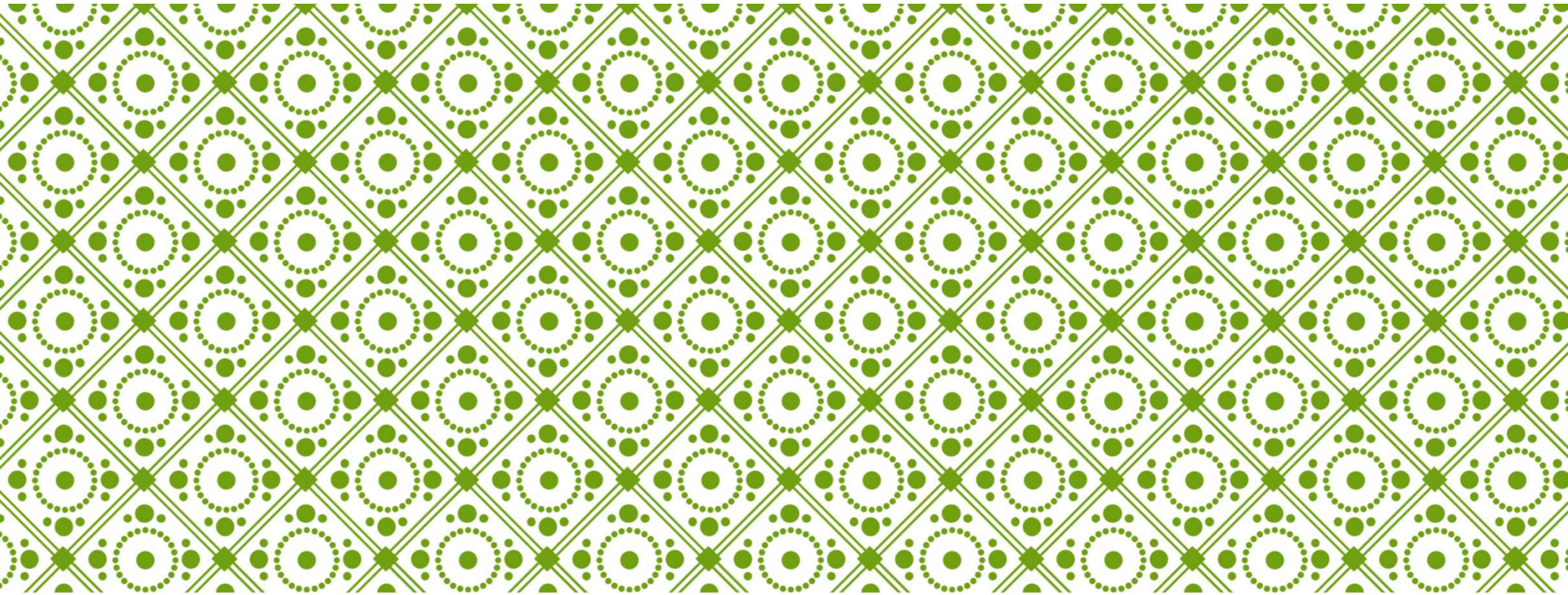


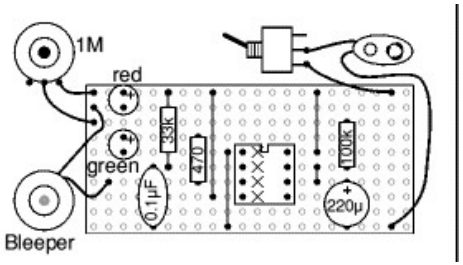
KNN AND KMEANS

INTRODUCTION TO MACHINE LEARNING

What is Pattern?

PATTERN

A pattern is the opposite of a chaos, it is an entity that can be given a name



RECOGNITION

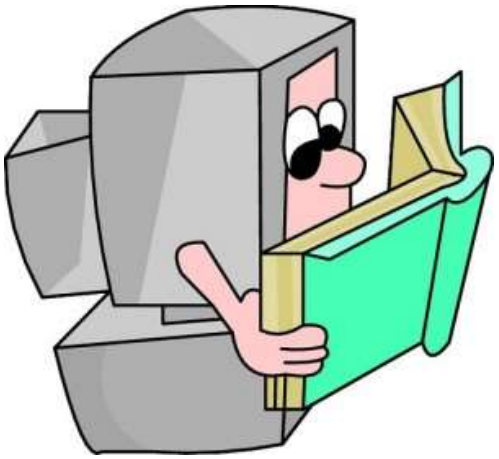
Identification of a pattern as a member of a category

- Classification (Supervised: known categories)
- Clustering (Unsupervised: learning categories)

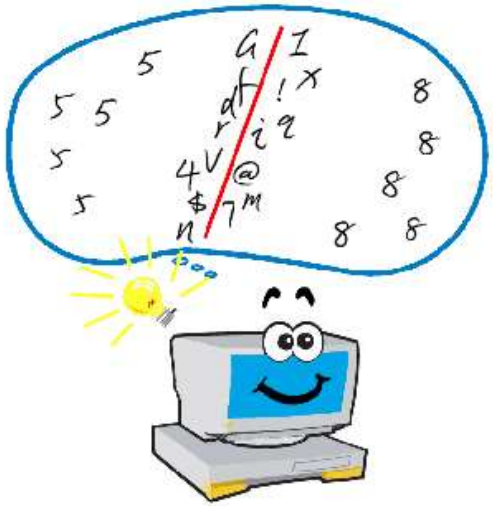
PATTERN RECOGNITION

Given an input pattern, make a decision about the "category" or "class" of the pattern

WHAT IS MACHINE LEARNING?



Make the machine '*learn*' some thing



Evaluate how good the machine has '*learned*'

MACHINE LEARNING

Field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Sameul (1959)

MACHINE LEARNING

Machine learning is programming computers to optimize a performance criterion using example data or past experience.

Tom Mitchell (1998)

LEARNING PROBLEMS – EXAMPLES

Learning = Improving with experience over some task

- Improve over task T,
- With respect to performance measure P,
- Based on experience E.

Example

- T = Play checkers
- P = % of games won in a tournament
- E = opportunity to play against itself

MACHINE LEARNING

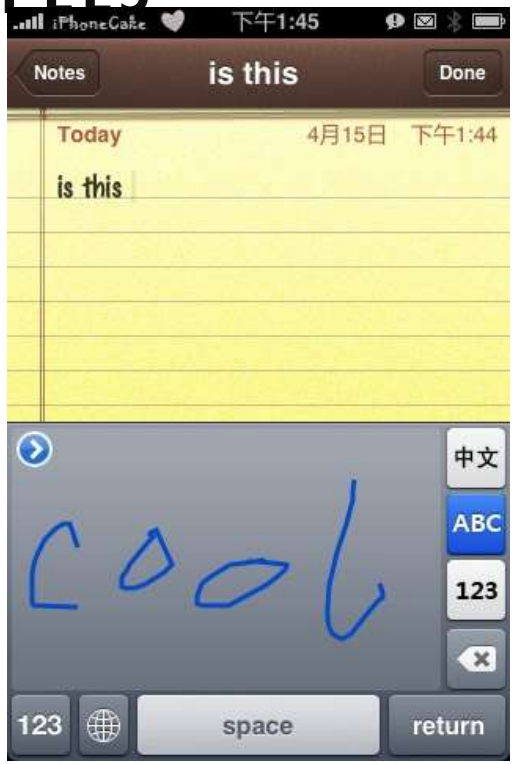
Learning = Improving with experience over some task

A computer program is said to ***learn*** from experience *E* with respect to some task *T* and performance measure *P*, if its performance at task *T*, as measured by *P*, improves with experience *E*

LEARNING PROBLEMS – EXAMPLES

Handwriting recognition learning problem

- Task *T*: recognizing handwritten words within images
- Performance measure *P*: percent of words correctly recognized
- Training experience *E*: a database of handwritten words with given classifications



LEARNING PROBLEMS – EXAMPLES

A robot driving learning problem

- Task *T*: driving on public four-lane highways using vision sensors
- Performance measure *P*: average distance traveled before an error (as judged by human overseer)
- Training experience *E*: a sequence of images and steering commands recorded while observing a human driver

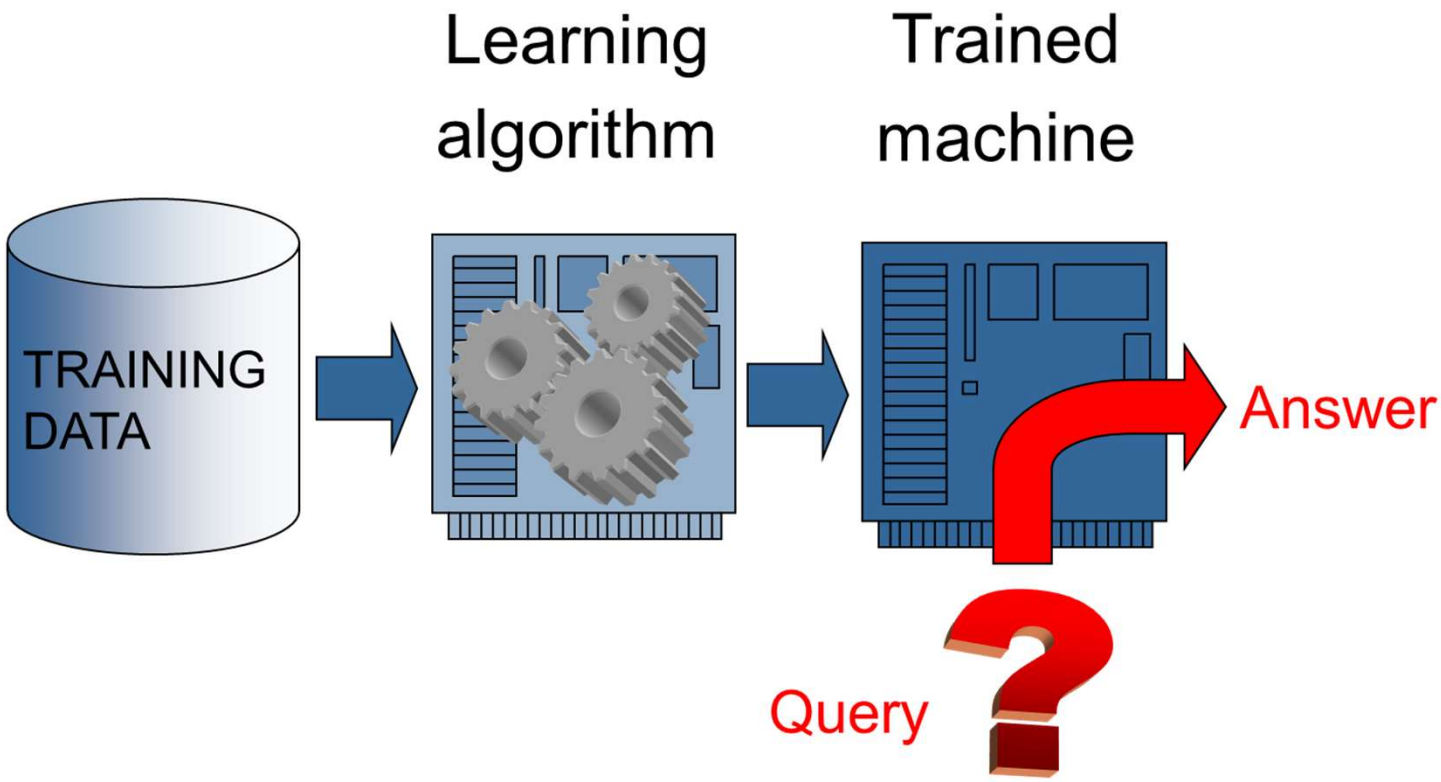
MACHINE LEARNING

There is no need to “learn” to calculate payroll

Learning is used in:

- Data mining programs that learn to detect fraudulent credit card transactions
- Programs that learn to filter spam email
- Programs that learn to play checkers/chess
- Autonomous vehicles that learn to drive on public highways
- Self customizing programs
- And many more…

MACHINE LEARNING



MACHINE LEARNING

Supervised

Unsupervised

Reinforcement

SUPERVISED LEARNING

Requires a labelled training set. A training set comprises of object belonging to different labels/classes.

Example (face recognition):

- Training Set: 10 persons or labels with 3 face images per person, each from different pose. We have 30 images in total. Labels in this case are names of each person (string).
- Testing Phase: Take a random person from those 10, take his/her image from random pose and random facial expression (sad, smily etc) and ask which person that face belongs to.

SUPERVISED LEARNING

Example (regression, estimating price of a plot):

- Training Set: Imagine from pwd housing society, we have 10 plots and for each plot we know its covered area in square meters alongwith the price. Price in this case in the label (floating number)
- Testing Phase: Estimate the price (label) of a random plot given its covered area.

UNSUPERVISED LEARNING

No labelled training set required.

Example (Clustering):

- Imagine you have 50 students and for each student you know his/her marks out of 100. Divide these 50 students in good, average and worst categories based upon their marks.
- Label in this case is one of the three strings (“good”, “average” or “bad”). Here you have a training set (a list of marks of those 50 students) but you don’t know the label of each of that student (whether that student belongs to a “good” category or one of the other two. That is what you have to find out!

REINFORCEMENT LEARNING

In **reinforcement learning** the agent learns from a series of reinforcements—rewards or punishments.

For example, the lack of a tip at the end of the journey gives the taxi agent an indication that it did something wrong. The two points for a win at the end of a chess game tells the agent it did something right.

SUPERVISED - KNN

KNN CLASSIFICATION

Given a training dataset with classes i.e. data and their labels are known

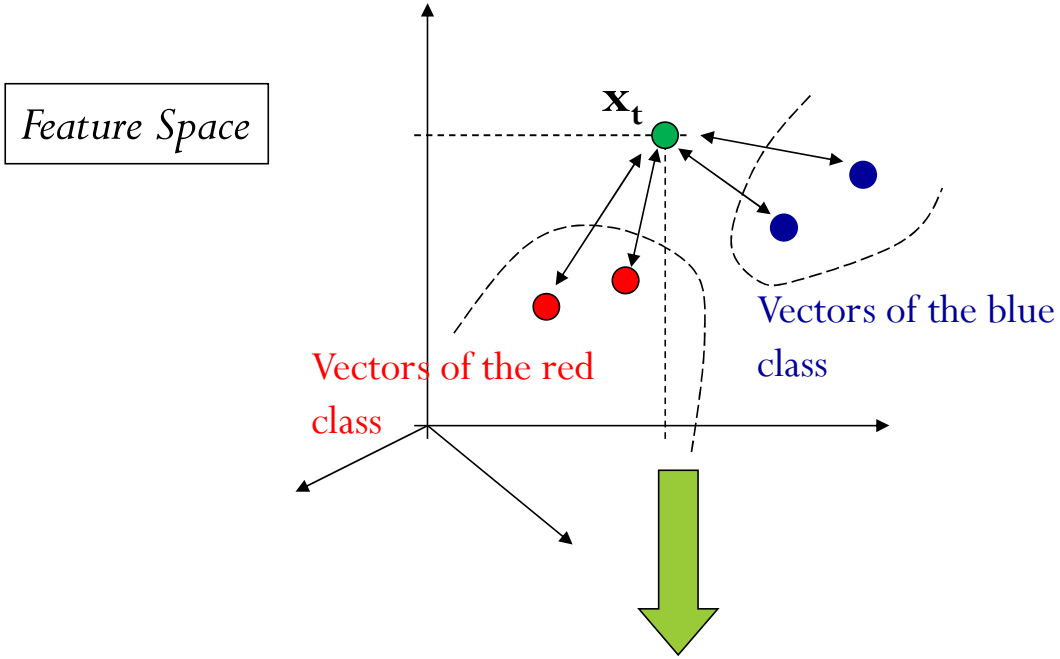
$$\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), (\mathbf{x}_3, \mathbf{y}_3), \cdots\cdots, (\mathbf{x}_n, \mathbf{y}_n)\}, \quad \mathbf{x} \in \mathbb{R}^d$$

Objective: Decide the class (or label) of a test vector $\mathbf{x}_t$

Method:

- Compute the distance between the vector $\mathbf{x}_t$ and all examples of the training dataset
- Decision: assign the class of the nearest neighbor (closest point) from training dataset to the vector $\mathbf{x}_t$

KNN CLASSIFICATION



Class ($\mathbf{x}_t$)

=

Class of the nearest neighbor

K-NEAREST NEIGHBOR RULE

Given a training dataset with classes i.e. data and their labels are known

$$\left\{\left(\mathbf{x}_1, \mathbf{y}_1\right), \left(\mathbf{x}_2, \mathbf{y}_2\right), \left(\mathbf{x}_3, \mathbf{y}_3\right), \cdots\cdots, \left(\mathbf{x}_n, \mathbf{y}_n\right)\right\}, \quad \mathbf{x} \in \square^{d}$$

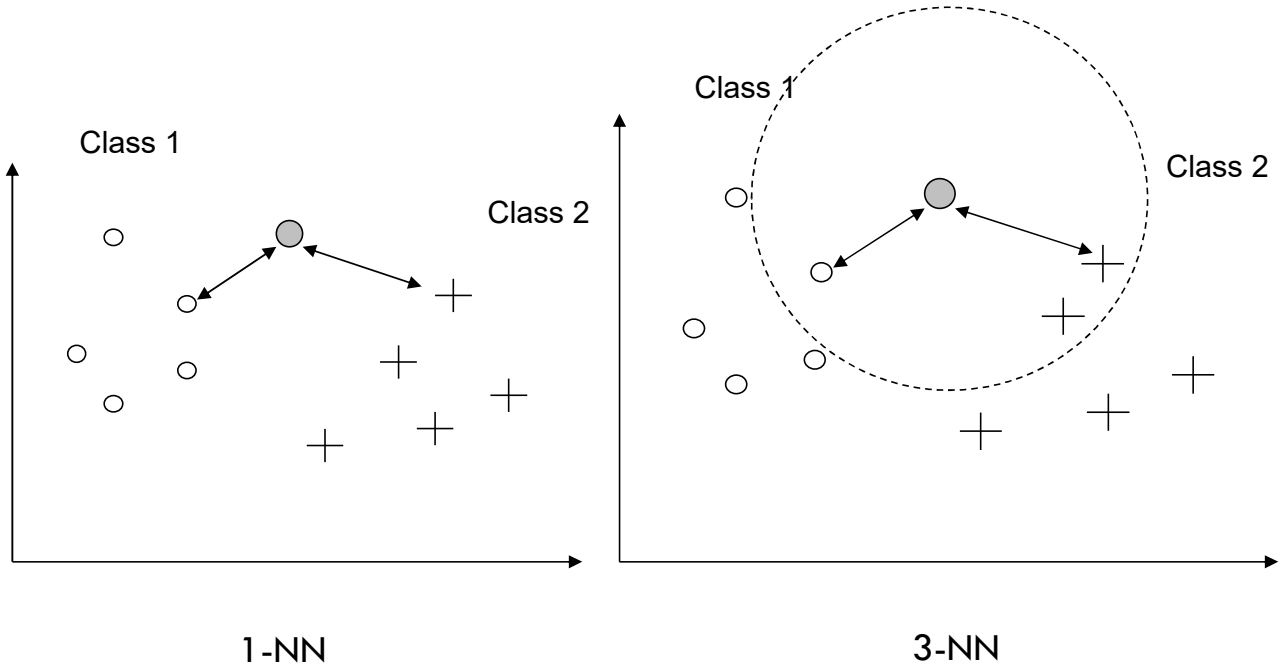
Objective: Decide the class (or label) of a test vector $\mathbf{x}_t$

Method:

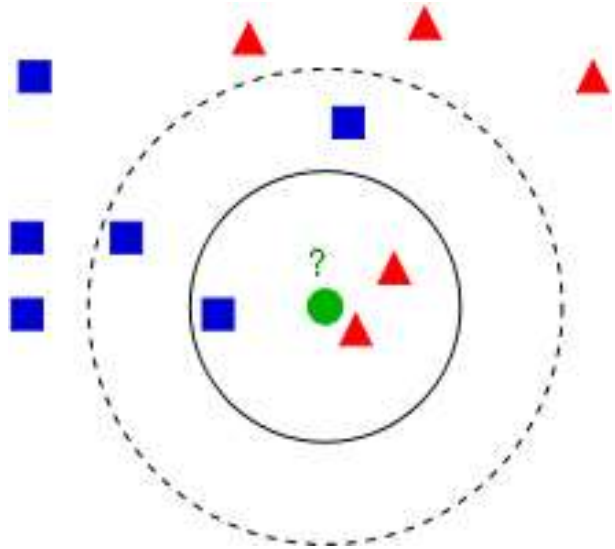
- Compute the distances between the vector $\mathbf{x}_t$ and all examples of the training dataset
- Determine k nearest neighbors of $\mathbf{x}_t$
- Decision: class($\mathbf{x}_t$) = majority vote of k nearest neighbors
- The choice of distance is very important

K-NEAREST NEIGHBOR RULE

Example



K-NEAREST NEIGHBOR RULE



The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If $k = 3$ it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If $k = 5$ it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

Image Source: Wikipedia

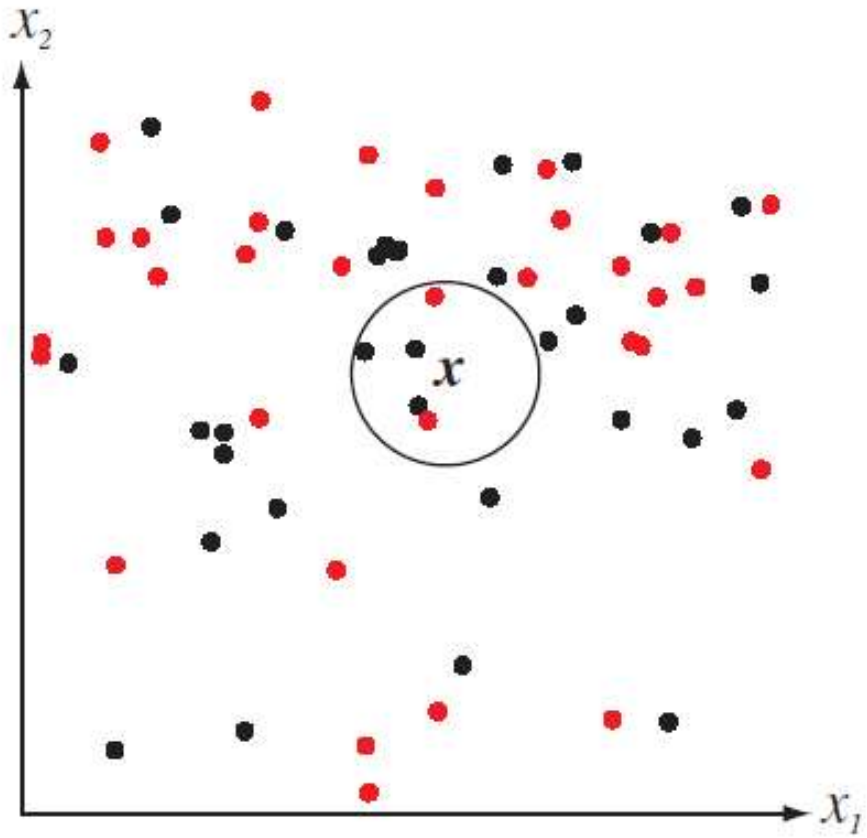


Figure 4.15: The k-nearest-neighbor query starts at the test point and grows a spherical region until it encloses k training samples, and labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point x would be labelled the category of the black points.

DISTANCE METRIC

A distance metric $D(.,.)$ is merely a function that gives generalized distance between two patterns

For three vectors **a, b** and **c,** a distance metric should hold following properties

Non-negativity:

$$D(\mathbf{a},\mathbf{b}) \geq 0$$

Reflexivity:

$$D(\mathbf{a},\mathbf{b}) = 0 \text{ if and only if } \mathbf{a} = \mathbf{b}$$

Symmetry:

$$D(\mathbf{a},\mathbf{b}) = D(\mathbf{b},\mathbf{a})$$

Triangle inequality:

$$D(\mathbf{a},\mathbf{b}) + D(\mathbf{b},\mathbf{c}) \geq D(\mathbf{a},\mathbf{c})$$

DISTANCE METRIC

Euclidean distance possesses all the four properties

$$D(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\| = \left(\sum_{k=1}^{d}(a_k - b_k)^2\right)^{1/2}$$

Euclidean distance in d-dimensions

DISTANCE METRIC

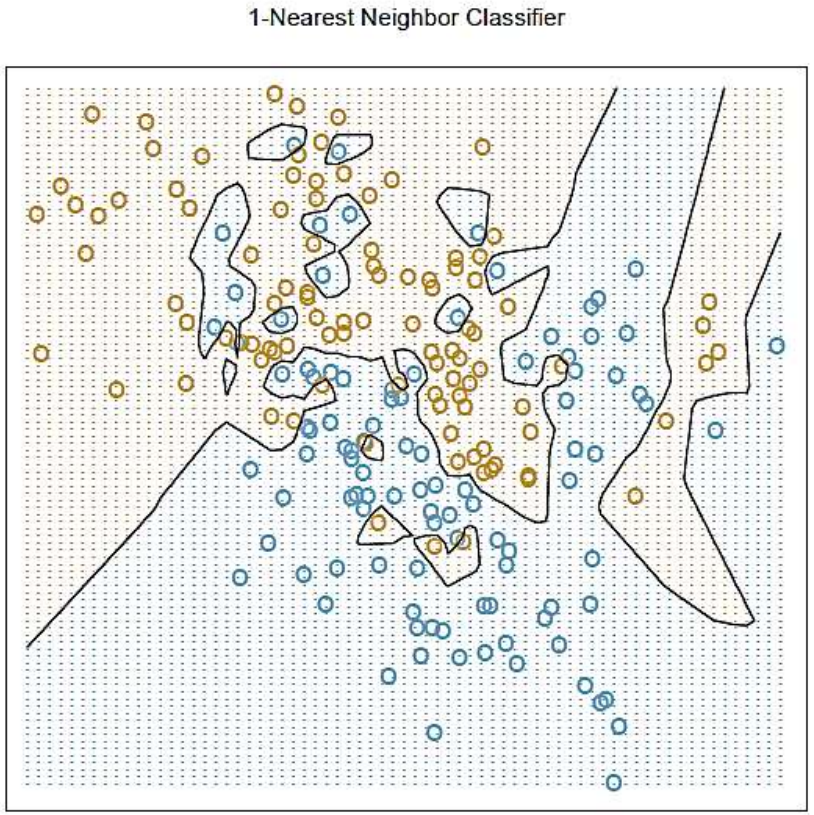
Data normalization

- If there is a large disparity in the ranges of the full data in each dimension, a common procedure is to rescale all the data to equalize such ranges

Euclidean distance is frequently employed as the distance metric in nearest neighbor classifier

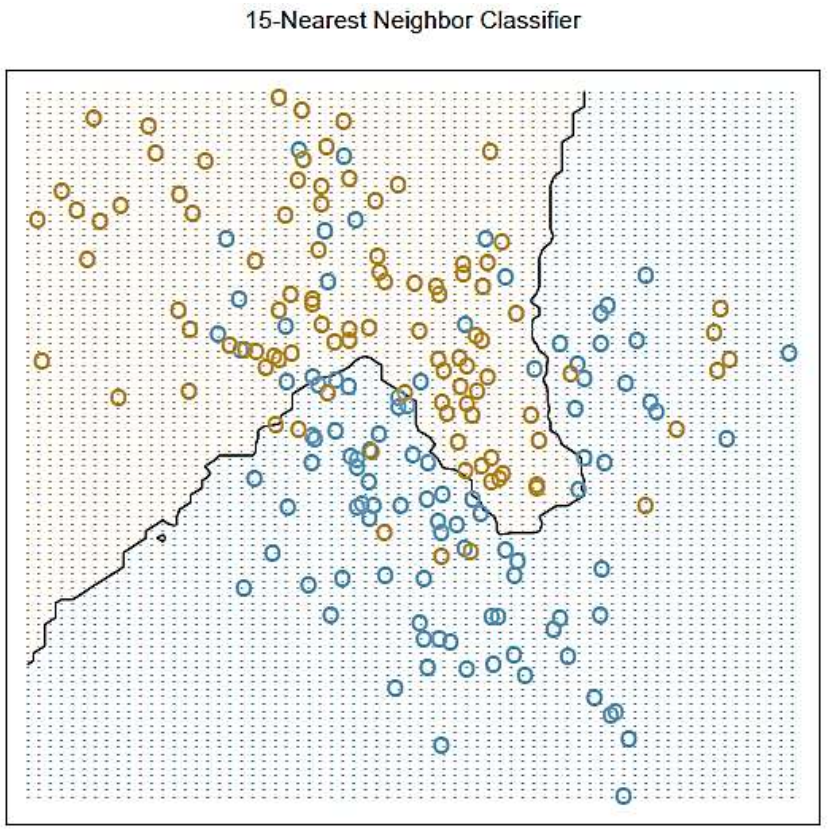
K-NEAREST NEIGHBOR CLASSIFIER EXAMPLE:

k =1
Decision boundary is complex



K-NEAREST NEIGHBOR CLASSIFIER EXAMPLE:

k =15
Decision boundary is relatively simple



Example

We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7. Without another expensive survey, can we guess what the classification of this new tissue is?

1. Determine parameter K = number of nearest neighbors

Suppose use K = 3

2. Calculate the distance between the query-instance and all the training samples

Coordinate of query instance is (3, 7), instead of calculating the distance we compute square distance which is faster to calculate (without square root)

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	**Square Distance to query instance (3, 7)**
7	7	$\mathbf{(7-3)^2+(7-7)^2=16}$
7	4	$(7-3)^2+(4-7)^2=25$
3	4	$(3-3)^2+(4-7)^2=9$
1	4	$(1-3)^2+(4-7)^2=13$

3. *Sort the distance and determine nearest neighbors based on the K-th minimum distance*

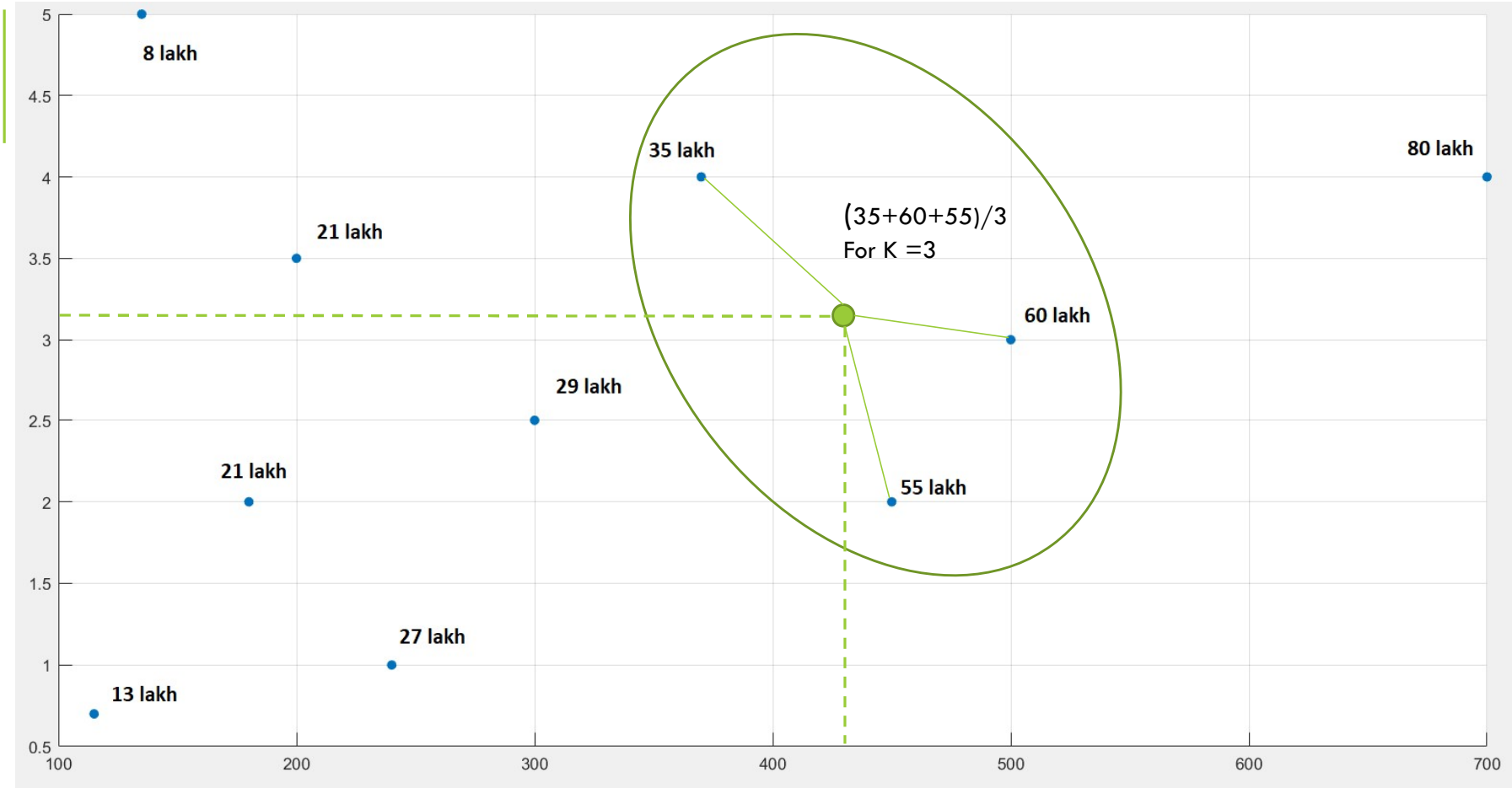
X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Square Distance to query instance (3, 7)	**Rank minimum distance**	**Is it included in 3-Nearest neighbors?**
7	7	$(7-3)^2+(7-7)^2=16$	**3**	**Yes**
7	4	$(7-3)^2+(4-7)^2=25$	**4**	**No**
3	4	$(3-3)^2+(4-7)^2=9$	**1**	**Yes**
1	4	$(1-3)^2+(4-7)^2=13$	**2**	**Yes**

4. Gather the category Y of the nearest neighbors. Notice in the second row last column that the category of nearest neighbor (Y) is not included because the rank of this data is more than 3 (=K).

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Square Distance to query instance (3, 7)	Rank minimum distance	Is it included in 3-Nearest neighbors?	**Y = Category of nearest Neighbor**
7	7	$(7-3)^2+(7-7)^2=16$	3	Yes	**Bad**
7	4	$(7-3)^2+(4-7)^2=25$	4	No	-
3	4	$(3-3)^2+(4-7)^2=9$	1	Yes	**Good**
1	4	$(1-3)^2+(4-7)^2=13$	2	Yes	**Good**

5. *Use simple majority of the category of nearest neighbors as the prediction value of the query instance*

We have 2 good and 1 bad, since 2>1 then we conclude that a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7 is included in **Good** category.


8 lakh
35 lakh
80 lakh
(35+60+55)/3
For K =3
21 lakh
60 lakh
29 lakh
21 lakh
55 lakh
27 lakh
13 lakh
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
100
200
300
400
500
600
700

K MEANS CLUSTERING

1. Choose the number (K) of clusters and randomly select the centroids of each cluster.
2. For each data point:
   I. Calculate the distance from the data point to each cluster.
   II. Assign the data point to the closest cluster.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until there is no further change in the assignment of data points (or in the centroids).

ANIMATION
1

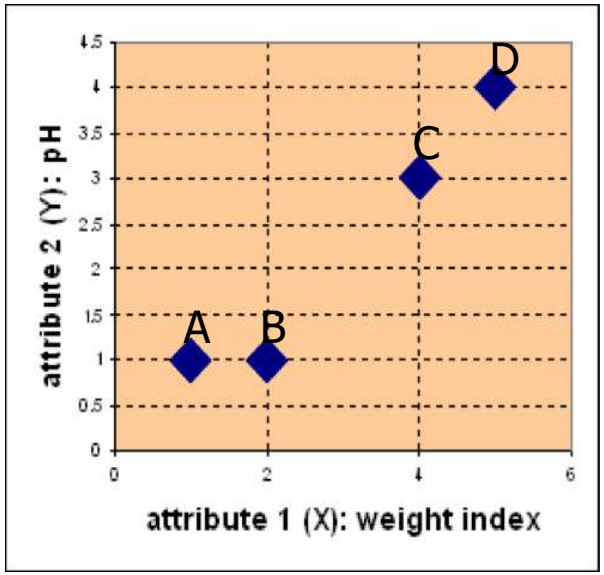
ANIMATION
2

ANIMATION
3

K MEANS – EXAMPLE 2

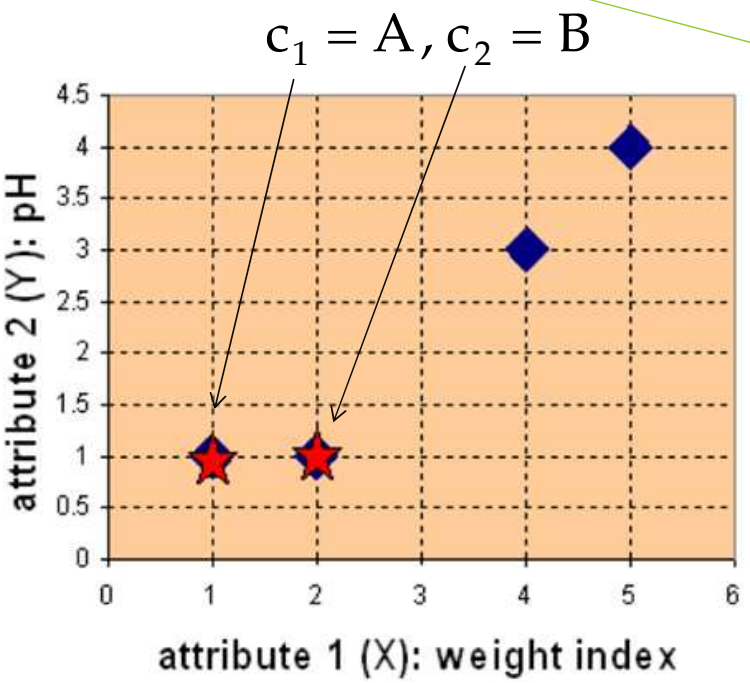
- Suppose we have 4 medicines and each has two attributes (pH and weight index). Our goal is to group these objects into K=2 clusters of medicine

Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4



K MEANS – EXAMPLE 2

- Compute the distance between all samples and *K* centroids



$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & & & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

$$d(D,c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

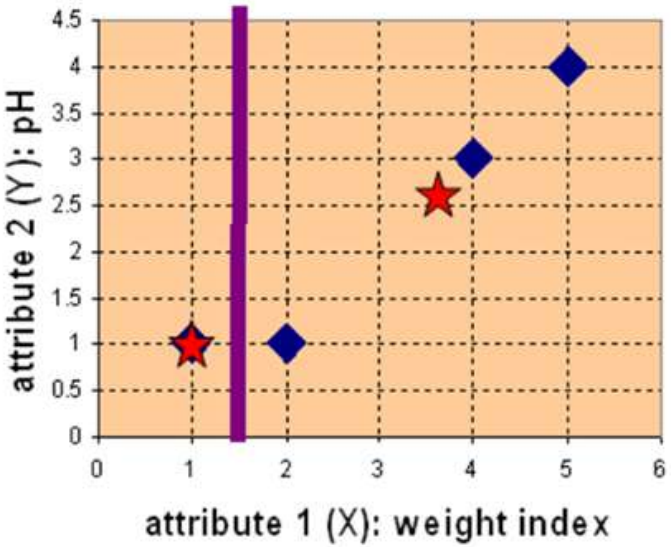
$$d(D,c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

K MEANS – EXAMPLE 2

- Assign the sample to its closest cluster

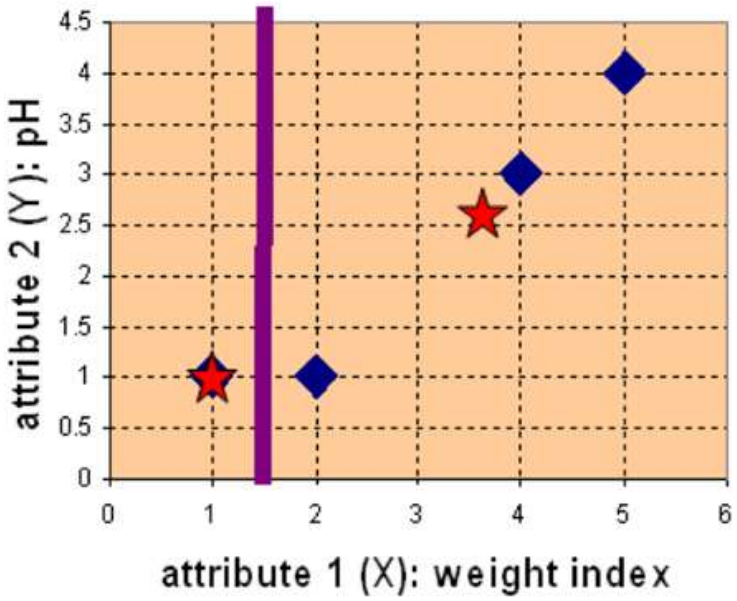
- An elements in a row of the Group matrix below is 1 if and only if the object is assigned to that group

$$\mathbf{G}^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} group-1 \\ group-2 \end{matrix}$$
$$A \quad B \quad C \quad D$$



K MEANS – EXAMPLE 2

- Re-calculate the K-centroids



Knowing the members of each cluster, now we compute the new centroid of each group based on these new memberships.

$$c_1 = (1,\ 1)$$

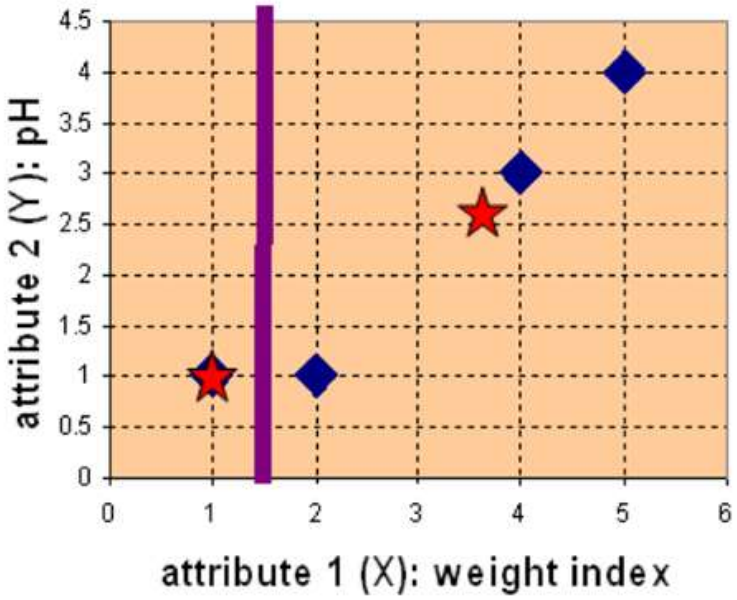
$$c_2 = \left(\frac{2+4+5}{3},\ \frac{1+3+4}{3}\right)$$

$$= (11/3,\ 8/3)$$

$$= (3.67,\ 2.67)$$

K MEANS – EXAMPLE 2

Repeat the above steps



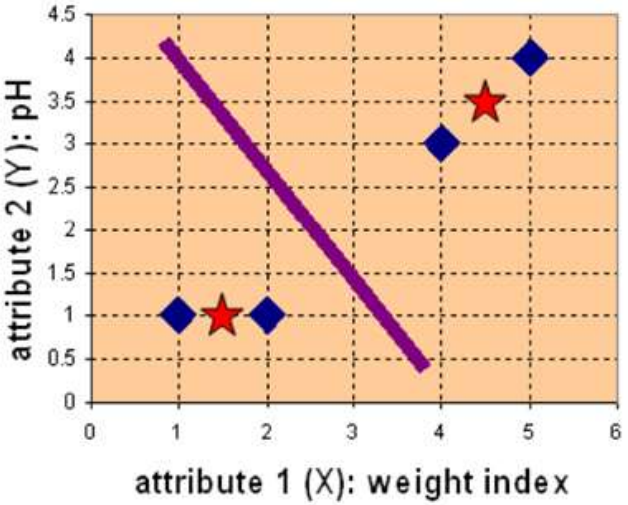
Compute the distance of all objects to the new centroids

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

K MEANS – EXAMPLE 2

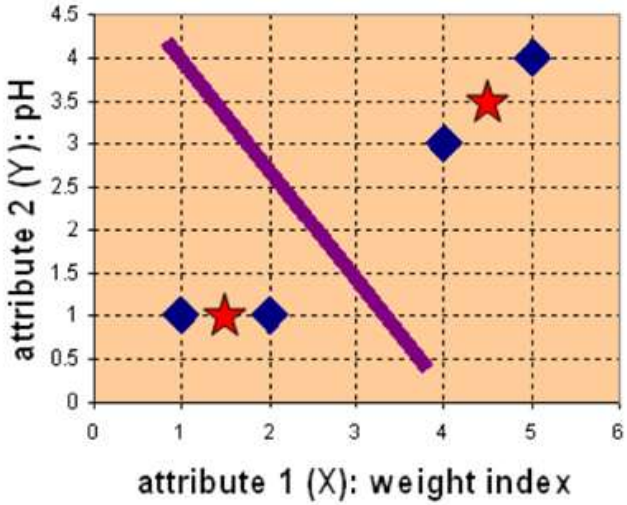


Assign the membership to objects

$$\mathbf{G}^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$A \quad B \quad C \quad D$$

K MEANS – EXAMPLE 2

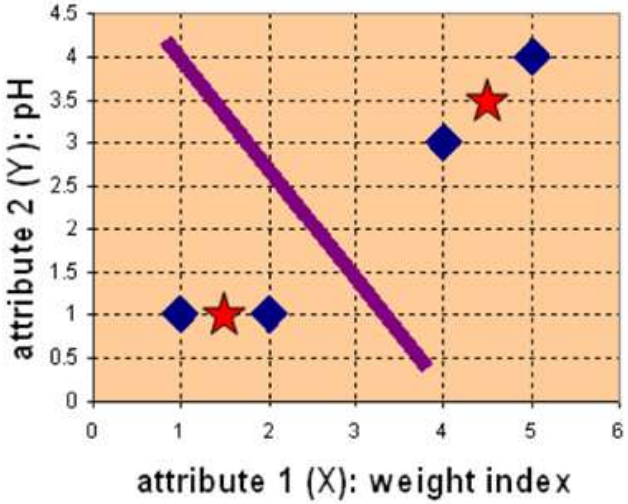


Knowing the members of each cluster, now we compute the new centroid of each group based on these new memberships.

$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1\frac{1}{2}, 1)$$

$$c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4\frac{1}{2}, 3\frac{1}{2})$$

K MEANS – EXAMPLE 2

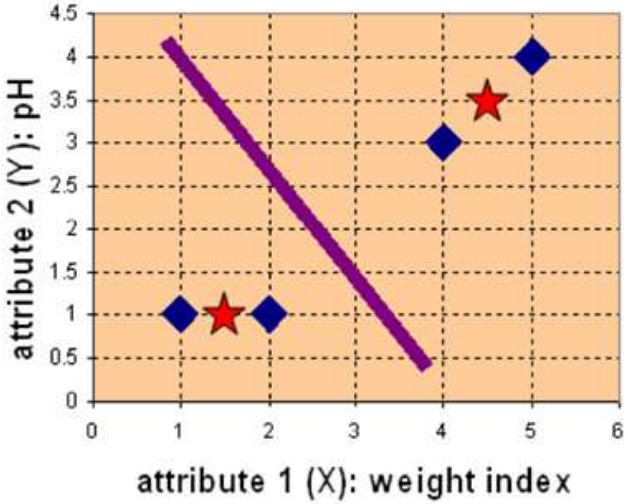


$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1\frac{1}{2}, 1) & group-1 \\ \mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

K MEANS – EXAMPLE 2



$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$A \quad B \quad C \quad D$$

K MEANS – EXAMPLE 2

We obtain result that $G^2=G^1$. Comparing the grouping of last iteration and this iteration reveals that the objects do not move group anymore.

Thus, the computation of the *k-mean* clustering has reached its stability and no more iterations are needed..

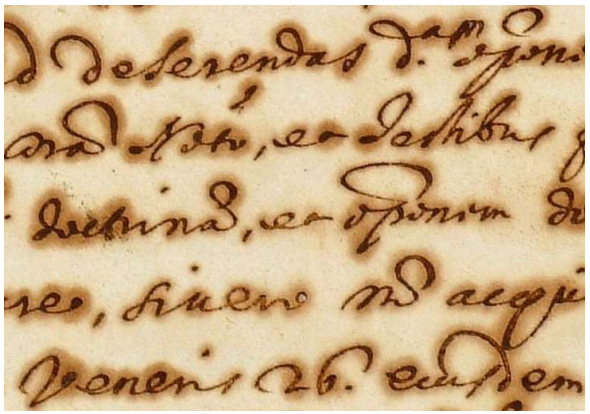
KMEANS - EXAMPLES

- Data Points – RGB Values of pixels
- Can be used for Image Segmentation

D. Comaniciu and P. Meer, *Robust Analysis of Feature Spaces: Color Image Segmentation*, 1997.

KMEANS - EXAMPLES

- Extraction of text in degraded documents



Original Image

Kmeans with k=3

KMEANS - EXAMPLES

Original

K=5

K=11

KMEANS - EXAMPLES

Quantization of colors