UCIrvine | The Paul Merage School of Business

2022.11.24

BANA 273 Session 8

Clustering

Prof. Vibs Abhishek

The Paul Merage School of Business

University of California, Irvine

Agenda

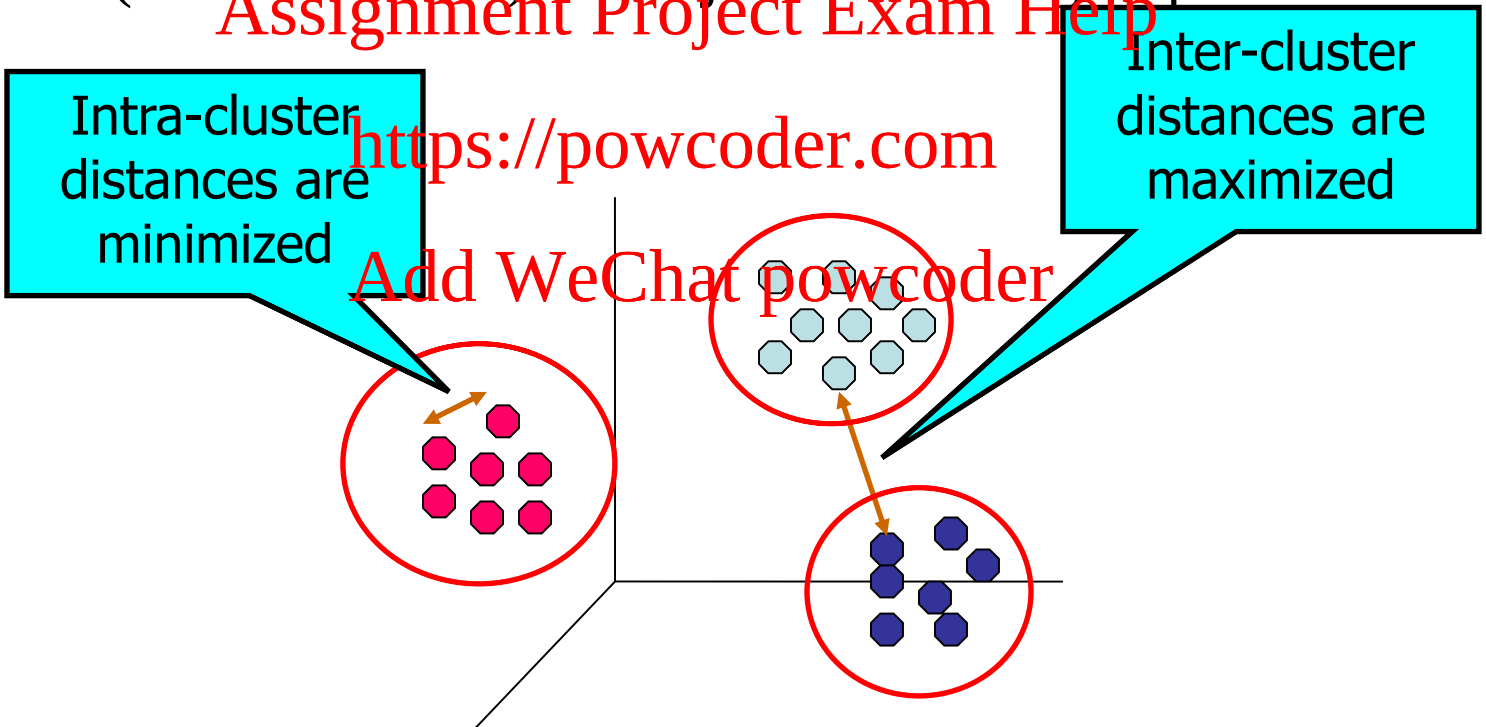
- Assignment 4 due on Canvas soon
- Please work on your projects
- Clustering using k-means algorithm

Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Intra-cluster distances are minimized

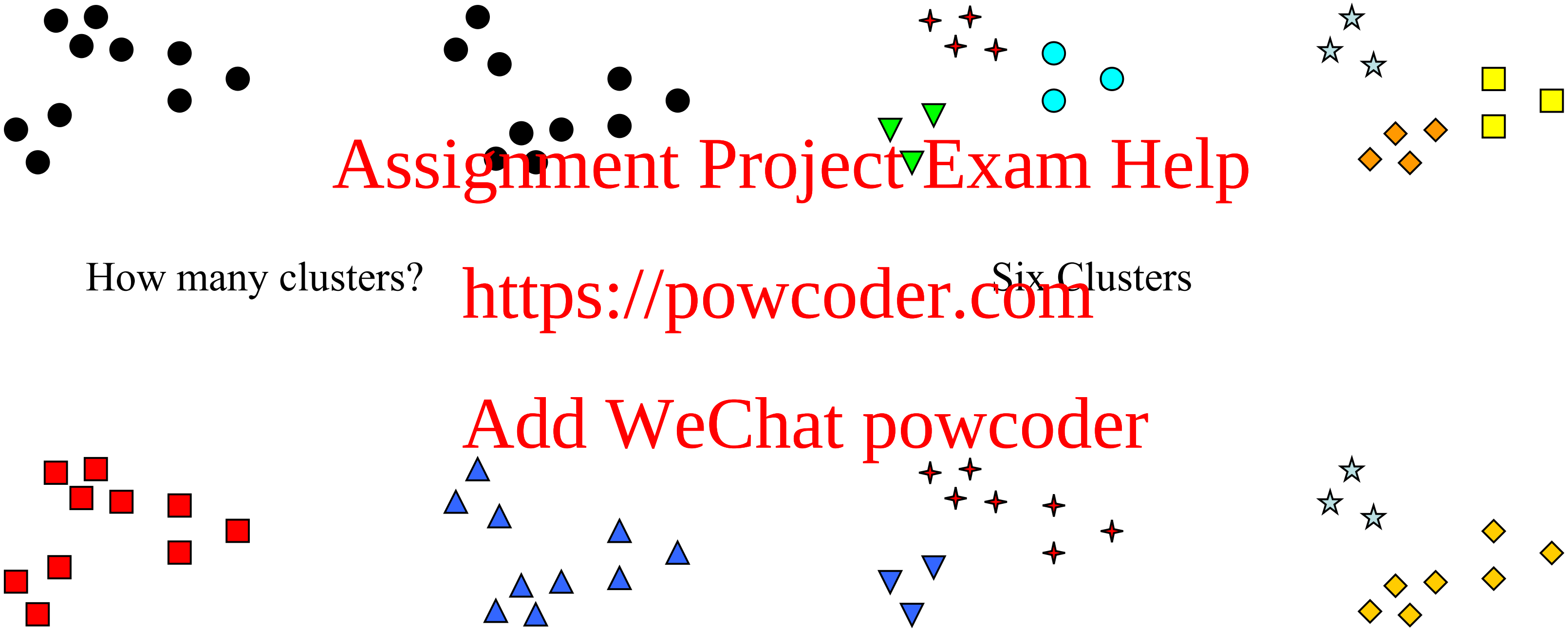
Inter-cluster distances are maximized

 Euclidean Distance Based Clustering in 3-D space.

Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Notion of a Cluster can be Ambiguous



How many clusters?

Six Clusters

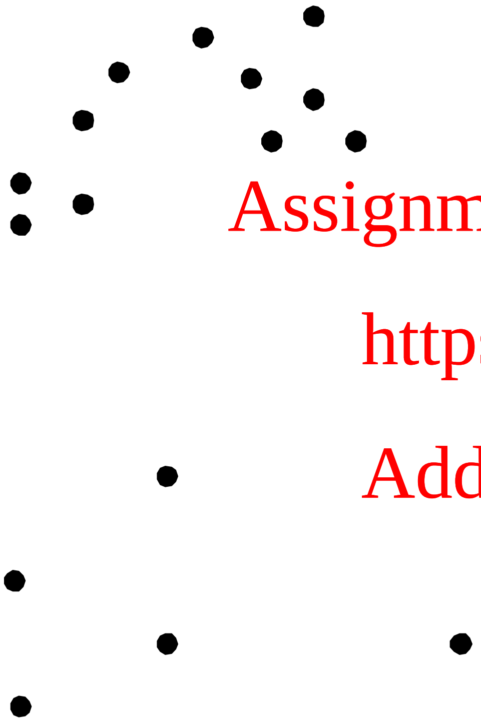
Two Clusters

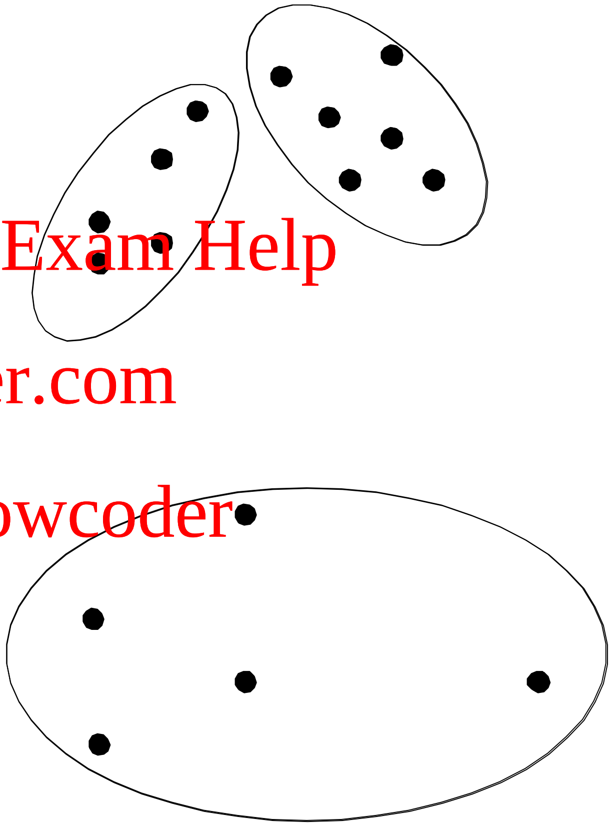
Four Clusters

Types of Clusterings

- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
- Partitional Clustering
  - A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

Partitional Clustering





Original Points

A Partitional Clustering

K-Means Clustering

1. Begin by specifying K, the number of clusters
2. Select K points as initial cluster centroids
3. Assign each point to the cluster whose centroid is closest using similarity measure (Euclidean Distance)
4. Re-compute the centroids of the clusters
5. Repeat steps 3 and 4 until points stop moving between clusters

Similarity Measure

- Need a distance measure
- Example of a distance measure:
  - Manhattan distance:

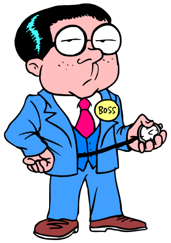
$$D(X,Y)=\sum_{i=1}^{n}|x_i-y_i|$$

Similarity Metric

- Example for a distance measure:
  - Euclidean distance

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Example of Euclidean Distance



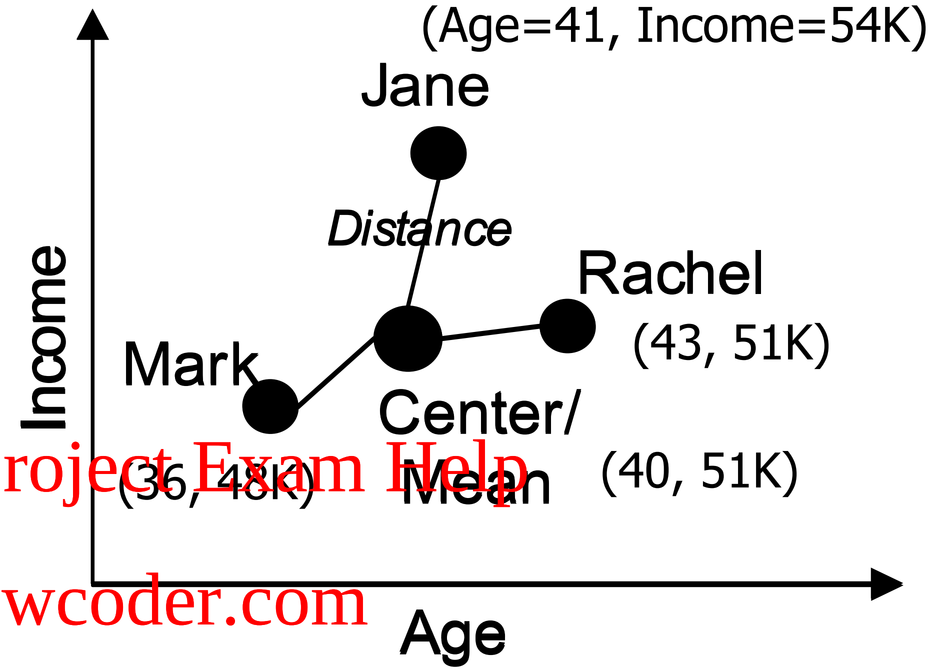
John:
Age=35
Income=95K
no. of credit cards=3

Rachel:
Age=41
Income=215K
no. of credit cards=2

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Distance (John, Rachel)=sqrt $[(35-41)^2+(95-215)^2+(3-2)^2]$

K-Means

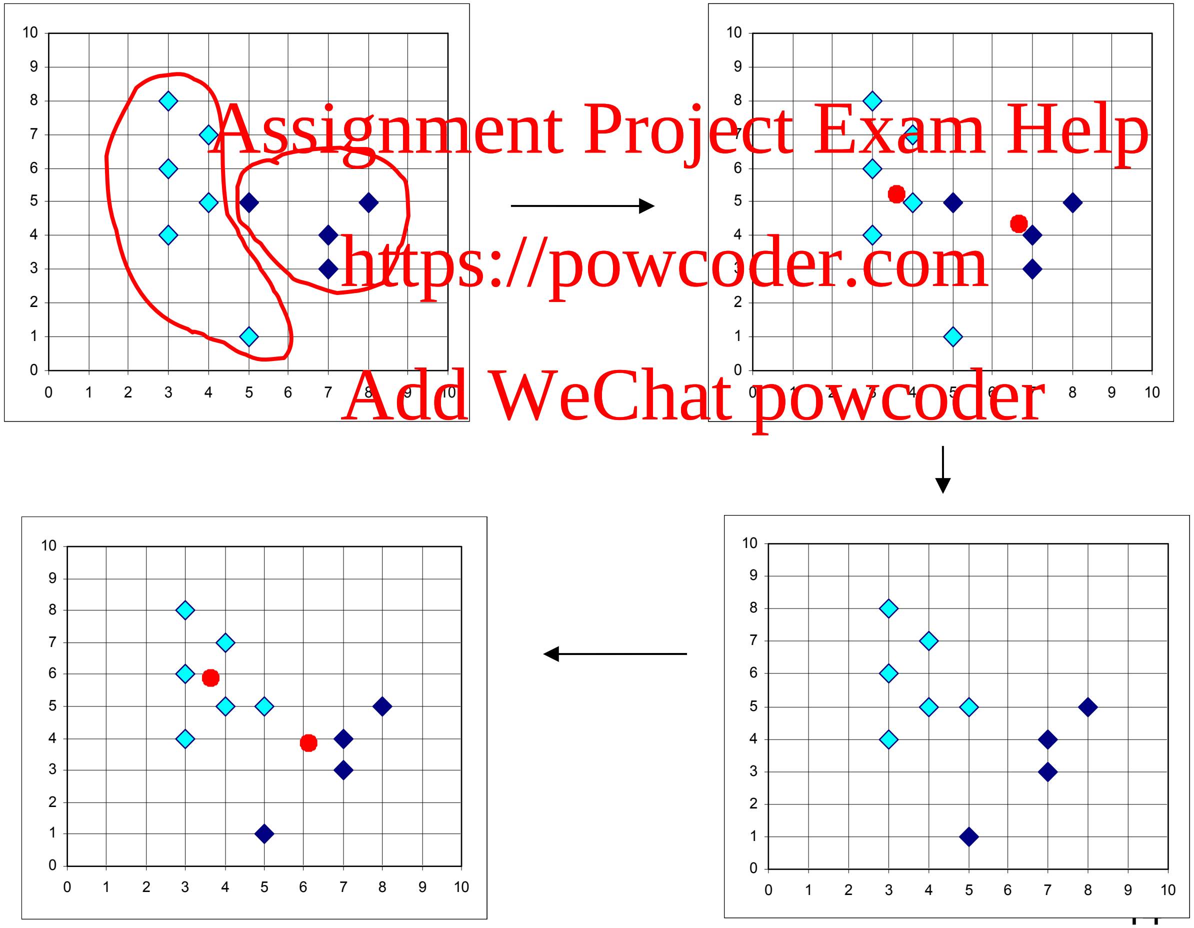


Cluster center is

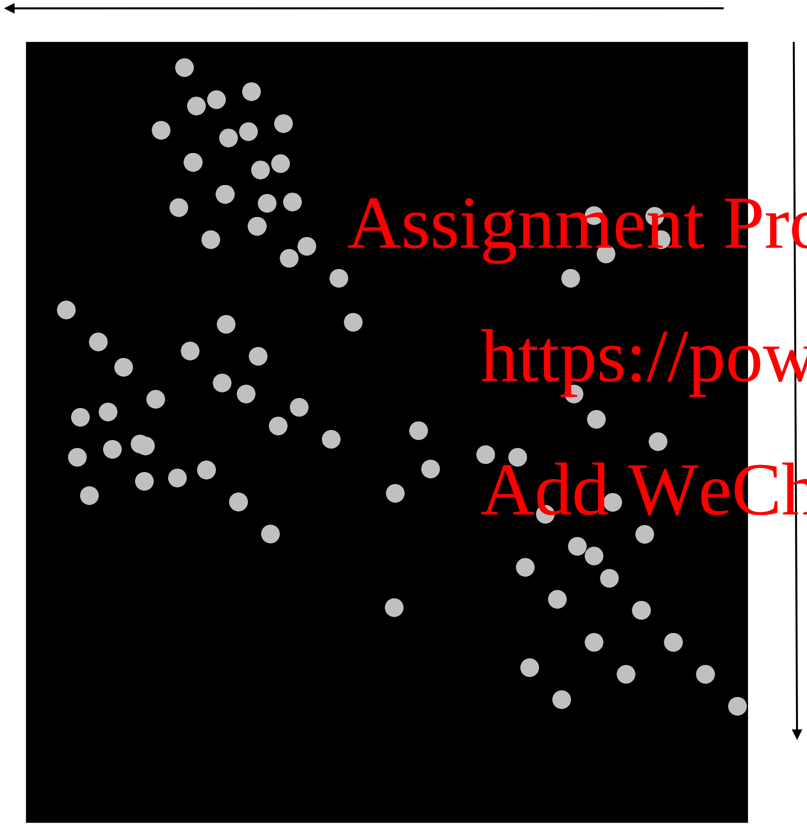
Age=(36+41+43)/3=40

Income=(48K+51K+54K)/3=51K

Example: 2-Means

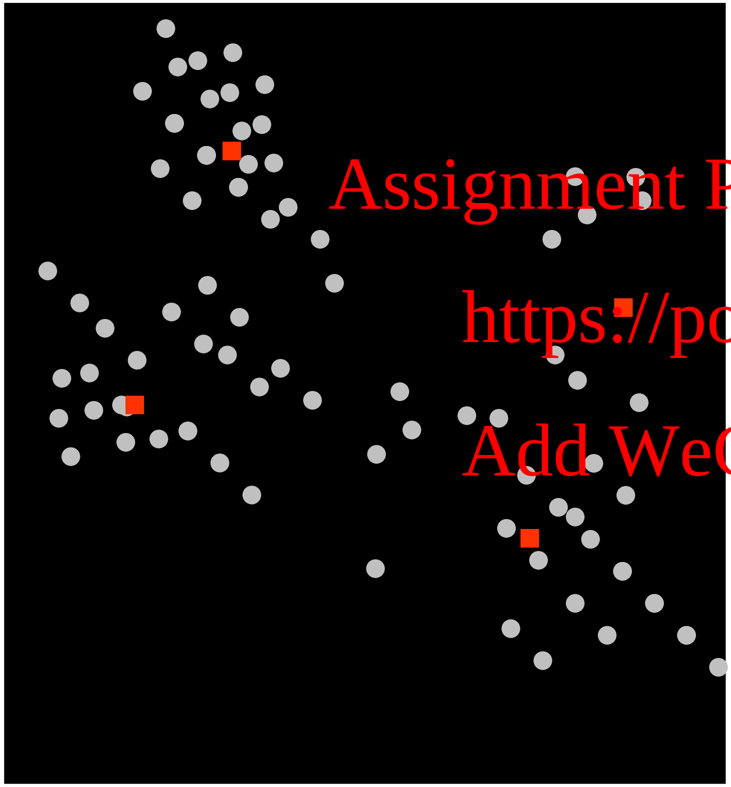


K-means clustering



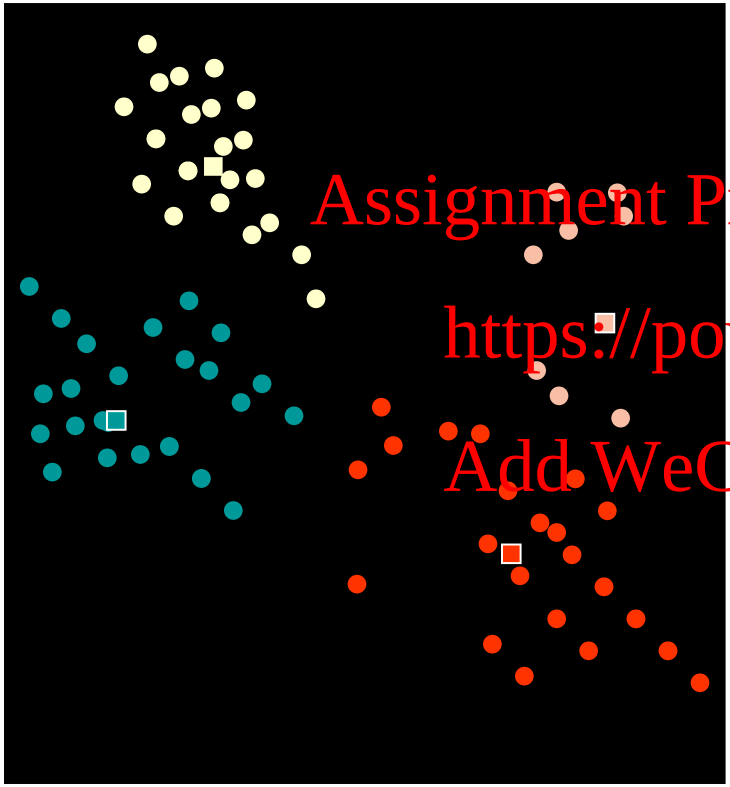
1. Select inputs

K-means clustering



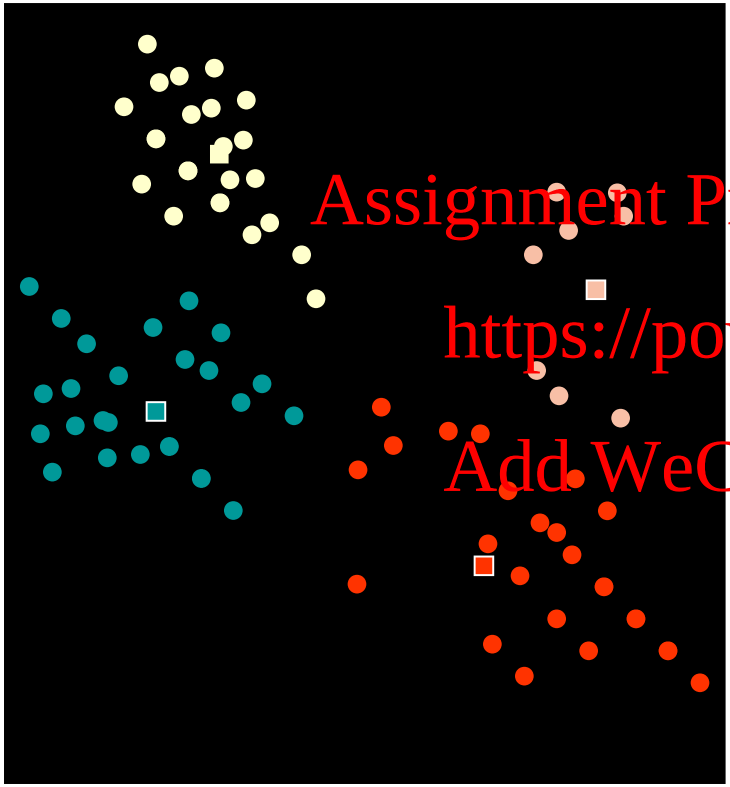
1. Select inputs
2. Select k cluster centers

K-means clustering



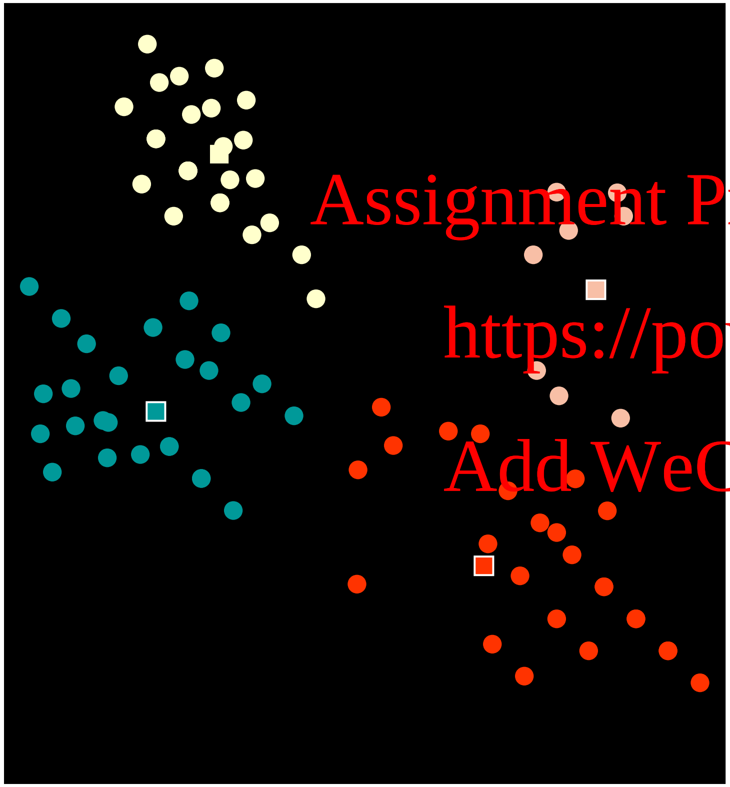
1. Select inputs
2. Select k cluster centers
3. Assign cases to closest center
   - Need to define "close"

K-means clustering



1. Select inputs
2. Select k cluster centers
3. Assign cases to closest center
   - Need to define “close”
4. Update cluster centers

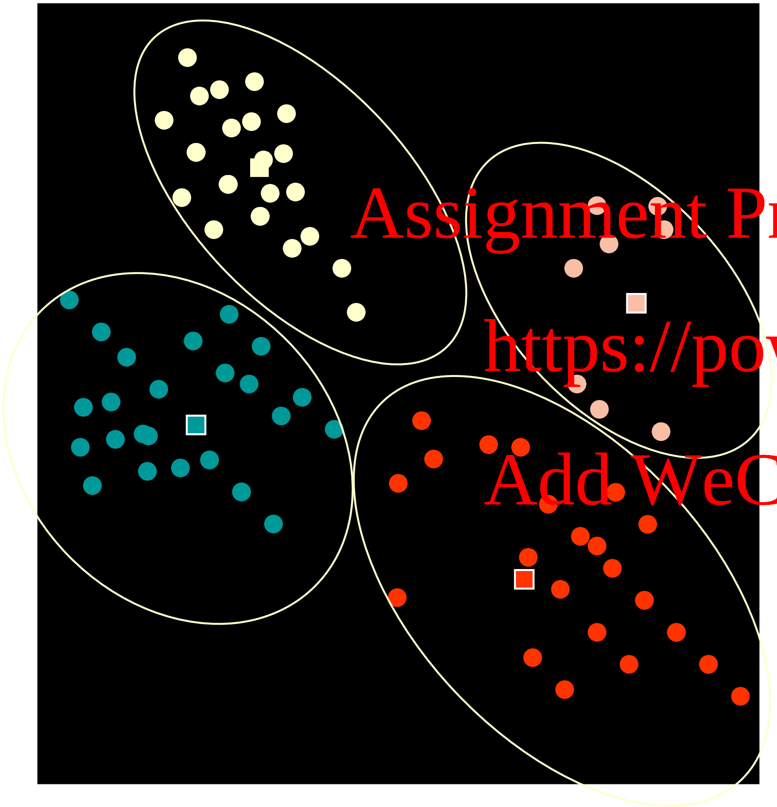
K-means clustering



1. Select inputs
2. Select k cluster centers
3. Assign cases to closest center
   - Need to define “close”
4. Update cluster centers
5. Re-assign cases

UCIrvine The Paul Merage School of Business

K-means clustering



1. Select inputs
2. Select k cluster centers
3. Assign cases to closest center
   - Need to define "close"
4. Update cluster centers
5. Re-assign cases
6. Repeat steps 4 and 5 until changes in cluster centers & assigned cases are insignificant

“k” in k-means clustering

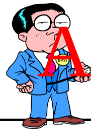
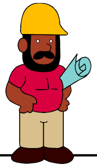
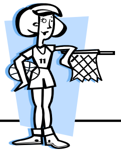
- Generally, *k* is set in advance
- If not known, a good idea is to try out different values of *k* that are near the number of clusters one expects from the data, to see how the sum of distances (in the clusters) reduces with larger *k*’s

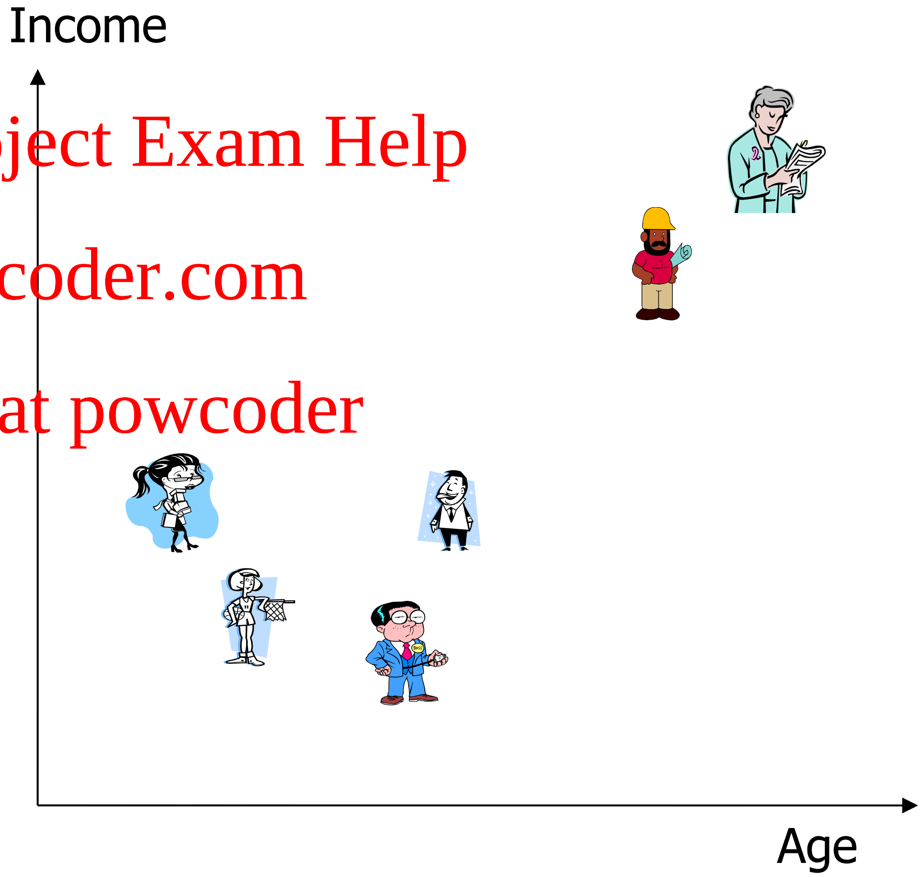
Cluster Validity

- Compute ratio
  = [sum of squared distances for a given *k*] / [sum of squared distances to the mean of all the records (*k* = 1)]
  - If the ratio is near 1.0 the clustering has not been very effective
  - If it is small we have well-separated groups
- Weka reports sum of squared errors (Intra cluster distance)

Example

Note: Both Age and Income are normalized.

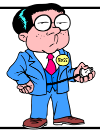
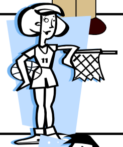
Customer	Age	Income (K)
John 	0.55	0.175
Rachel	0.34	0.25
Hannah	1	1
Tom 	0.93	0.85
Nellie 	0.39	0.2
David	0.58	0.25

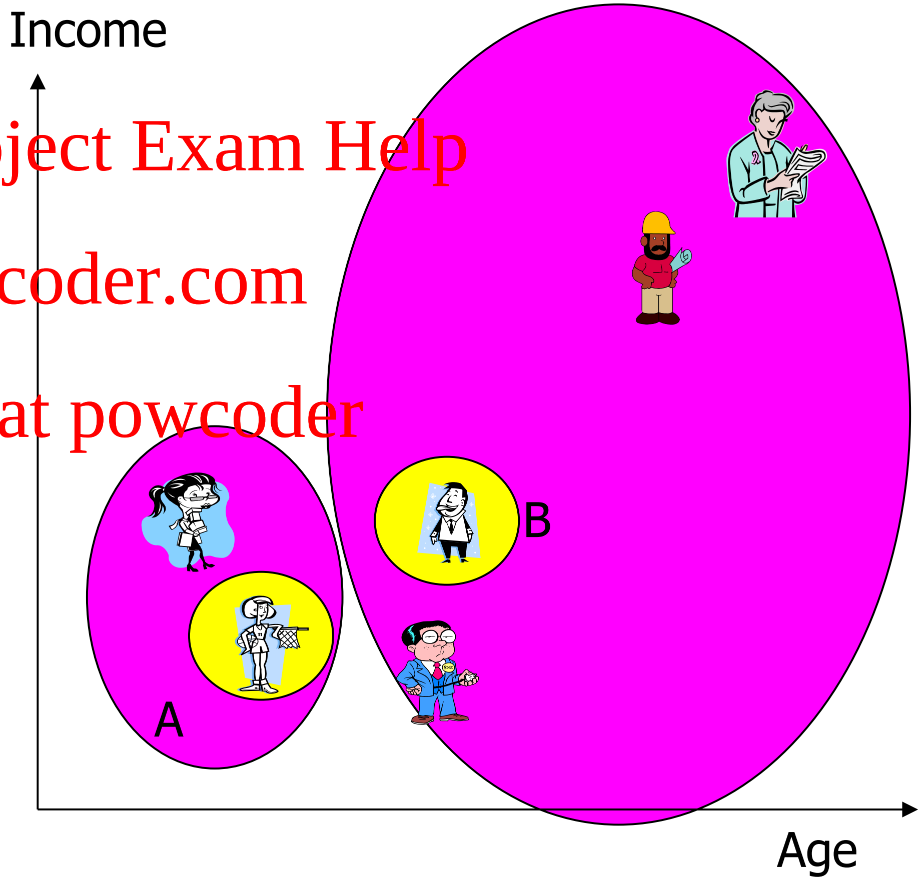


K-Means Algorithm: Example

Step 1:

Nellie and David are selected as cluster centers A and B respectively

Customer	Distance from David	Distance from Nellie
John 	0.08	0.16
Rachel	0.24	0.07
Hannah	0.86	1.01
Tom	0.69	0.85
Nellie 		
David		



K-Means Algorithm: Example

Calculate cluster center:

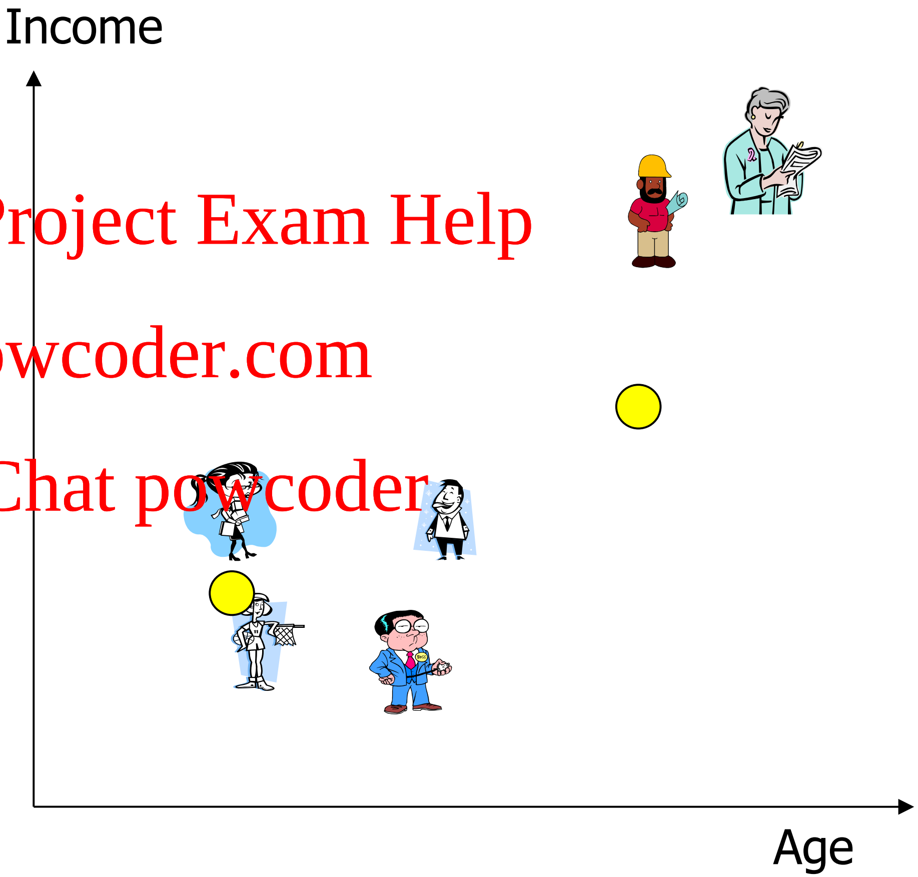
Cluster A center:

- Age 0.37, Income=0.23

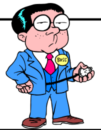
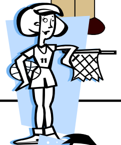
Cluster B center:

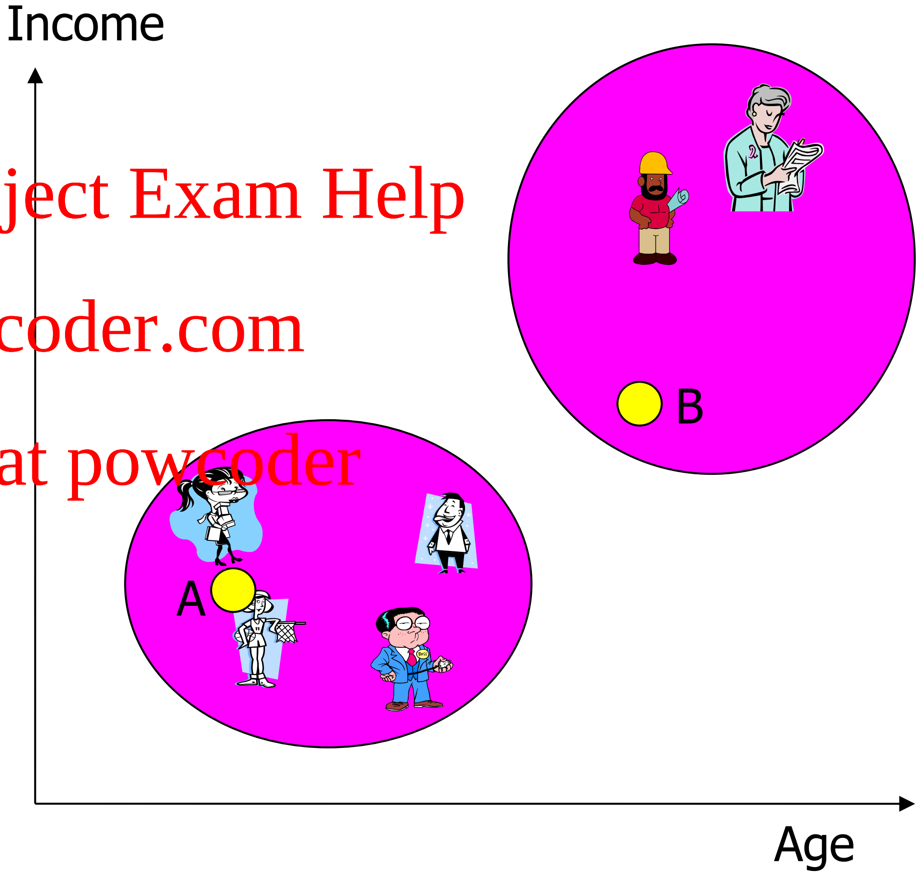
- Age 0.77, Income=0.57

Assign customers to clusters based on new cluster centers



K-Means Algorithm: Example

Customer	Distance A	Distance B
John 	0.19	0.45
Rachel	0.04	0.54
Hannah	0.99	0.49
Tom	0.84	0.32
Nellie 	0.04	0.53
David	0.21	0.37



K-Means Algorithm: Example

Calculate cluster center:

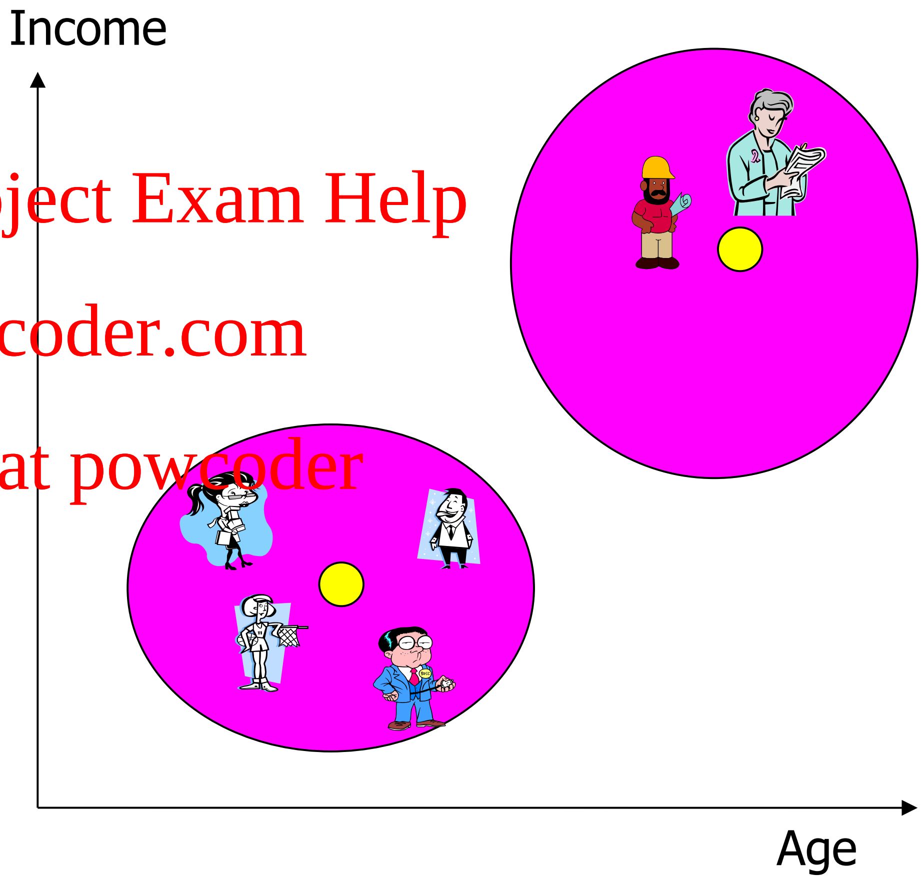
Cluster A center:

- Age 0.47, Income=0.22

Cluster B center:

- Age 0.97, Income= 0.93

- Clusters do not change



Scale and Weigh Data

- Scaling makes sure that the distance is not biased by units (1K, 1M, etc.)

- Weighting can add the information that one variable is more (or less) important than others.

- After scaling to get rid of biases caused by different units, use weights to introduce bias based on knowledge of the business context.
  - (eg. Two households with the same income are more similar than two households with the same number of pets.)

- Common way to scale:
  - Range: (value-min)/(max-min); [0,1]
    - E.g. {11,8,4,6,10,1} → {1, 0.7, 0.3, 0.5, 0.9, 0}

What is a “Good” cluster?

A. Inter-cluster distance is maximized and intra-cluster distance is maximized

B. Inter-cluster distance is minimized and intra-cluster distance is maximized

C. Inter-cluster distance is maximized and intra-cluster distance is minimized

D. Inter-cluster distance is minimized and intra-cluster distance is minimized

E. None of the Above

Clustering in Weka

Utility Example

East West Airlines

http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/k-means.html

Clustering Exercise

Start with indivuduals 1 and 4 as initial centroids

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Strengths and Weaknesses of the *K-Means*

- Strength
  - Relatively efficient
  - Simple implementation

- Weakness
  - Need to specify *k*, the number of clusters, in advance
  - Unable to handle noisy data and outliers well
  - Euclidian Distance does not work for nominal variables.

Applications of Clustering

- **Marketing**: Customer segmentation (discovery of distinct groups of customers) for target marketing. Create product differentiation: different offers for different segments (It’s not always possible to offer personalization.)
- **Car insurance**: Identify customer groups with high average claim cost
- **Property**: Identify houses in the same city with similar characteristics
- **Image recognition**
- Creating **document collections**, or grouping web pages

Review of Assignments

Next Session

- Review of Assignment 4
- Review of sample final exam
- Other Data mining techniques
  - Text Mining
  - Collaborative Filtering