Lesson B03 (A)

K — Nearest Neighbor Algorithm (KNN)

K = 5, find the five closest samples.

KNN is a classification strategy based on instance-based learning (Lazy learning).

Classification.

Function: Domain → Range (Discrete)

Name	Age	Gender	Sport	Distance to Anna
Alex	32	M	Football	27.02
Mark	40	M	~~Tennis~~ ϕ	35.01
Sam	16	F	Tennis.	11
Zoe	34	F	Tennis	9
Sally	~~35~~	M	ϕ	50
Richerd	40	M	Tennis	35
Paul	20	F	ϕ	15
Smith	15	M	tennis	10
Larry	56	F	Football	50
Mike	15	M	Football	10.05
Anna	5	F	?	

Numerical Data ← Discrete data.

e.g. male = 0 : female = 1.

Find the distance between Anna to other samples.

The Distance measurements:

- Euclidean Distance $((x_1, y_1) \leftrightarrow (x_2, y_2))$

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

Other Distance:

- Manhattan Distance.
- Minkowski Distance.

Alex: male, age = 32.
Anne: femal, age = 5.

$$\text{Alex} \leftrightarrow \text{Anne} = \sqrt{(5-32)^2 + (1-0)^2}$$
$$= \sqrt{729 + 1} = 27.2.$$

Compute the distance between Anne with all other sample

$K = 3$. find the 3 closest sample

Zoe	9	$\longrightarrow$	Tennis
Smith	10.	$\longrightarrow$	Tennis
Mike	10.05	$\longrightarrow$	Football

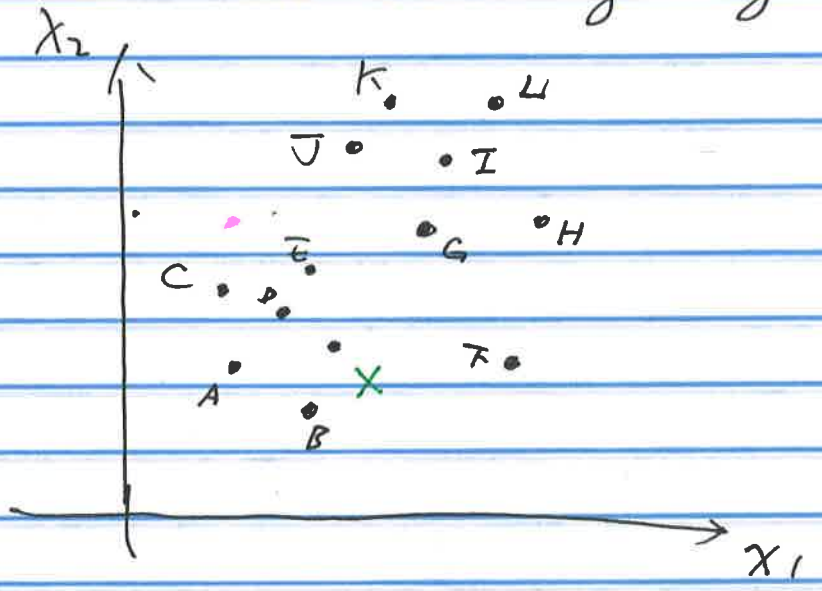
$\Longrightarrow$ Tennis dominates in 3 samples

$\Longrightarrow$ Anna will play Tennis.

Lesson B03(B)

K — Means..

(A most representative clustering algorithm)



Concept:

1. Centroid.
2. Assignment
3. Mean of Observation (New Centroid).