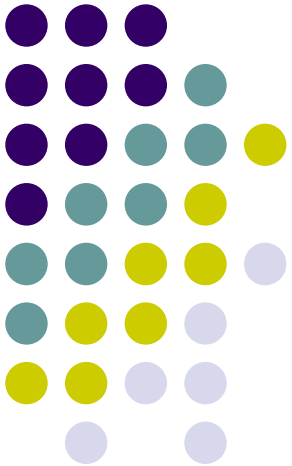
Clustering

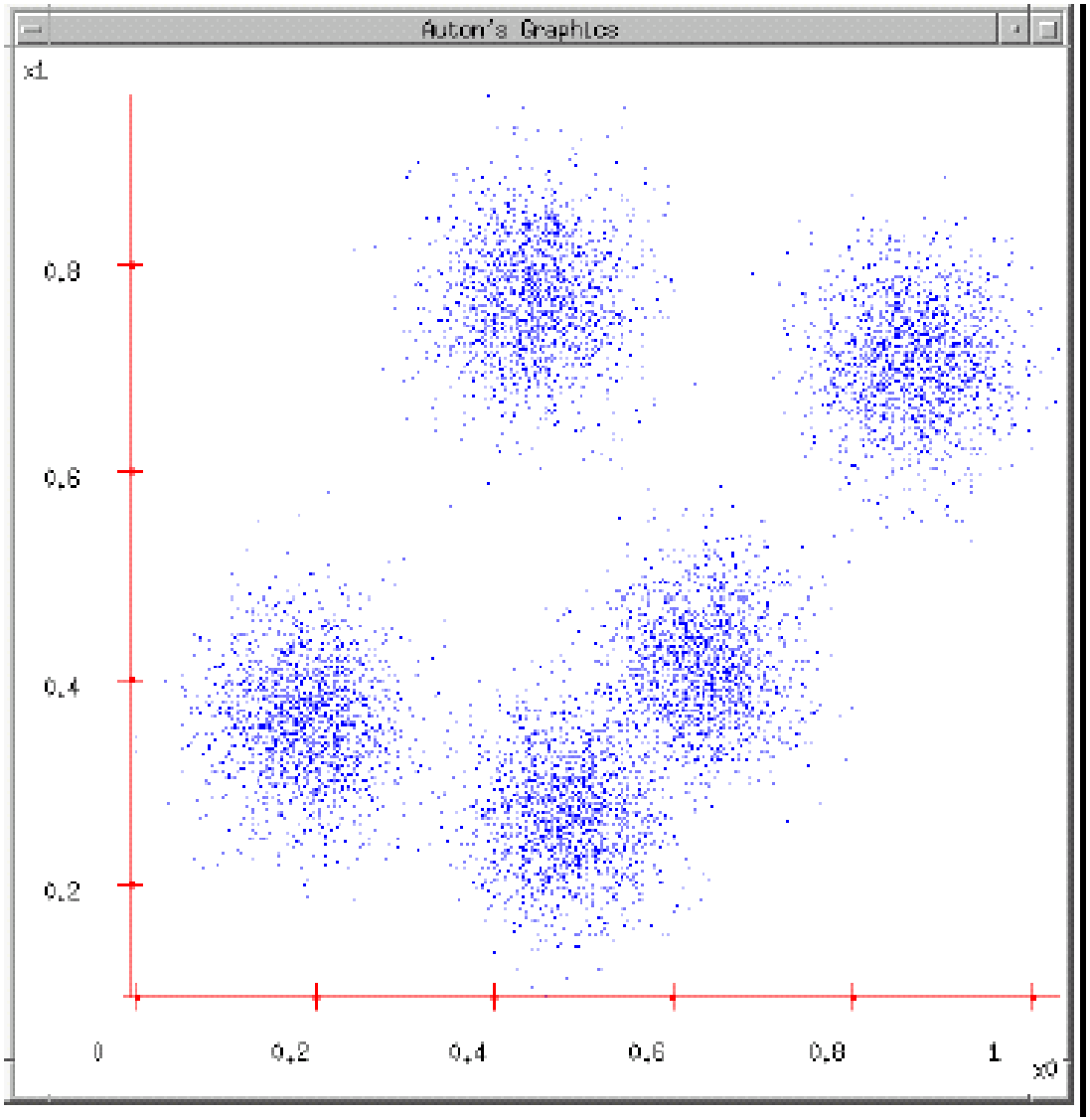
a.j.m.m. (ton) weijters

The main idea is to define k centroids, one for each cluster

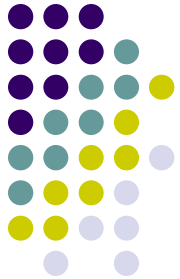
(Example from a K-clustering tutorial of Teknomo, K.

http://people.revoledu.com/kardi/tutorial/index.html)



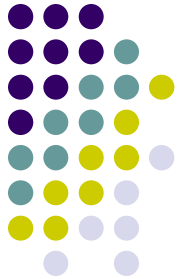

Auton's Graphics
x1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
x0

Example



- A trainer of a running group has 220 runners. For practical reasons, he likes to split up the group in 8 homogenous sub groups that can perform more or less the same training program.
- Think about relevant properties:
  - Running distance during Cooper test
  - Weight
  - ...

K-means clustering



- K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

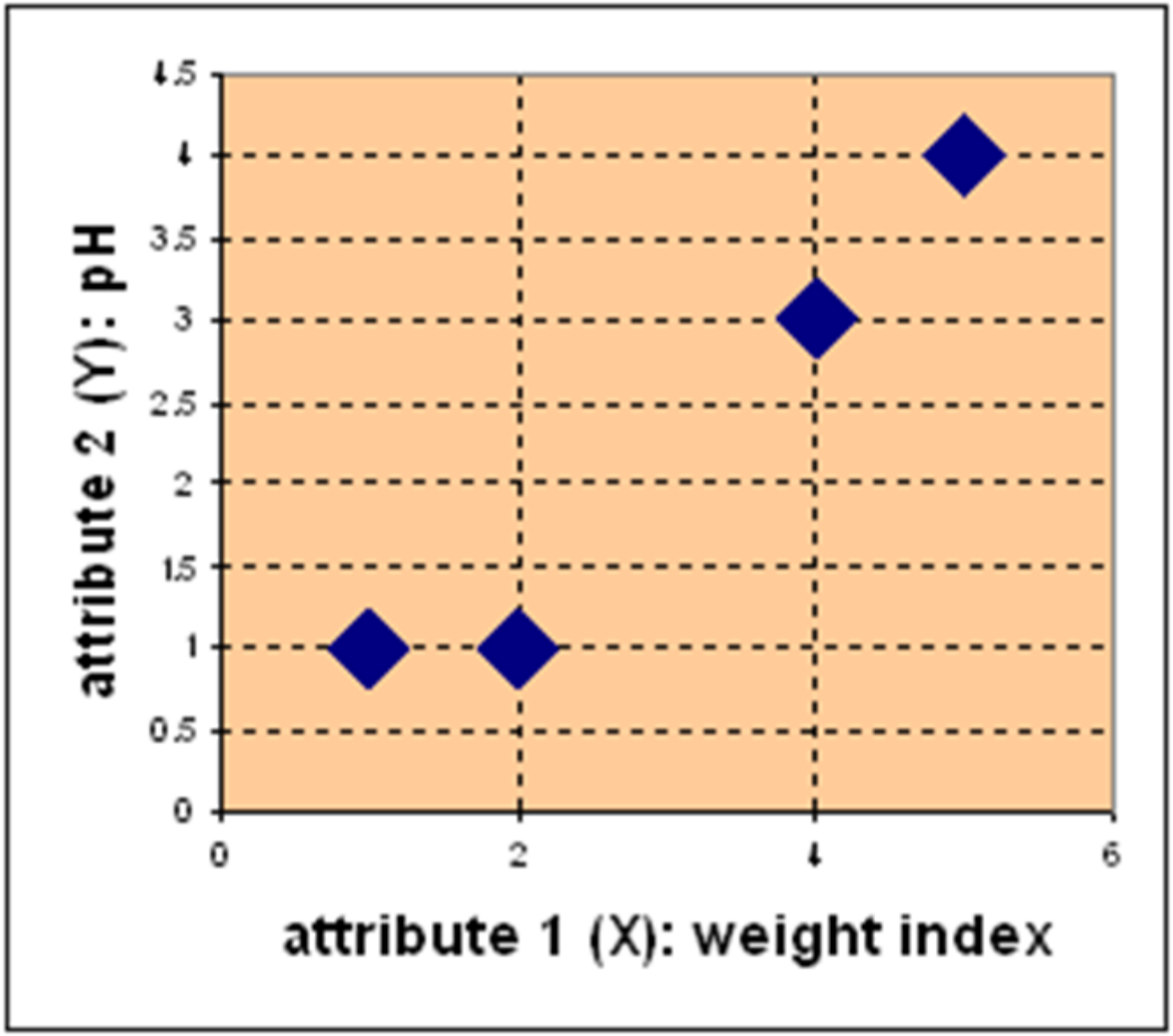
Steps of the algorithm

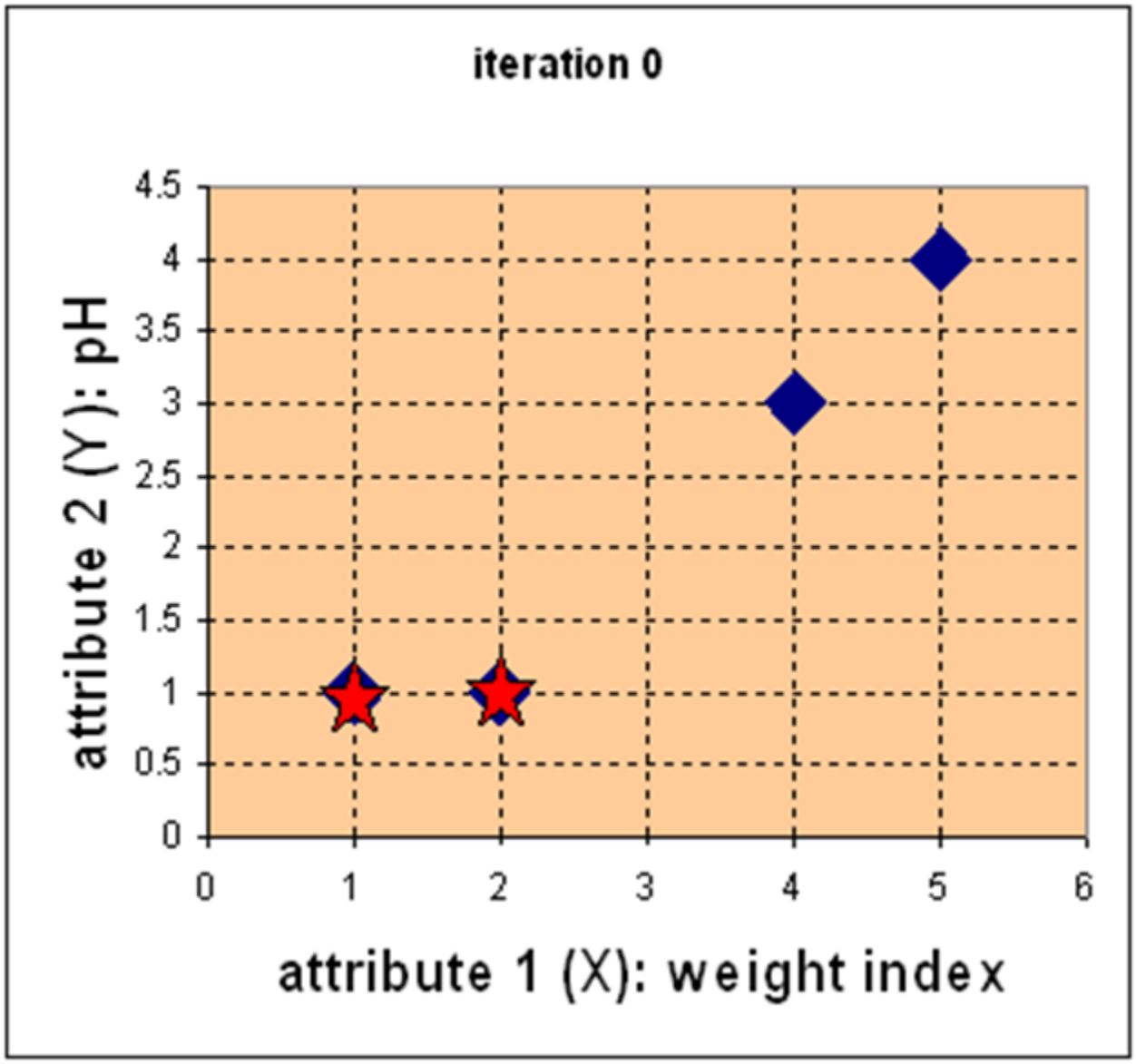
- Determine K centroids (randomly?)
- Iterate until *stable* (= no object move group)
  - Determine the distance of each object to the centroids
  - Group the object based on minimum distance
  - When all objects have been assigned, recalculate the positions of the K centroids

Example

- The numerical example below is given to understand this simple iteration

Object	attribute 1 (X):	attribute 2 (Y):
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

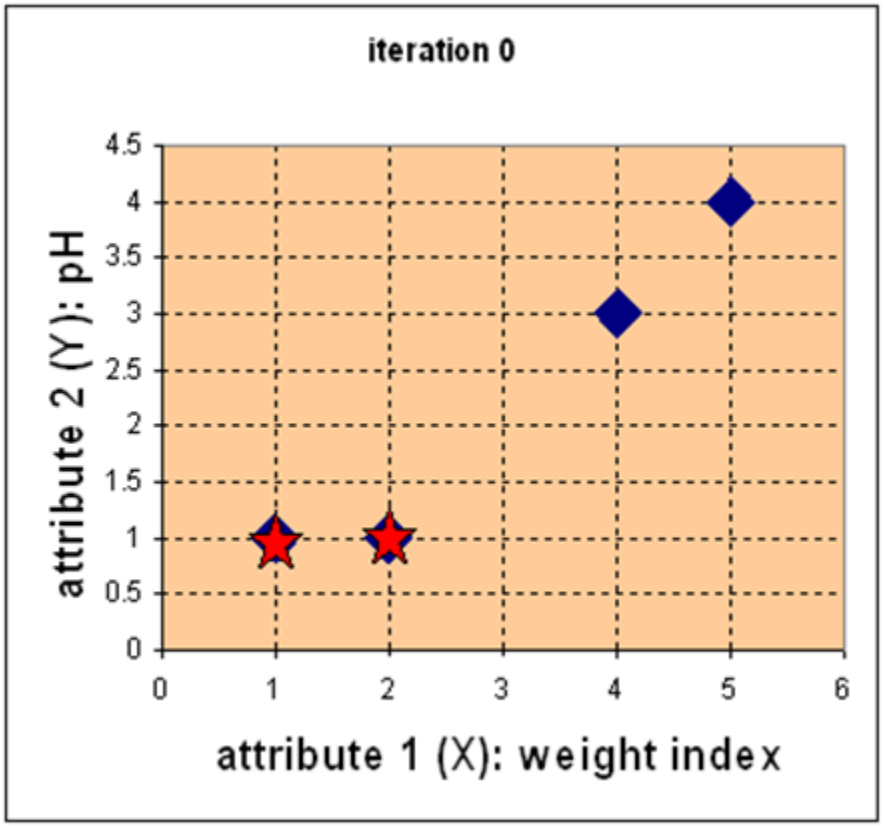

4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
2
4
6
attribute 2 (Y): pH
attribute 1 (X): weight index



K=2

Centronic 1 = (1,1)

Centronic 2 = (2,1)

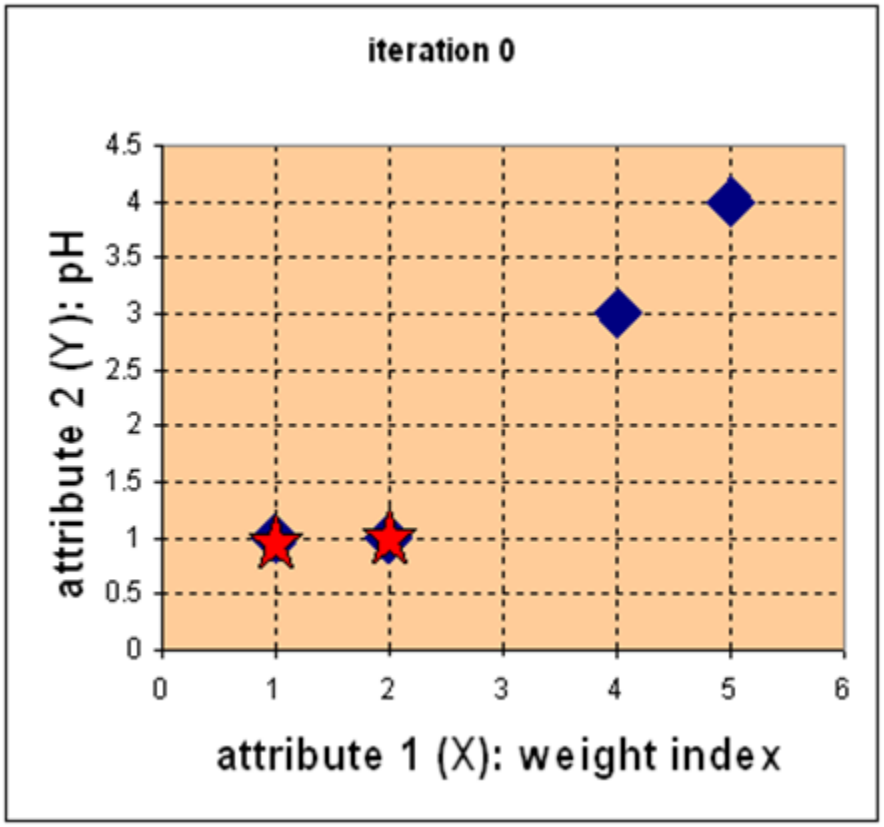


For example, distance from medicine C = (4, 3) to the first centroid is $\sqrt{(4-1)^2+(3-1)^2} = 3.61$

and its distance to the second centroid is,

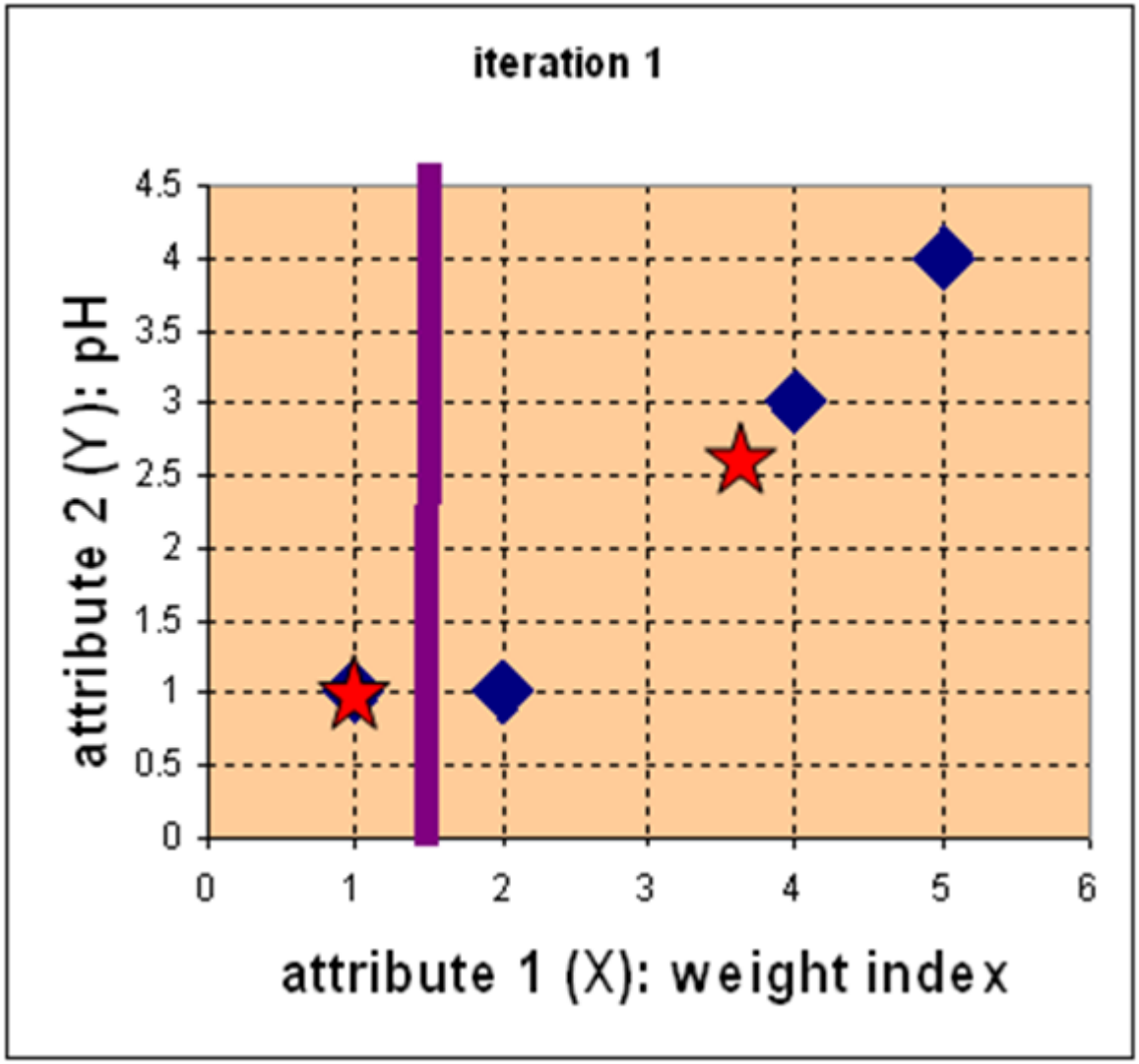
etc. $\sqrt{(4-2)^2+(3-1)^2} = 2.83$

Medicine C belongs to centroid C2

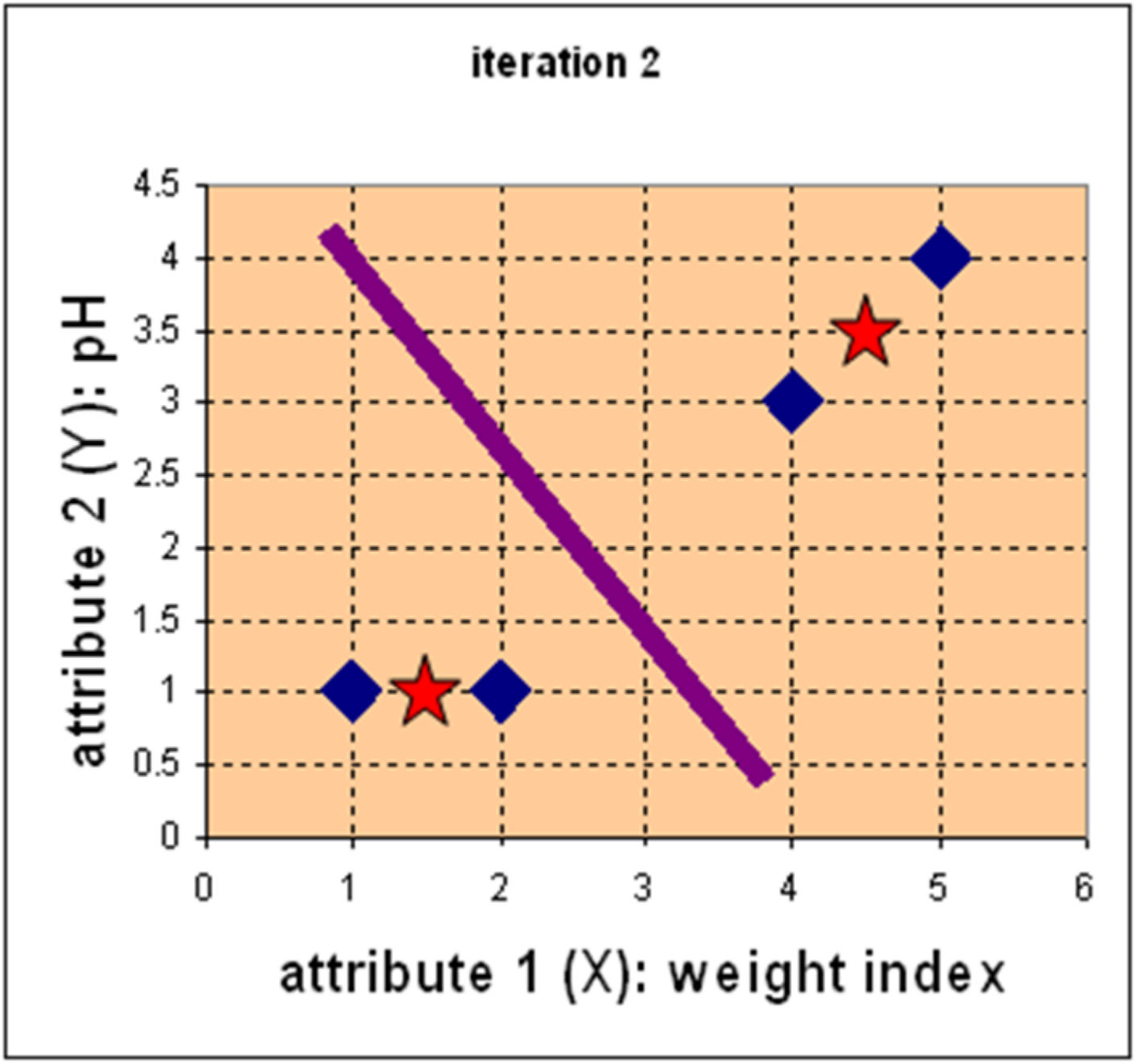


The two groups are C1 = {A}, C2 ={B,C,D}
Calculate new C1 and C2. New C1 = old C1

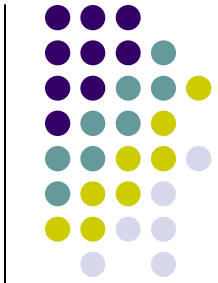
$$C2 = (\frac{2+4+5}{3}, \frac{1+3+4}{3}) = (\frac{11}{3}, \frac{8}{3})$$


iteration 1
attribute 2 (Y): pH
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6
attribute 1 (X): weight index

C1={A,B}, C2={C,D}


iteration 2
attribute 2 (Y): pH
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6
attribute 1 (X): weight index

Important Issues



- Normalization (age, weight, distance Cooper-test)
- Nominal attributes (male, female)
- Weighting