KHAI PHÁ DỮ LIỆU

VO DUC QUANG – VINH UNIVERSITY

QUANGVD@VINHUNI.EDU.VN

Nội dung

- Chương 1: Tổng quan về Data Mining
- Chương 2: Dữ liệu và tiền xử lý dữ liệu
- Chương 3: Bài toán phân lớp dữ liệu
- Chương 4: Bài toán phân cụm dữ liệu
- Chương 5: Khai phá luật kết hợp

Chương 3 – Phân lớp (Classification)

- Bài toán Phân lớp
- Phân lớp sử dụng Cây quyết định
- Một số giải thuật khác
  - Naïve Bayes
  - KNN
- Đánh giá mô hình phân lớp

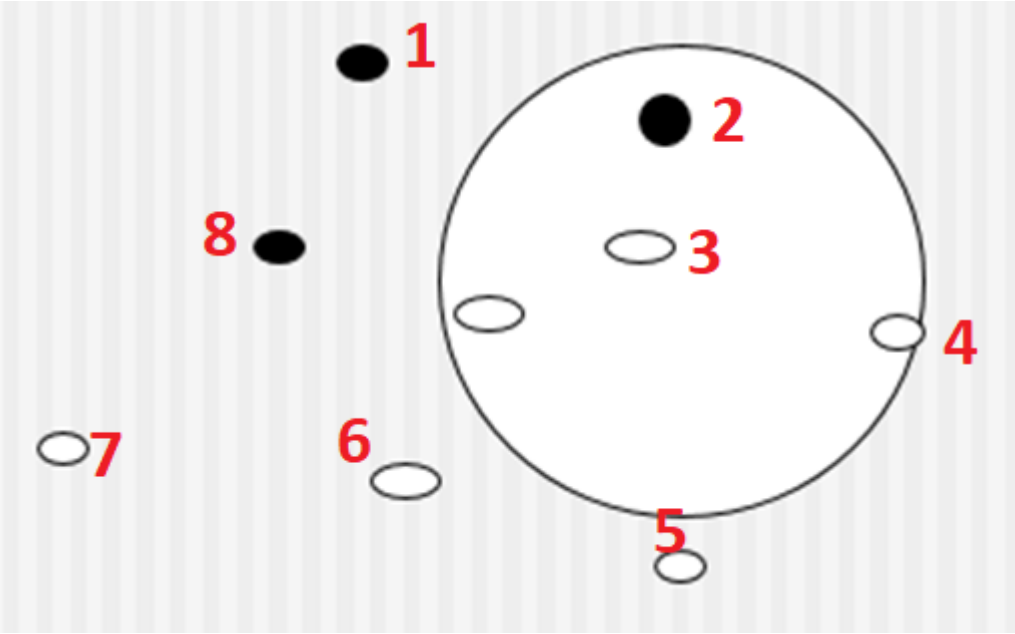
Ôn tập khái niệm

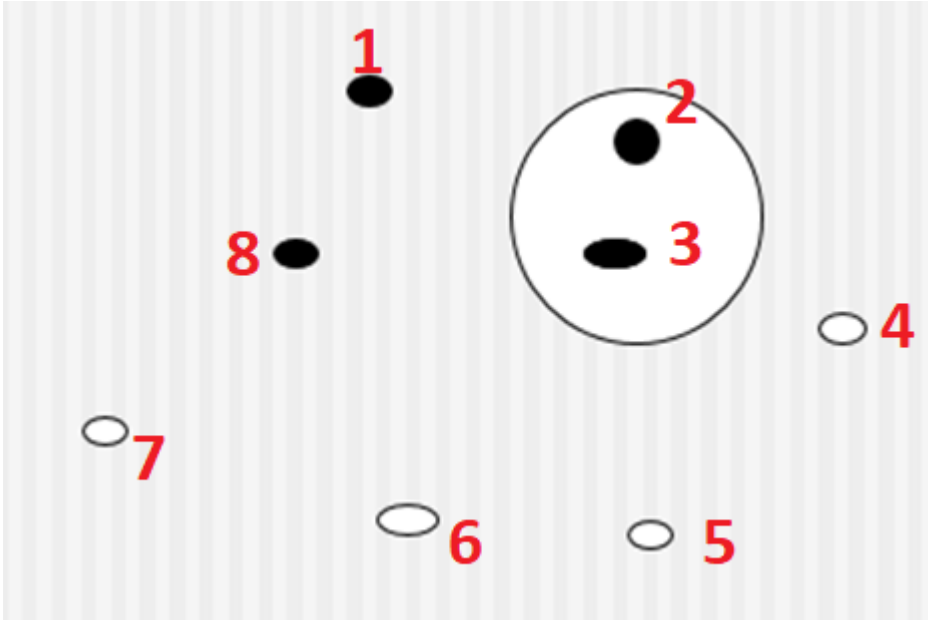
- Xem ví dụ (Machinelearningcoban.com)

Ngôn ngữ người	Ngôn ngữ Máy Học	in Machine Learning
Câu hỏi	Điểm dữ liệu	Data point
Đáp án	Đầu ra, nhãn	Output, Label
Ôn thi	Huấn luyện	Training
Tập tài liệu mang vào phòng thi	Tập dữ liệu tập huấn	Training set
Đề thi	Tập dữ liểu kiểm thử	Test set
Câu hỏi trong dề thi	Dữ liệu kiểm thử	Test data point
Câu hỏi có đáp án sai	Nhiễu	Noise, Outlier
Câu hỏi gần giống	Điểm dữ liệu gần nhất	Nearest Neighbor

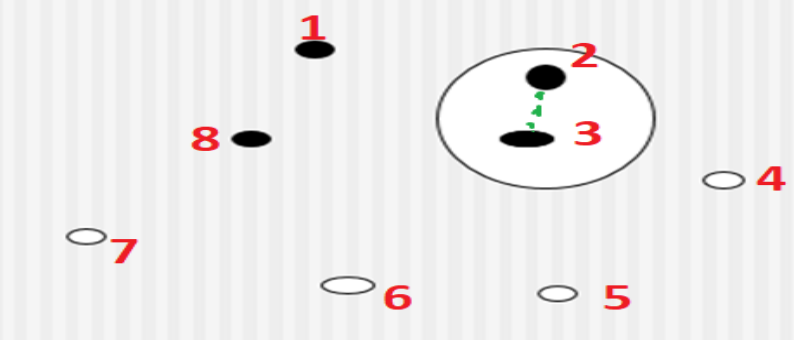
K-nearest neighbor (KNN)

- lazy learning
- Neighbor: Láng giềng
- K-nearest neighbor ?





Khoảng cách



Điểm gần nhất là 3

Vì KC(2,3) → Min

Min 1 0

Min 2 4

Min 3 0

Điểm 3 là hàng xóm của 2

$$KC(2,3) = \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2}$$

KC(2,4) = ...

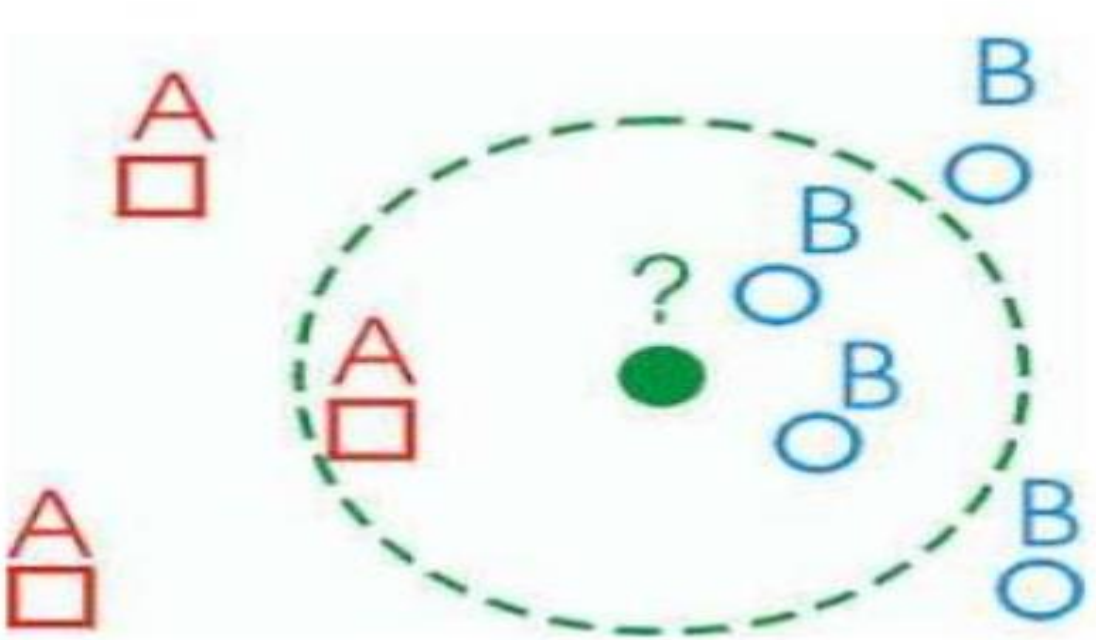
KC(2,1) = ...

Euclid

$$d(2,3) = \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2 + (z_3 - z_2)^2}$$

KNN

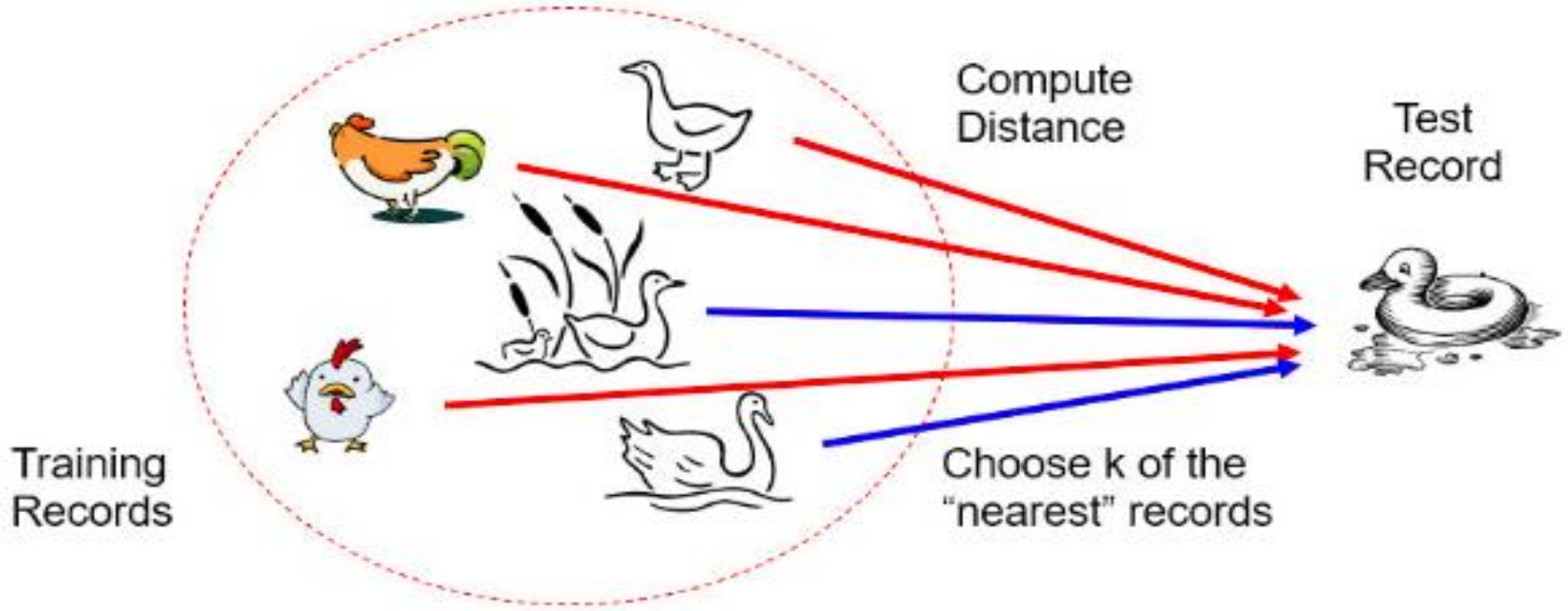
- Xác định nhãn cho một điểm?



KNN

K-Nearest Neighbors: *"Ask not what this is called, ask what this is like ?"*

- Nếu nó đi như một con vịt, lang thang như một con vịt, thì nó có thể là một con vịt.



KNN

- Demo code

Table 1. Euclidean distance matrix D listing all possible pairwise Euclidean distances between 19 samples.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_2	1.5																	
x_3	1.4	1.6																
x_4	1.6	1.4	1.3															
x_5	1.7	1.4	1.5	1.5														
x_6	1.3	1.4	1.4	1.5	1.4													
x_7	1.6	1.3	1.4	1.4	1.5	1.8												
x_8	1.5	1.4	1.6	1.3	1.7	1.6	1.4											
x_9	1.4	1.3	1.4	1.5	1.2	1.4	1.3	1.5										
x_{10}	2.3	2.4	2.5	2.3	2.6	2.7	2.8	2.7	3.1									
x_{11}	2.9	2.8	2.9	3.0	2.9	3.1	2.9	3.1	3.0	1.5								
x_{12}	3.2	3.3	3.2	3.1	3.3	3.4	3.3	3.4	3.5	3.6	1.6							
x_{13}	3.3	3.4	3.2	3.2	3.3	3.4	3.2	3.3	3.5	3.6	1.4	1.7						
x_{14}	3.4	3.2	3.5	3.4	3.7	3.5	3.6	3.3	3.5	3.6	1.5	1.8	0.5					
x_{15}	4.2	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	1.7	1.6	0.3	0.5				
x_{16}	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	1.6	1.5	0.4	0.5	0.4			
x_{17}	5.9	6.2	6.2	5.8	6.1	6.0	6.1	5.9	5.8	6.0	2.3	2.3	2.5	2.3	2.4	2.5		
x_{18}	6.1	6.3	6.2	5.8	6.1	6.0	6.1	5.9	5.8	6.0	3.1	2.7	2.6	2.3	2.5	2.6	3.0	
x_{19}	6.0	6.1	6.2	5.8	6.1	6.0	6.1	5.9	5.8	6.0	3.0	2.9	2.7	2.4	2.5	2.8	3.1	0.4

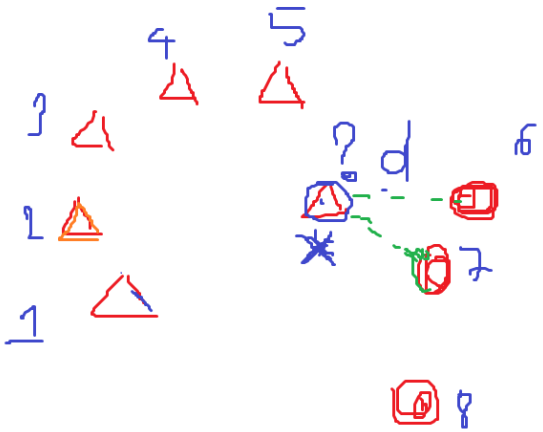
$k = 3$

x_{+k}

$\Rightarrow$ k

KNN

KNN



$= \sqrt{(\cdot)^2 + (\cdot)^2 + \dots}$

K = 7 ?

K = 6

A 3 / 3 B ?

⇒ Tiến dị phụ

- K = 1 → d(x, 7) min ⇒ C7 = tròn ⇒ X ⇒ tròn
- K = 3 ⇒ d(x, ·) → Tn; d(x, ·) → Tn; d(x, ·) → ▲ } ⇒ C3C ⇒ Tn