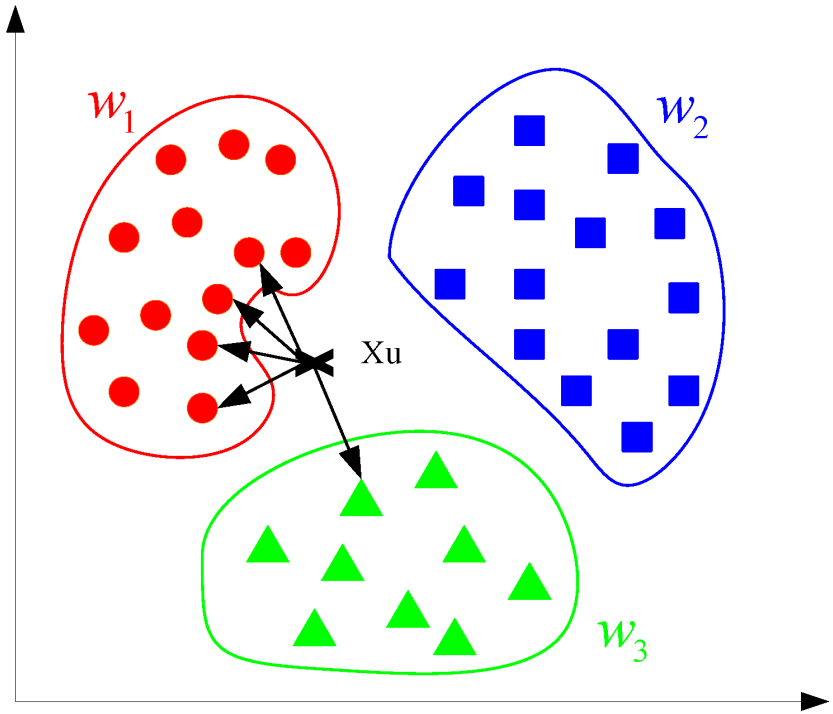
Yêu cầu 2:

- Tách tập dữ liệu data_wine thành 2 phần train – test theo tỷ lệ 75% - 25%



Yêu cầu 3:

- Sử dụng thuật toán KNN với các trường hợp: K=1, 5, 7, 11 cho biết độ chính xác ứng với từng trường hợp của K trên tập Test.
- Áp dụng thuật toán KNN với K=9 và có đánh trọng số các điểm lân cận. cho biết độ chính xác của thuật toán trên tập Test.

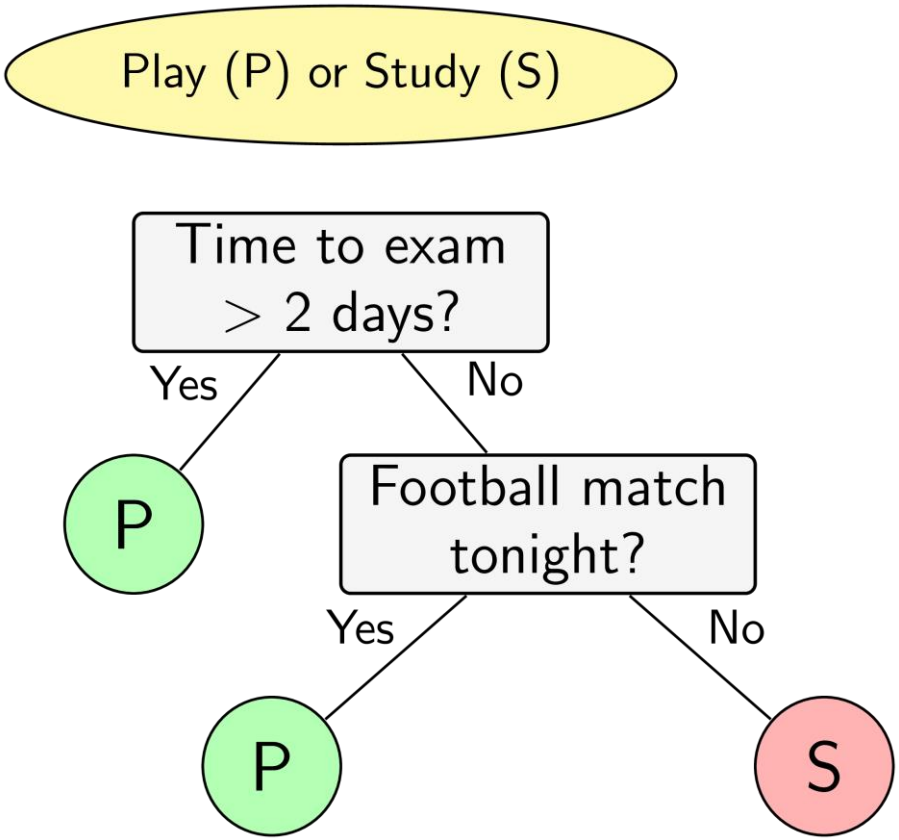


5.2 Cây quyết định (Decision Tree)

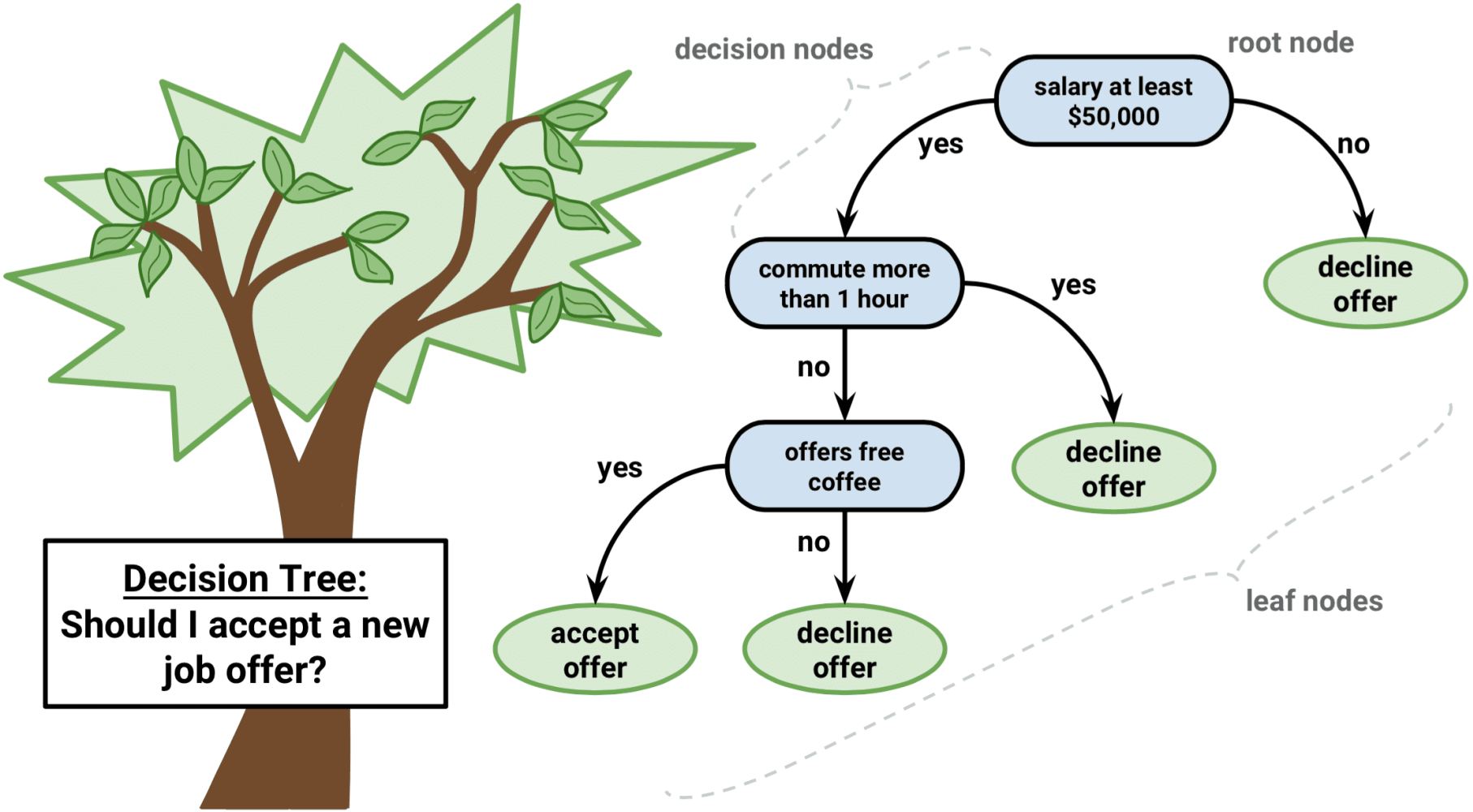
Cây quyết định:

- Việc quan sát, suy nghĩ và ra các quyết định của con người thường được bắt đầu từ các câu hỏi. Machine learning cũng có một mô hình ra quyết định dựa trên các câu hỏi. Mô hình này có tên là *cây quyết định* (*decision tree*).

- Decision tree là một mô hình học có giám sát, có thể được áp dụng vào cả hai bài toán classification và regression.

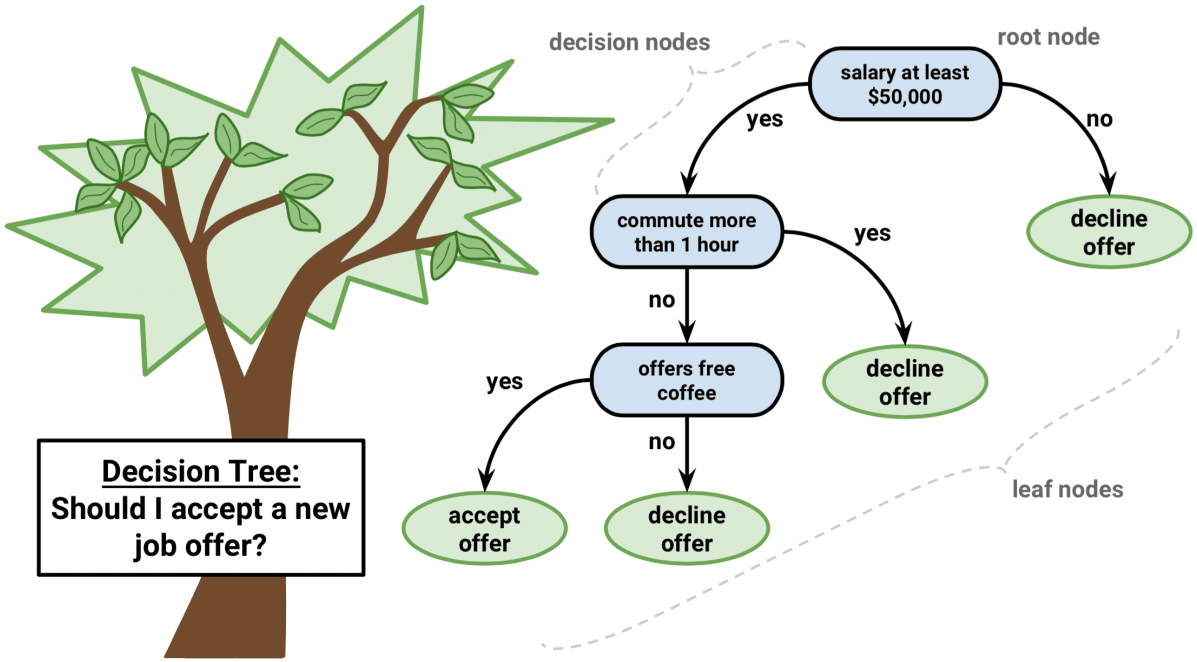


Cây quyết định:



Việc xây dựng một decision tree trên dữ liệu huấn luyện cho trước là việc đi xác định các *câu hỏi* và *thứ tự của chúng*

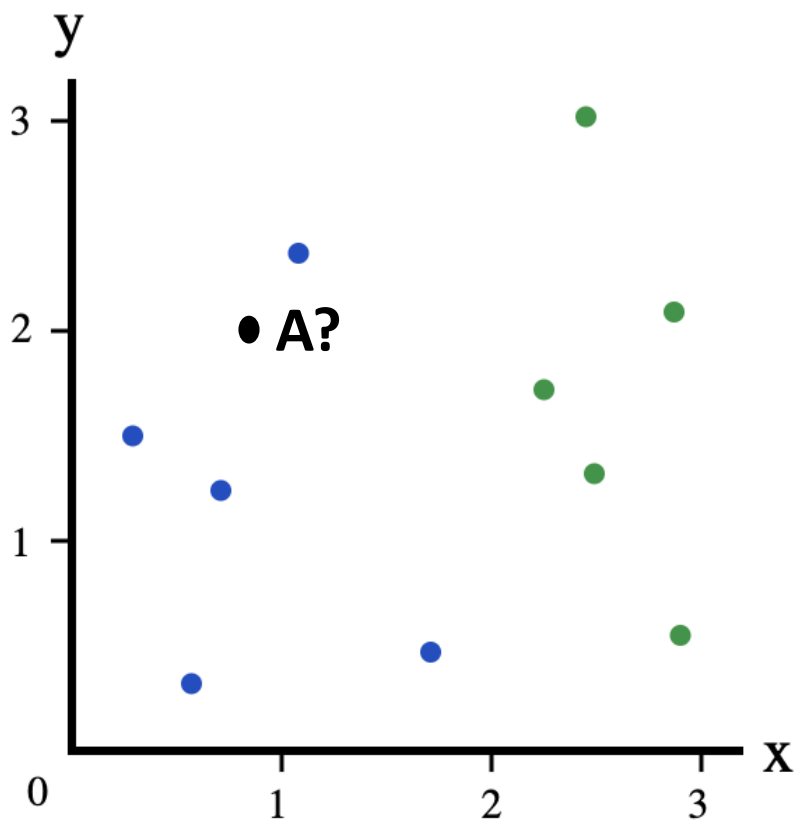
Cây quyết định:



- Dùng cấu trúc cây để đưa ra một hàm phân lớp cần học (hàm mục tiêu có giá trị rời rạc)
- Một cây quyết định có thể được biểu diễn (diễn giải) bằng một tập các luật IF-THEN (dễ đọc và dễ hiểu)
- Được áp dụng thành công trong rất nhiều các bài toán ứng dụng thực tế

Ví dụ cây quyết định:

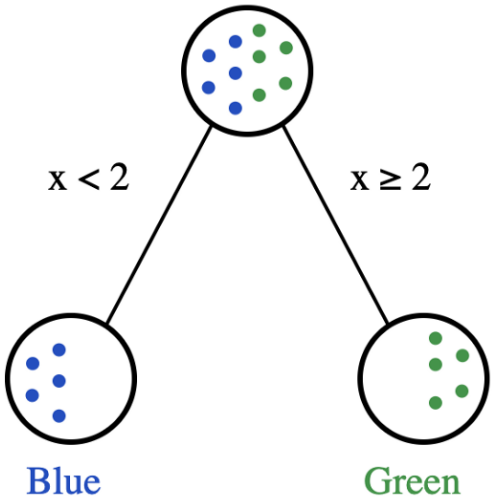
- Cho tập dữ liệu gồm 10 mẫu, thuộc 2 lớp:
  - Lớp blue: 5 mẫu
  - Lớp green: 5 mẫu

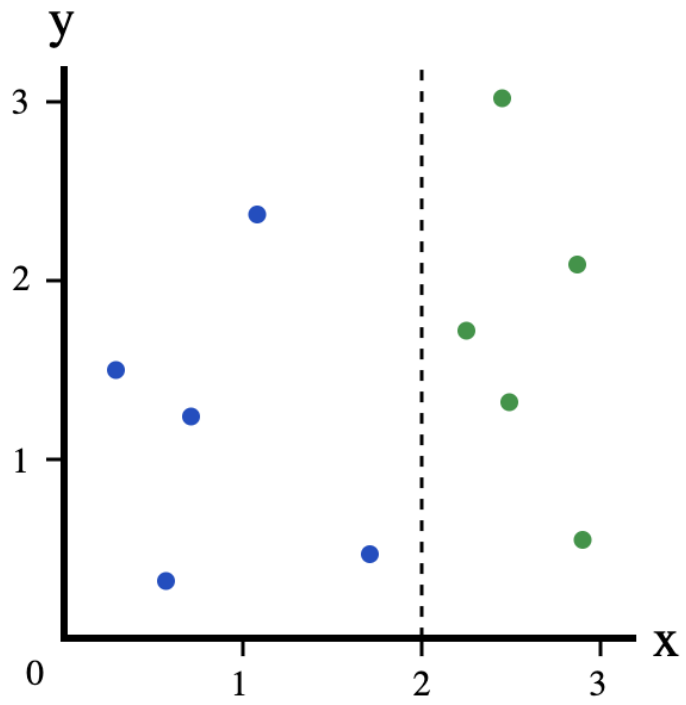


Có điểm dữ liệu mới với giá trị thuộc tính A(x = 1, y = 2) màu của điểm này nên là gì? (nên phân vào lớp nào?)

Ví dụ cây quyết định:

- Đây là cây quyết định đơn giản với một node phân loại kiểm tra xem x<2?

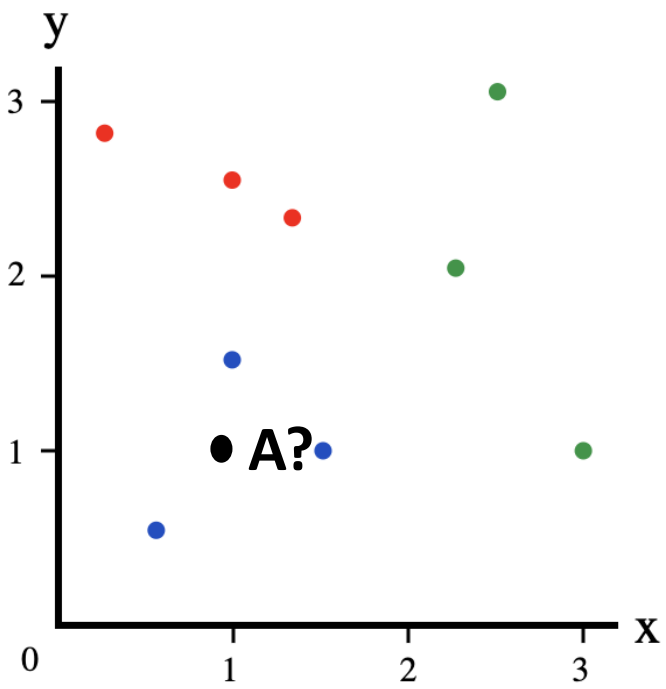




Nếu kiểm tra x < 2, chúng ta lấy nhánh trái và gán nhãn blue, nếu kiểm tra không đúng (x ≥ 2), chúng ta lấy nhánh phải và gán nhãn green.

Ví dụ cây quyết định:

- Ví dụ tập dữ liệu có 9 mẫu gồm 3 lớp:
  - Lớp Blue: 3 mẫu
  - Lớp Green: 3 mẫu
  - Lớp Red: 3 mẫu

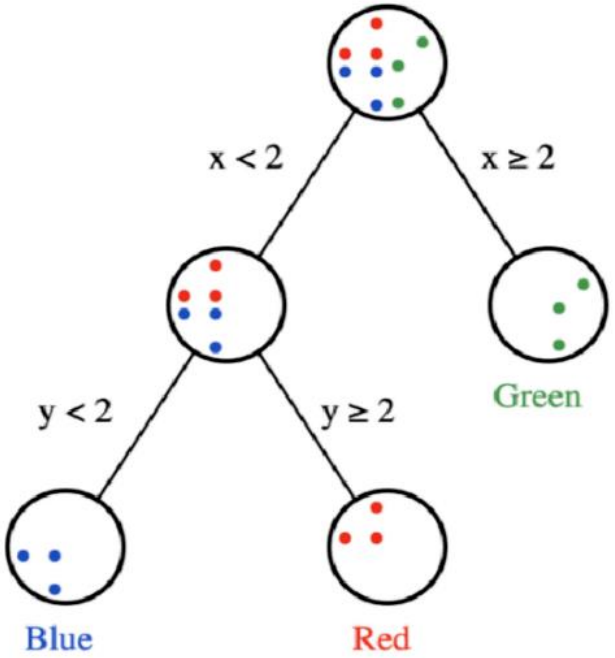


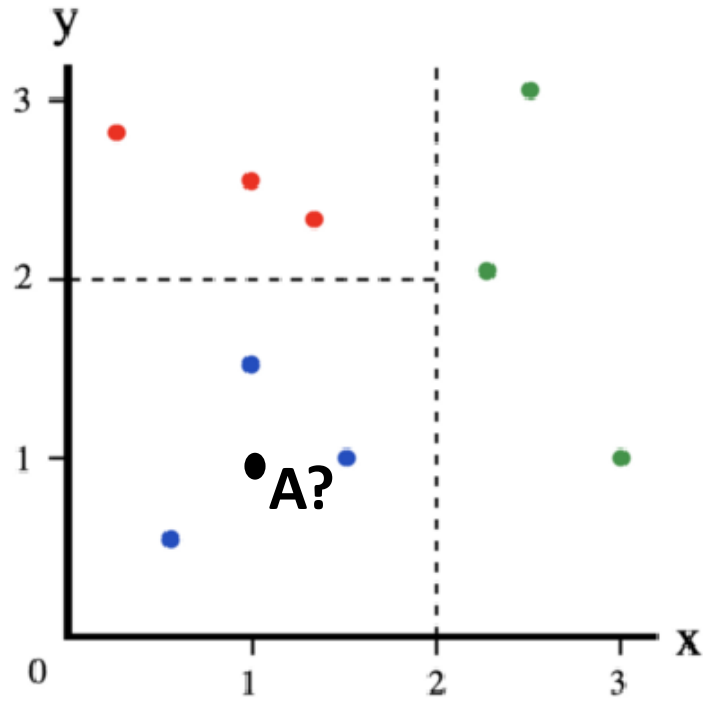
Cây quyết đĩnh cũ không hiệu quả, với mẫu dữ liệu mới A(x=1,y=1)

- Nếu x ≥ 2, chúng ta có thể vẫn tự tin phân loại vào Green
- Nếu x < 2, chúng ta không thể phân loại ngay vào Blue, nó cũng có thể vào Red.

Ví dụ cây quyết định:

- Chúng ta cần them node quyết định vào cây quyết định





Đó là ý tưởng chính của cây ra quyết định?

Các độ đo lựa chọn thuộc tính

- Một độ đo lựa chọn thuộc tính là một phương pháp tiên nghiệm (heuristic) để lựa chọn tiêu chí phân chia để phân tách tốt nhất phần dữ liệu D đã cho
- Một cách lý tưởng
  - Mỗi phần được chia ra nên thuần nhất
  - Mỗi phần thuần nhất là phần chứa các mẫu cùng thuộc một lớp
- Các độ đo phân chia thuộc tính (các luật phân chia)
  - Xác định các mẫu ở một node được phân chia thế nào
  - Đưa ra cách xếp hạng các thuộc tính
  - Thuộc tính với điểm cao nhất được lựa chọn
  - Xác định một điểm phân chia hoặc một tập con phân chia
- Các phương pháp
  - **Information gain (Entropy)**
  - **Gain ratio**
  - **Gini Index**

Các độ đo lựa chọn thuộc tính

Gini Index

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

p_j: proportion of the samples that belongs to class c for a particular node

Entropy

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$$

p_j: proportion of the samples that belongs to class c for a particular node.

*This is the the definition of entropy for all non-empty classes ($p \neq 0$). The entropy is 0 if all samples at a node belong to the same class.

Thư viện Sklearn sử dụng 2 độ đo: Gini (Mặc định), Entropy

Cây quyết định

- Cây quyết định có tốc độ học tương đối nhanh so với các phương pháp khác
- Đơn giản và dễ hiểu các luật phân loại trong cây ra quyết định
- Information Gain, Gain Ratio, và Gini Index là những phương pháp lựa chọn thuộc tính thông dụng nhất
- Cắt tỉa cây là cần thiết để loại bỏ những nhánh không tin cậy

- **Ưu điểm:**
  - Dễ hiểu: Cây biểu diễn trực quan
  - Hữu ích: Xác định được các biến quan trọng
  - Phi tham số: không cần giả định về phân phối
  - Không phục thuộc vào dữ liệu: Có thể áp dụng cả dữ liệu phân loại và liên tục
- **Nhược điểm:**
  - Dễ bị quá khớp (overfitting)

Ví dụ 2: Phân lớp hoa lan với Decision Tree

Ví dụ: Phân lớp hoa lan với Decision tree

- Tập dữ liệu bao gồm 150 mẫu về thông số chiều rộng, chiều dài của lá hóa và cánh hoa của 3 loại hoa Lan

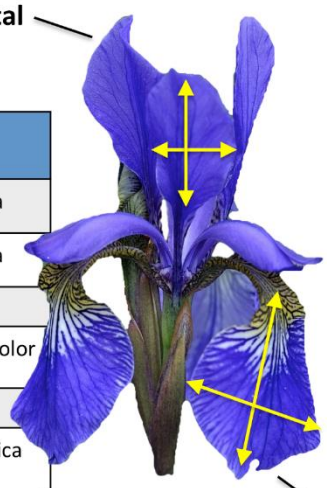
Iris Versicolor | Iris Setosa | Iris Virginica

Iris Dataset

Classes	3
Samples per class	50
Samples total	150
Dimensionality	4
Features	real, positive

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

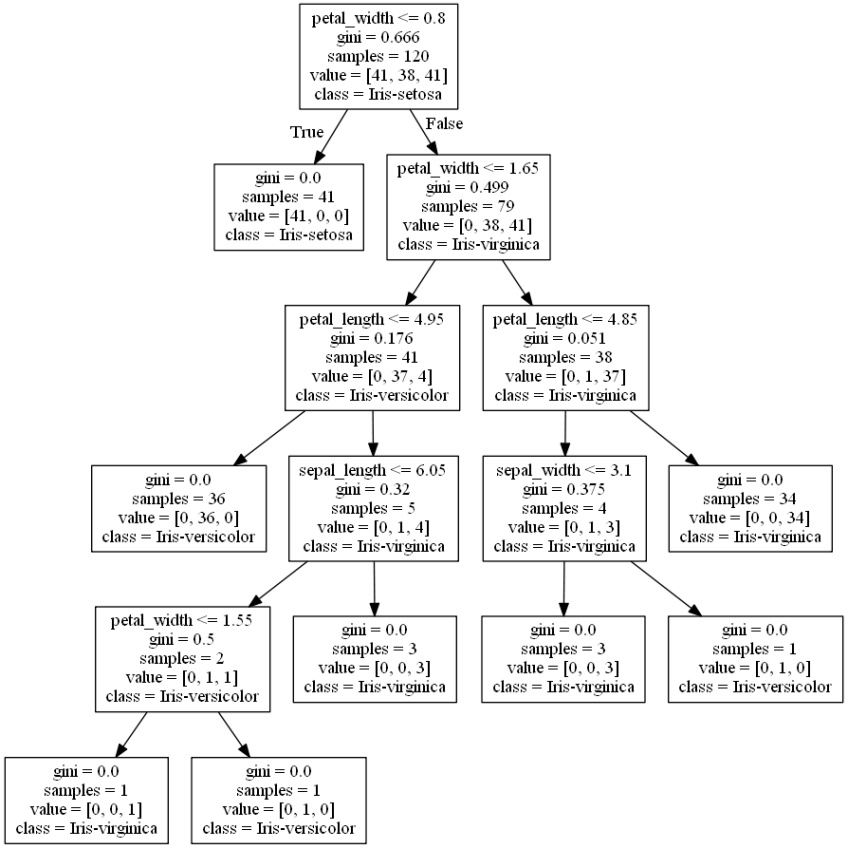


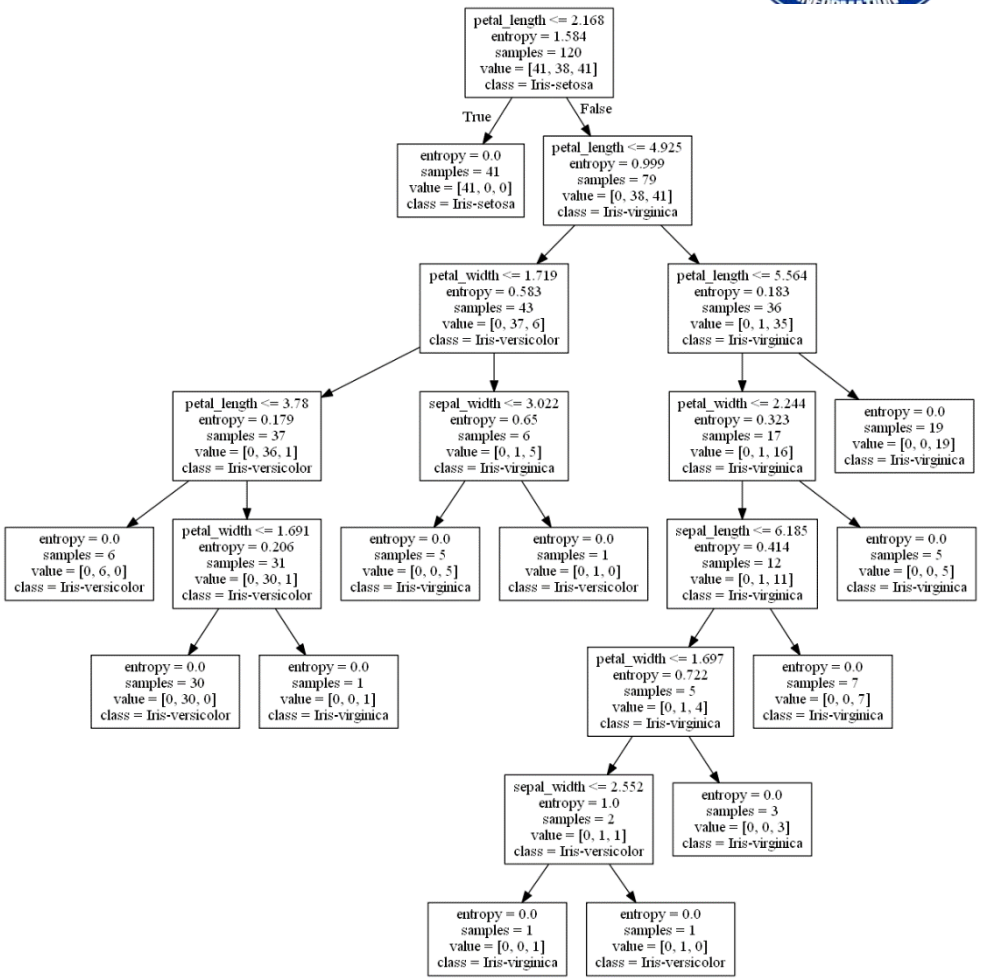
Features
(attributes, measurements, dimensions)

Class labels
(targets)

- **Tham khảo tiến trình thực hiện trong file code trên Jupyter Notebook**

Ví dụ: Phân lớp hoa lan với Decision tree





DecisionTreeClassifier(criterion= 'gini',
splitter='best',
random_state=0)

DecisionTreeClassifier(criterion= 'entropy',
splitter='random',
random_state=0)

THỰC HÀNH 3.2

Yêu cầu 1:

- Sinh viên tìm hiểu về tập dữ liệu mẫu wine trong Dataset của Sklearn (xác định các features và label)

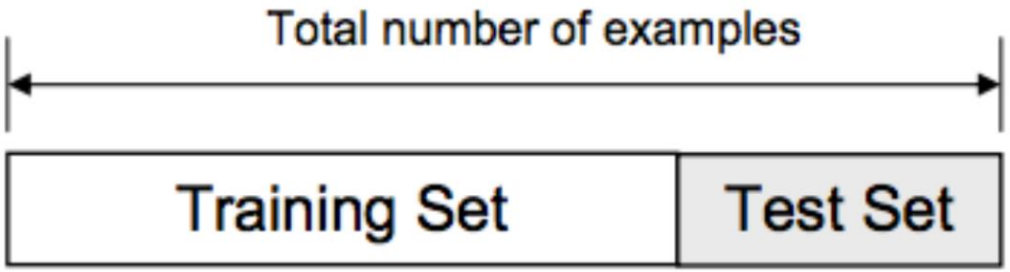
Number of Instances:	178 (50 in each of three classes)
Number of Attributes:	13 numeric, predictive attributes and the class
Attribute Information:	• Alcohol • Malic acid • Ash • Alcalinity of ash • Magnesium • Total phenols • Flavanoids • Nonflavanoid phenols • Proanthocyanins • Color intensity • Hue • OD280/OD315 of diluted wines • Proline

- **class:**
  - class_0
  - class_1
  - class_2

Classes	3
Samples per class	[59,71,48]
Samples total	178
Dimensionality	13
Features	real, positive

Yêu cầu 2:

- Tách tập dữ liệu data_wine thành 2 phần train – test theo tỷ lệ 75% - 25%



Yêu cầu 3:

- Sử dụng thuật toán Cây quyết định trong 2 trường hợp:
  - Sử dụng độ đo Entropy: Trực quan hóa cây quyết định thu được trên tập Huấn luyện, xác định thuộc tính quan trọng và vẽ biểu đồ; xác định độ chính xác của mô hình trên tập Test.
  - Sử dụng độ đo Gini: Trực quan hóa cây quyết định thu được trên tập Huấn luyện, xác định thuộc tính quan trọng và vẽ biểu đồ; xác định độ chính xác của mô hình trên tập Test.

Yêu cầu 4:

- Một mẫu rượu có các tham số như sau:

  - Alcohol : 12.7
  - Malic acid : 3.05
  - Ash : 1.88
  - Alcalinity of ash : 28.8
  - Magnesium : 101.1
  - Total phenols : 2.88
  - Flavanoids : 3.88
  - Nonflavanoid phenols : 0.44
  - Proanthocyanins :
    2.88
  - Color intensity : 8.8
  - Hue : 1.48
  - OD280/OD315 of diluted wines : 3.88
  - Proline : 888

Sử dụng model huấn luyện được cho biết mẫu rượu này thuộc loại nào?


HANOI UNIVERSITY OF MINING AND GEOLOGY
KHOA
CÔNG NGHỆ THÔNG TIN
FACULTY OF INFORMATION TECHNOLOGY

Thank you!