Machine Learning

- Lecture 8:
  - Clustering
    - k-mean Clustering
    - Fuzzy k-mean clustering

Clustering

- What is clustering?
- A statistical technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics.

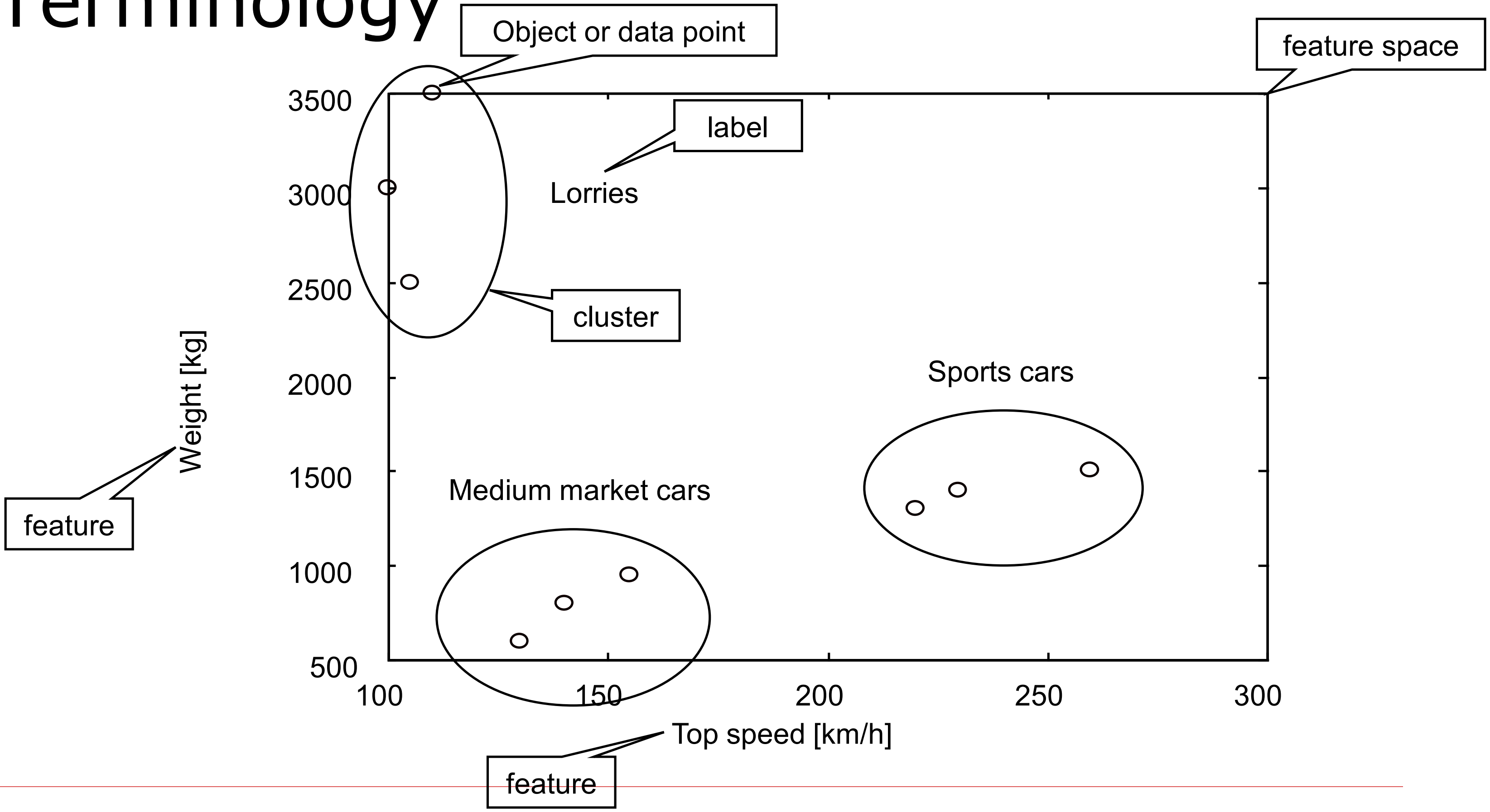
Clustering

- Example

Vehicle	Top speed km/h	Colour	Air resistance	Weight Kg
V1	220	red	0.30	1300
V2	230	black	0.32	1400
V3	260	red	0.29	1500
V4	140	gray	0.35	800
V5	155	blue	0.33	950
V6	130	white	0.40	600
V7	100	black	0.50	3000
V8	105	red	0.60	2500
V9	110	gray	0.55	3500

Clustering

- Terminology



k-mean Clustering

- What is k-mean clustering?
  - An algorithm to group some objects based on attributes/features into ***k*** number of group.
  - ***k*** is positive integer number.
  - The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

k-mean Clustering

- Example:
  Suppose we have 4 objects as training data points and each object have 2 attributes.

Medicine	Attrib1	Attrib2
A	1	1
B	2	1
C	4	3
D	5	4

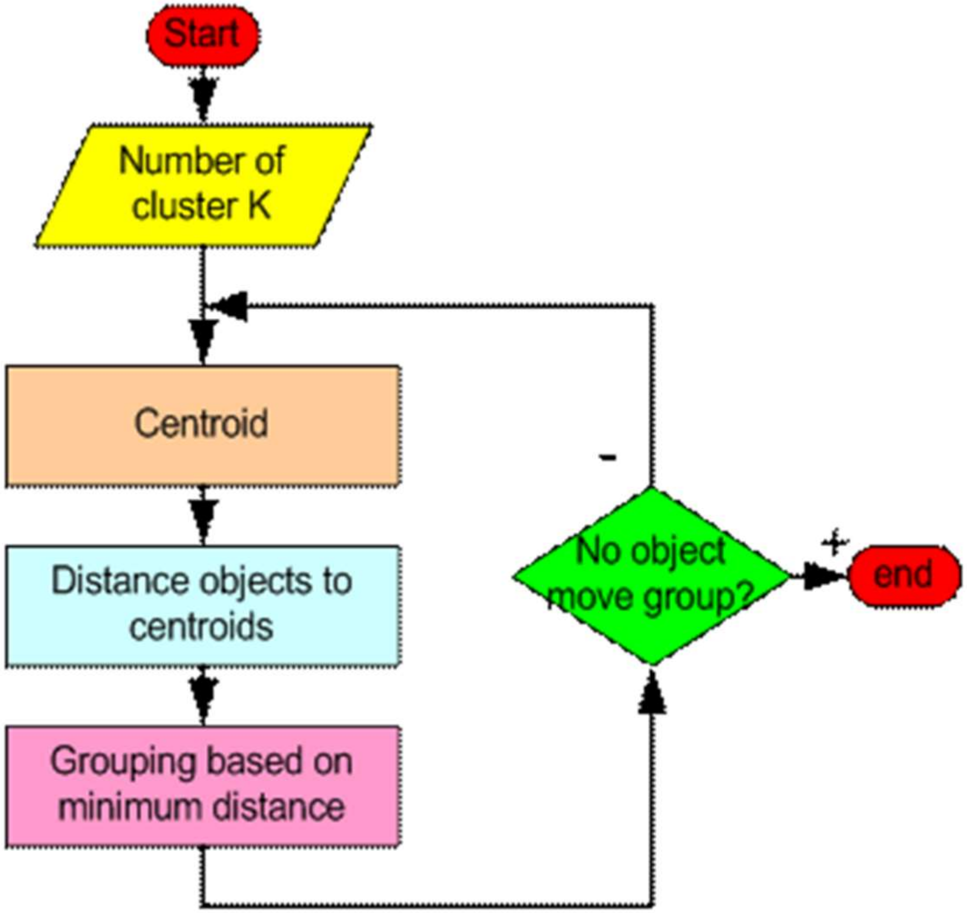
Our goal is to group these objects into K=2 group of medicine based on the two attributes

k-mean Clustering

- The k-mean algorithm
- 3 steps
  - Repeat
  1. Determine the centroid coordinate
  2. Determine the distance of each object to the centroids.
  3. Group the object based on minimum distance
  - Iterate until *stable*

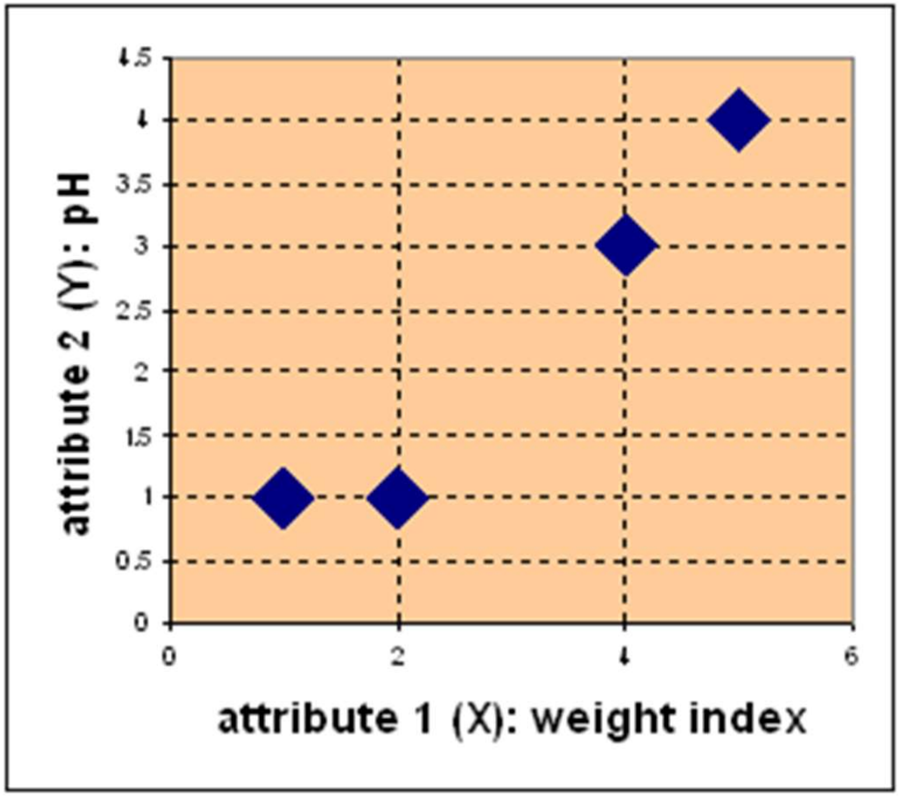
k-mean Clustering

- Flowchart


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

k-mean Clustering

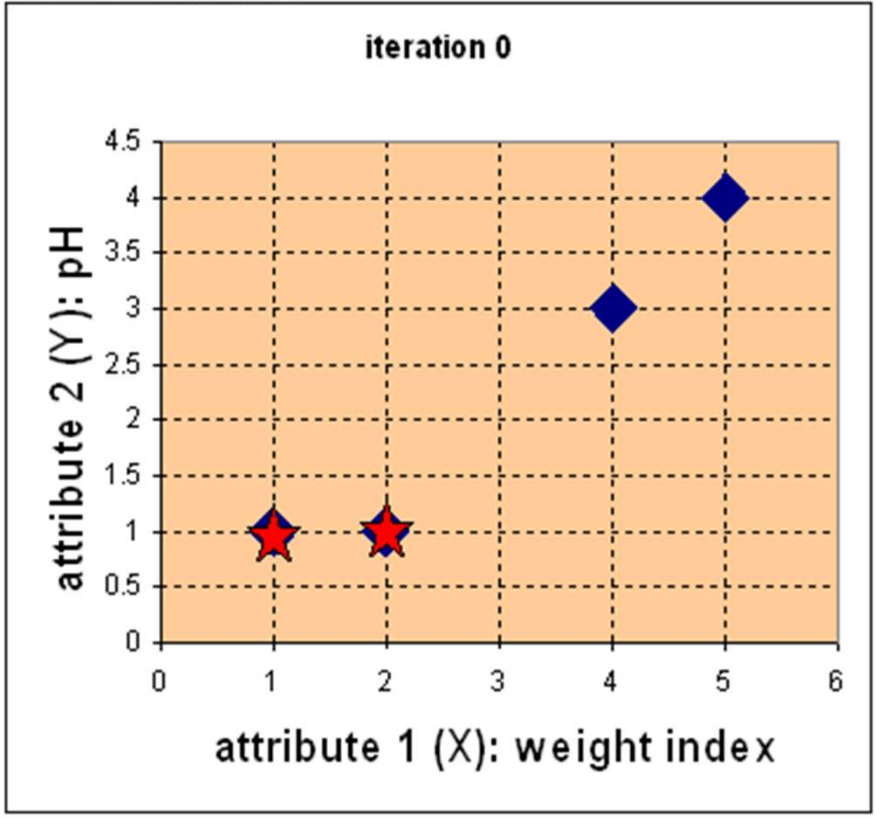
- The feature space



Medicine	Attrib1	Attrib2
A	1	1
B	2	1
C	4	3
D	5	4

k-mean Clustering

- Step1: Initial centroids:
  - c1=(1,1)
  - c2=(2,1)



k-mean Clustering

- Step2: Objects-Centroids distance *(Euclidean distance)*

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{array}{ll} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{array}$$

$$\begin{array}{cccc} A & B & C & D \end{array}$$

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

- Example distance from (4,3) to c(1,1)

$$\sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

k-mean Clustering

- Step3:The element of group matrix G

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{array}{ll} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{array}$$

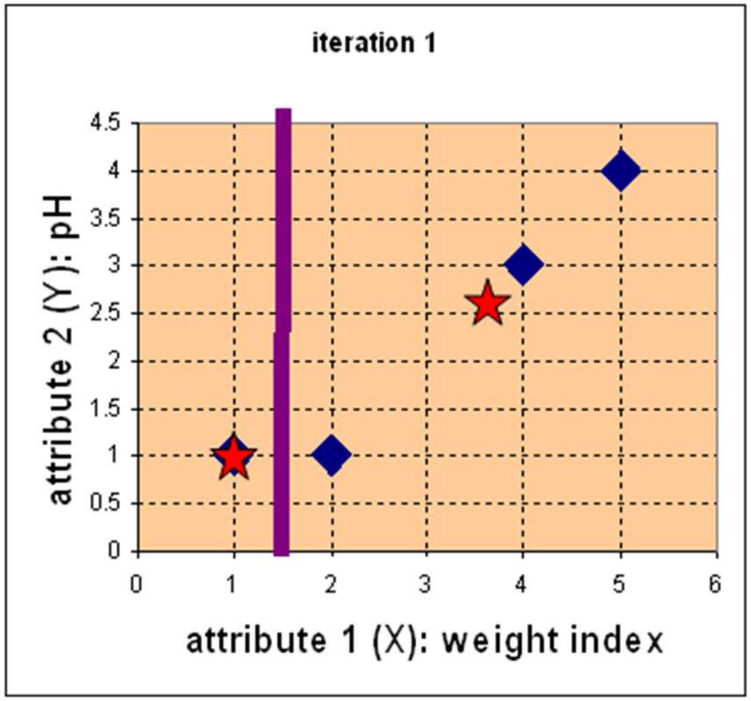
$$\begin{array}{cccc} A & B & C & D \end{array}$$

$$\begin{array}{cccc|c} 1 & 2 & 4 & 5 & X \\ 1 & 1 & 3 & 4 & Y \end{array}$$

$$\mathbf{G}^0 = \begin{array}{c} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{l} group-1 \\ group-2 \end{array}$$

k-mean Clustering

- Repeat step1: determine centroids



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

$$\mathbf{c}_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\tfrac{11}{3}, \tfrac{8}{3}\right)$$

k-mean Clustering

- Repeat step2: find distances

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3},\frac{8}{3}) & group-2 \end{matrix}$$

$$\begin{array}{cccc} A & B & C & D \end{array}$$

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

k-mean Clustering

□ Repeat step3: object clustering

$$\mathbf{D}^1 = \begin{array}{c} \left[\begin{array}{cccc} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{array}\right] \\ \begin{array}{cccc} A & B & C & D \end{array} \\ \left[\begin{array}{cccc} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{array}\right] \begin{array}{c} X \\ Y \end{array} \end{array} \quad \begin{array}{ll} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3},\frac{8}{3}) & group-2 \end{array}$$

$$\mathbf{G}^1 = \begin{array}{c} \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right] \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{l} group-1 \\ group-2 \\ \end{array}$$

k-mean Clustering

- Should we repeat again?

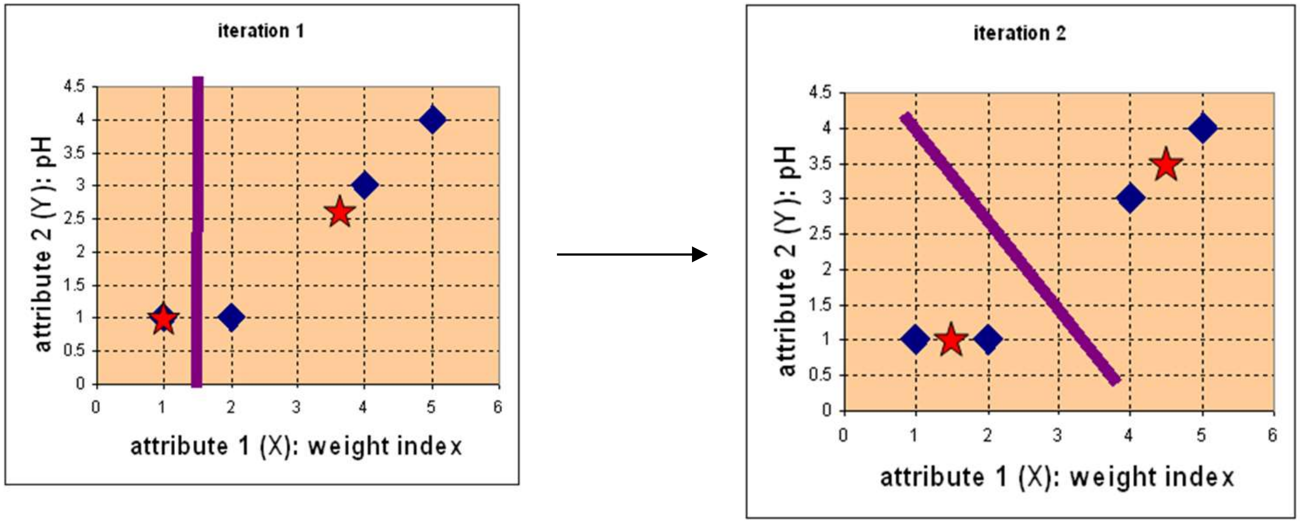
$$\mathbf{G}^0 = \begin{array}{cc} \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{array}\right] & \begin{array}{l} group-1 \\ group-2 \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \end{array} \longrightarrow \mathbf{G}^1 = \begin{array}{cc} \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right] & \begin{array}{l} group-1 \\ group-2 \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \end{array}$$

- yes

k-mean Clustering

- Repeat step1: find new centriods

$$\mathbf{c}_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1\tfrac{1}{2}, 1) \qquad \mathbf{c}_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4\tfrac{1}{2}, 3\tfrac{1}{2})$$



- □ Repeat step 2 and 3

$$\mathbf{D}^2 = \begin{array}{c} \left[\begin{array}{cccc} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{array}\right] \\ \begin{array}{cccc} A & B & C & D \end{array} \\ \left[\begin{array}{cccc} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{array}\right] \end{array} \begin{array}{l} \mathbf{c}_1 = (1\tfrac{1}{2}, 1) \quad group-1 \\ \mathbf{c}_2 = (4\tfrac{1}{2}, 3\tfrac{1}{2}) \quad group-2 \\ \\ X \\ Y \end{array}$$

$$\mathbf{G}^2 = \begin{array}{c} \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right] \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{l} group-1 \\ group-2 \\ \\ \end{array}$$

- □ No change…STOP