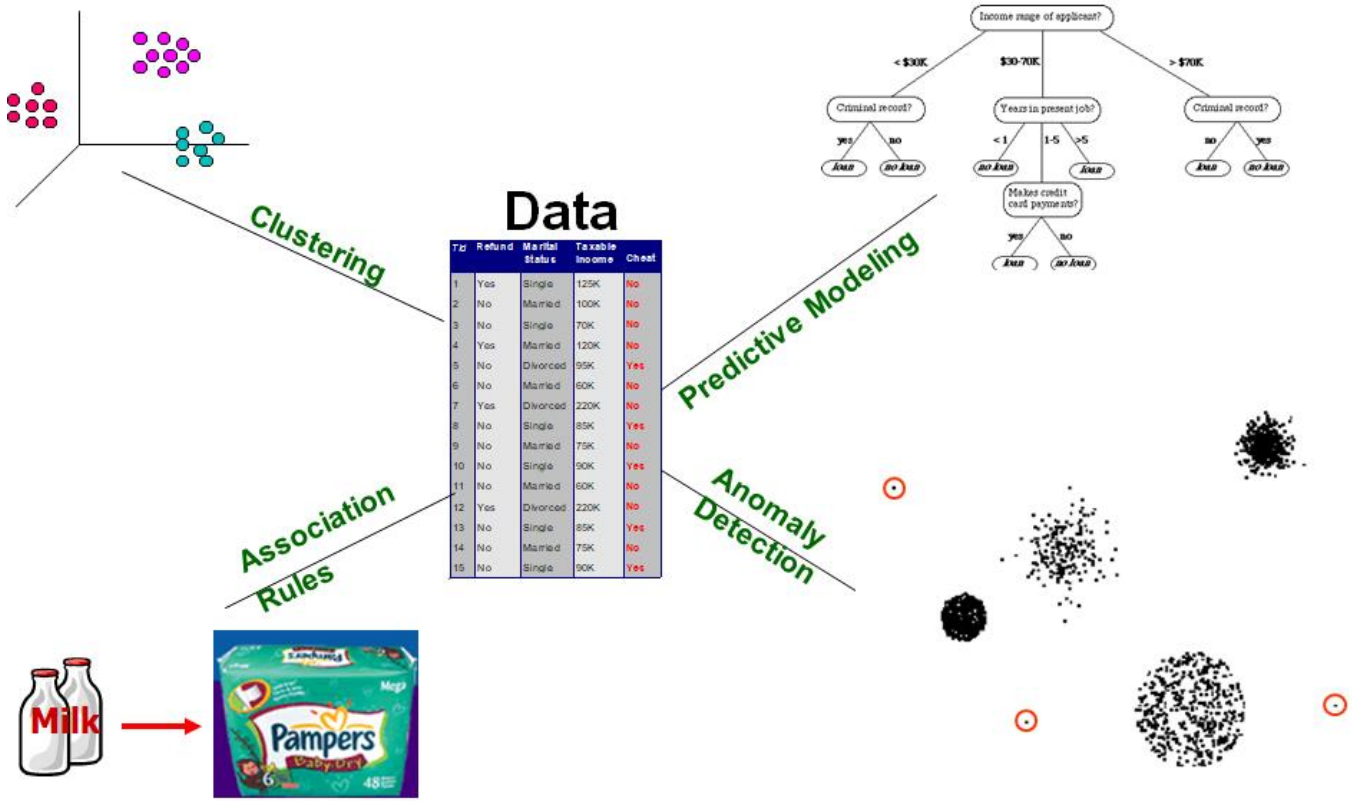
Foundations of Data Science & Analytics: K-means Clustering

Ezgi Siir Kibris

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

Tasks

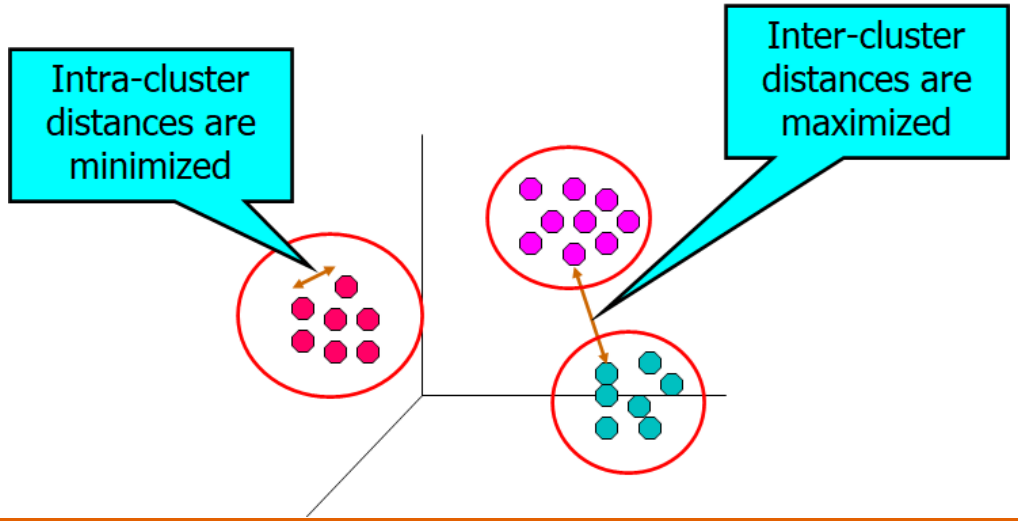

Clustering
Data
Predictive Modeling
Association Rules
Anomaly Detection
Milk
Pampers
Income range of applicant?
< $30K
$30-70K
> $70K
Criminal record?
Years in present job?
Criminal record?
Makes credit card payments?
Tid Refund Marital Status Taxable Income Cheat
1 Yes Single 125K No
2 No Married 100K No
3 No Single 70K No
4 Yes Married 120K No
5 No Divorced 95K Yes
6 No Married 60K No
7 Yes Divorced 220K No
8 No Single 85K Yes
9 No Married 75K No
10 No Single 90K Yes
11 No Married 60K No
12 Yes Divorced 220K No
13 No Single 85K Yes
14 No Married 75K No
15 No Single 90K Yes

Clustering vs Predictive Modeling

	Training Data
Predictive Modeling (Supervised Learning):	Independent Variables + Dependent Variables
Clustering (Unsupervised Learning):	Independent Variables

What is Clustering?

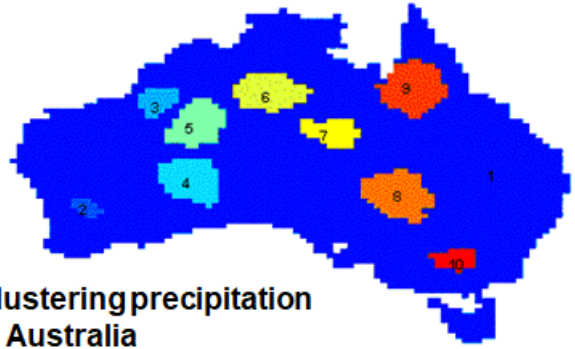
Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Applications

- **Understanding**
  - Group related documents for browsing,
  - group genes and proteins that have similar functionality,
  - or group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large datasets

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



Clustering precipitation in Australia

No Single Correct Answer

How many clusters?

Six Clusters

Two Clusters

Four Clusters

Types of Clusterings

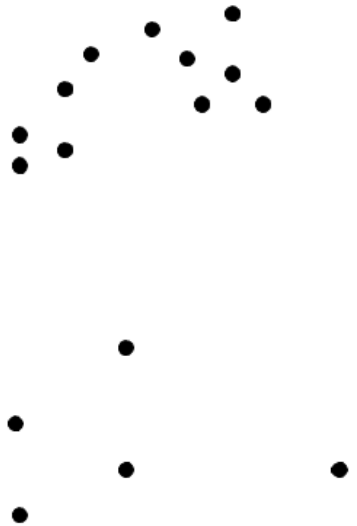
Partitional Clustering

- A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Usually relies on user to decide the number of clusters beforehand

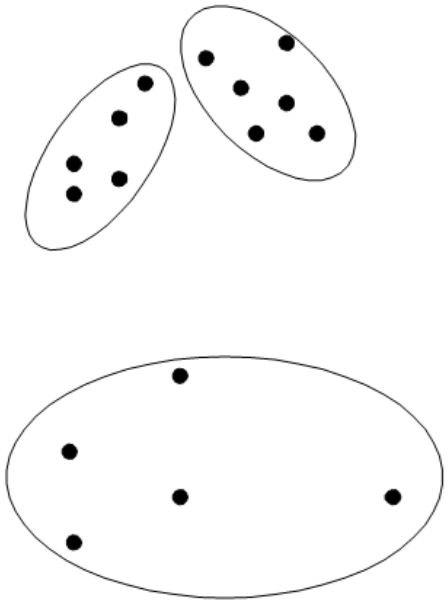
Hierarchical clustering

- A set of nested clusters organized as a hierarchical tree
- Can provide more information on the appropriate number of clusters

Partitional Clustering

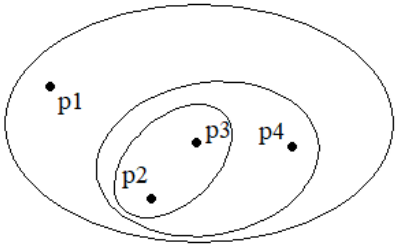


Original Points

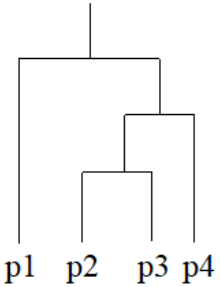


A Partitional Clustering

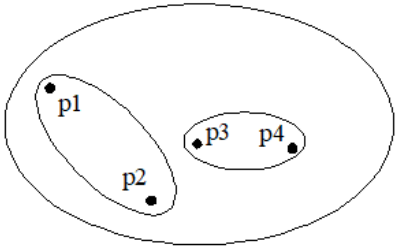
Hierarchical Clustering



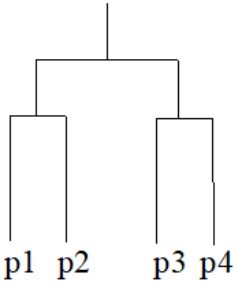
Traditional Hierarchical Clustering



Traditional Dendrogram



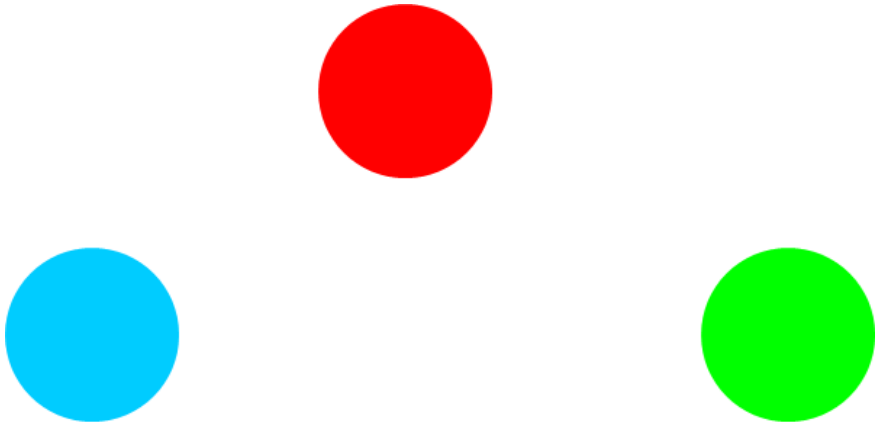
Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Well-separated clusters

A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

Center-based clusters

- A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
- The center of a cluster is often a **centroid**, the average of all the points in the cluster, or a medoid, the most “representative” point of a cluster

4 center-based clusters

K-means

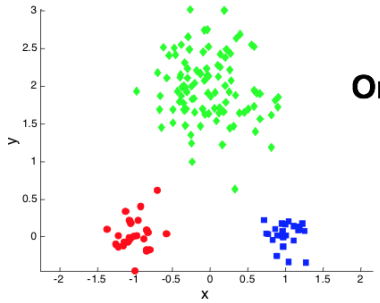
- Partitional clustering approach
- Number of clusters, K, must be specified
- Each cluster is associated with a **centroid** (center point)
- Each point is assigned to the cluster with the closest centroid
- The basic algorithm is very simple

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

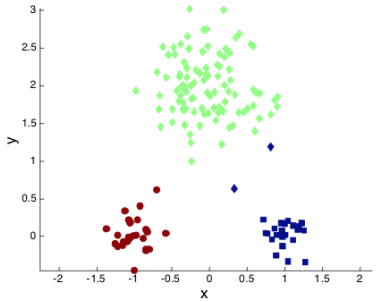
K-means Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- **Closeness** is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to "**until relatively few points change clusters**"
- **Complexity** is O(n * K * I * d)
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of features

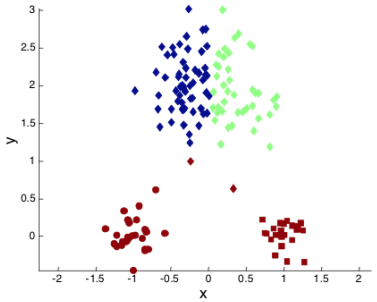
Two Different K-means Runs



Original Points



Optimal Clustering

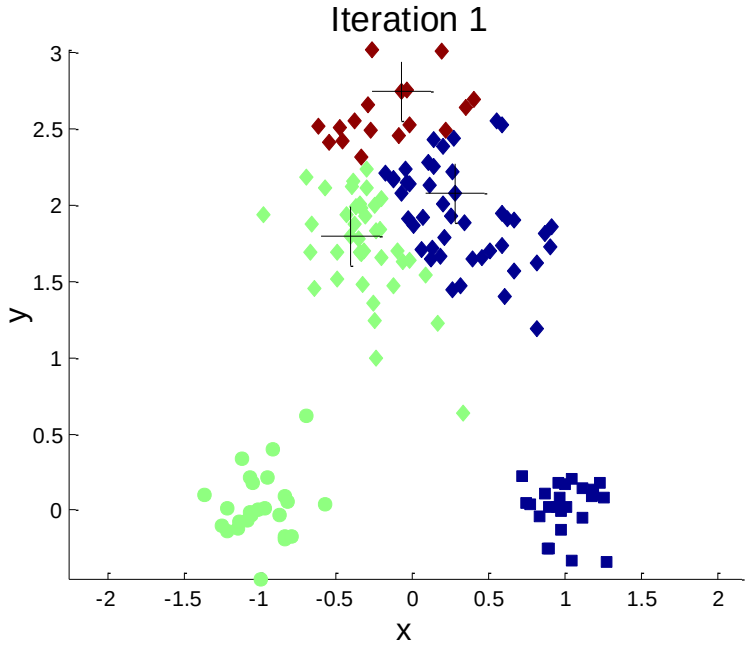


Sub-optimal Clustering

Initial Centroid Problem

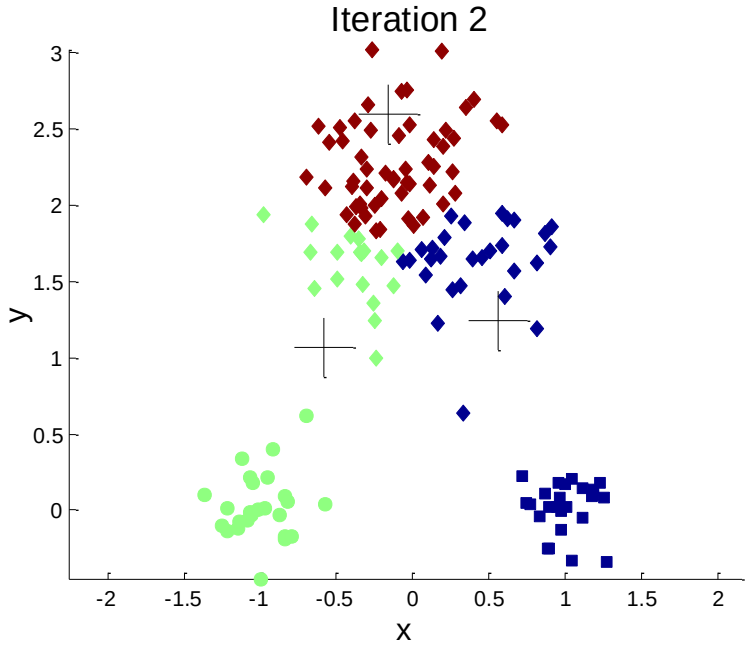
- Initial centroids are chosen **at random**
- Results in potentially large variability in the quality of clusters/clusterings created by K-means
- Due to this problem, it is **important** to run K-means multiple times!

Run 1 – Iteration 1



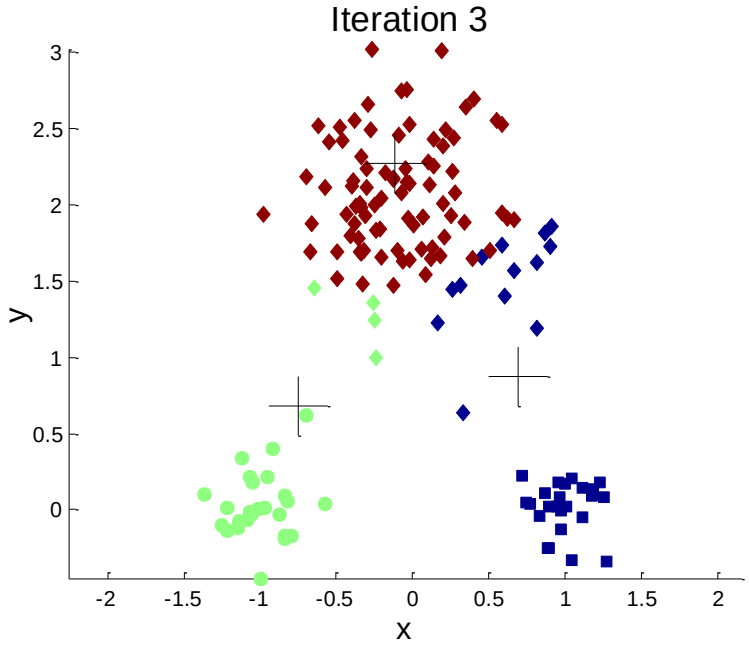
Three data instances are randomly chosen as initial centroids and three clusters are computed (brown, green, blue)

Run 1 – Iteration 2



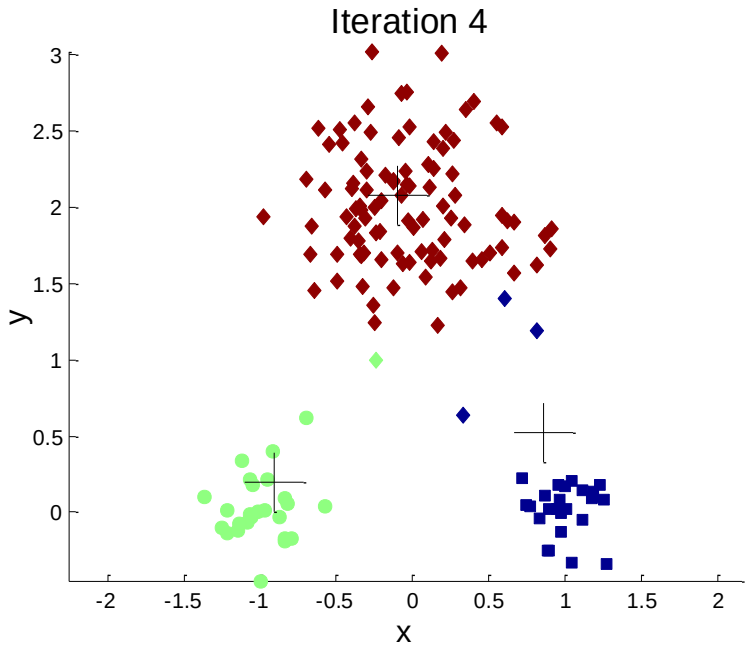
New centroids are computed based upon cluster membership established in Iteration 1 and new membership is computed

Run 1 – Iteration 3

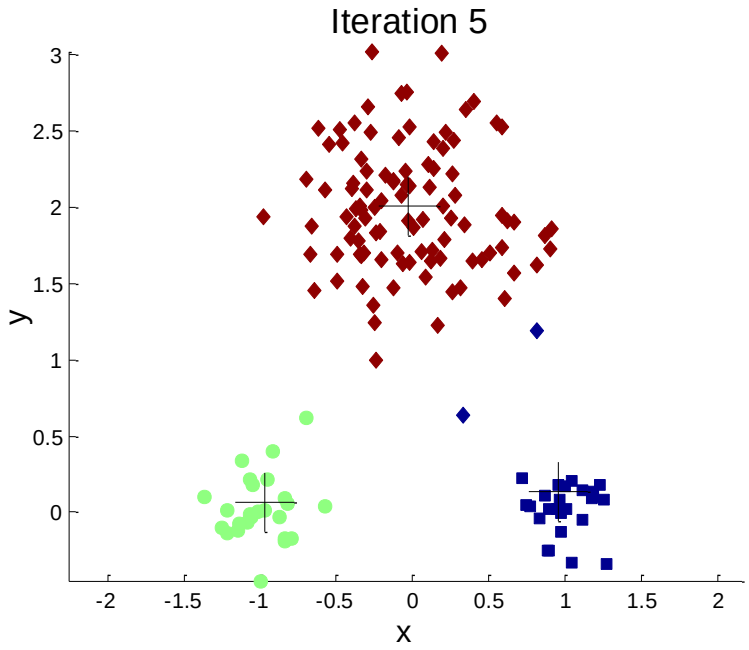


Cluster evolution continues as the lower two centroids move down and to the left and right, respectively

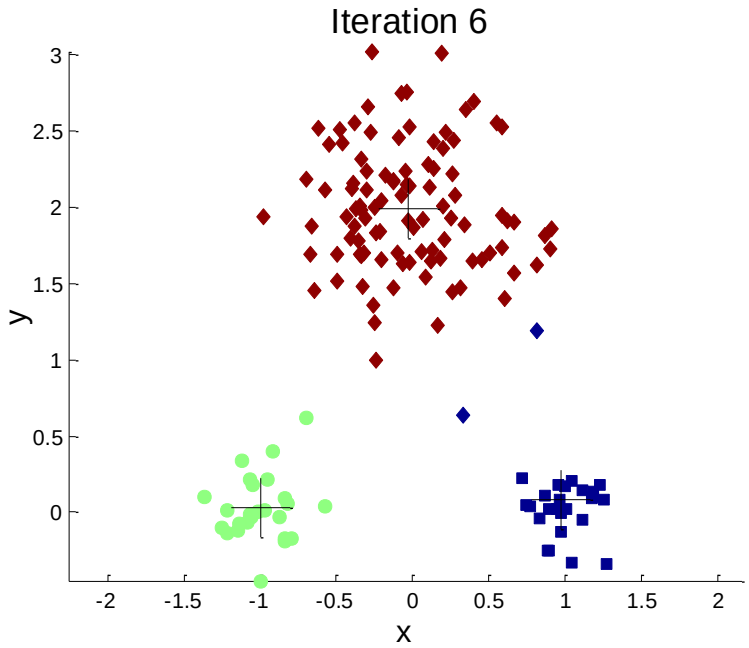
Run 1 – Iteration 4



Run 1 – Iteration 5

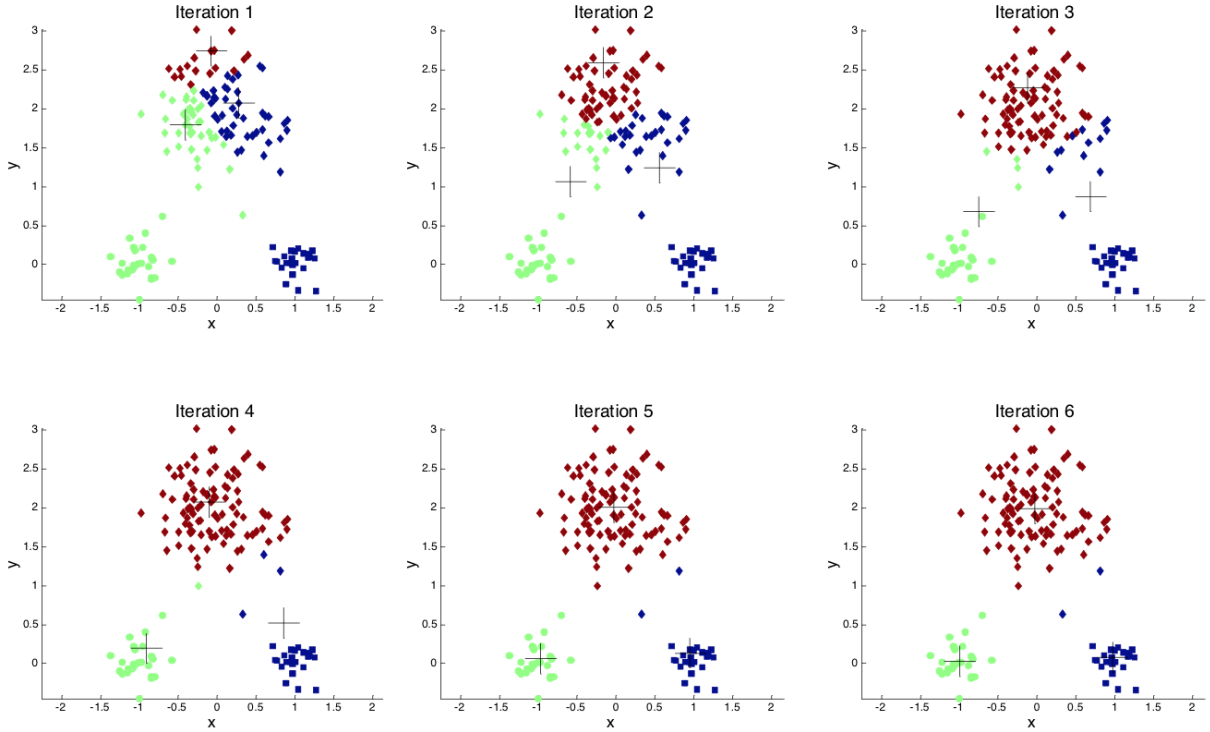


Run 1 – Iteration 6



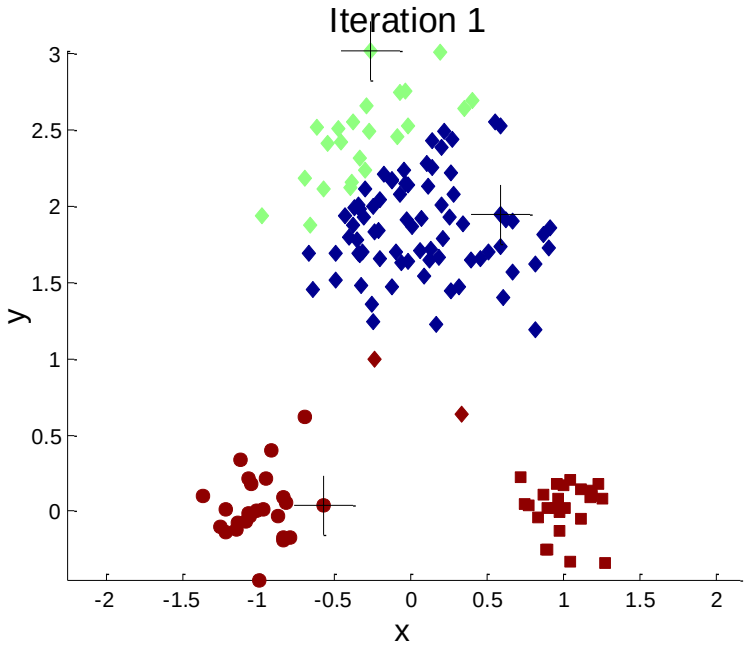
Process stops since only a slight change in cluster centroids occurred and cluster membership did not change this time

All Iterations of Run 1



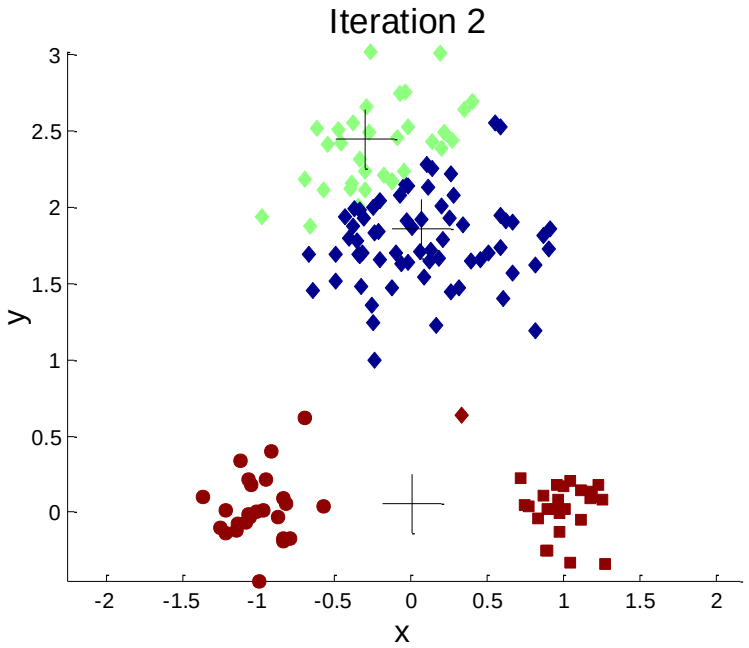
For the most part, the resultant three clusters are well separated

Run 2 – Iteration 1



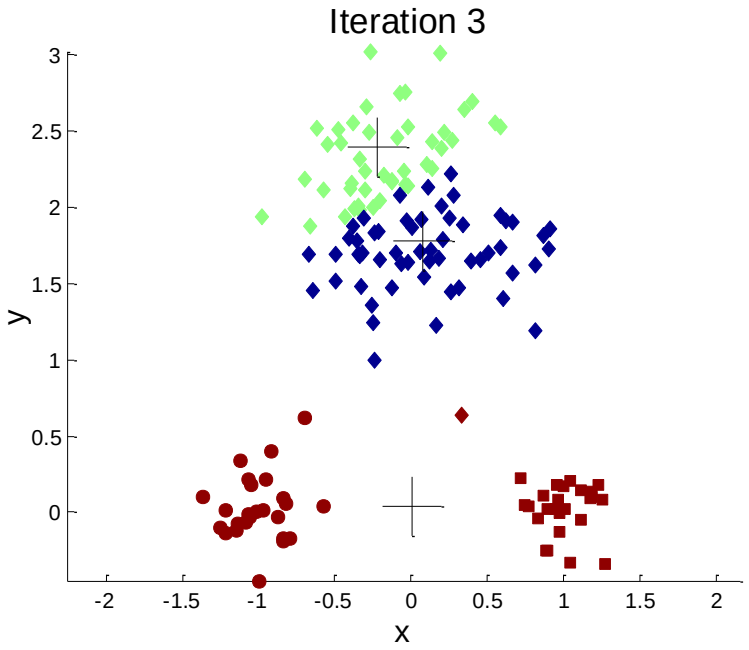
Three **different** data instances are randomly chosen as initial centroids and three clusters are computed (brown, green, blue)

Run 2 – Iteration 2



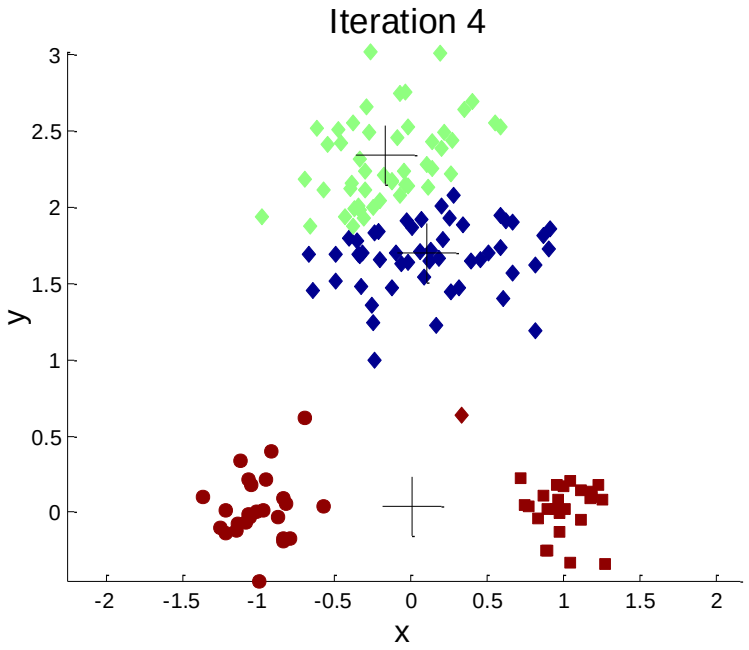
The bottom centroid has formed a single cluster from what was two separate clusters in Run 1

Run 2 – Iteration 3

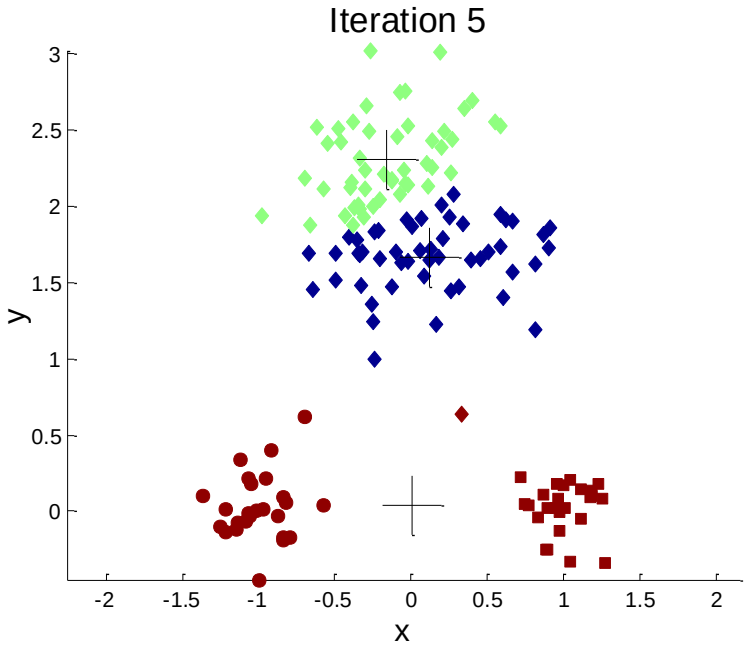


There is only slight movement seen among the centroids

Run 2 – Iteration 4

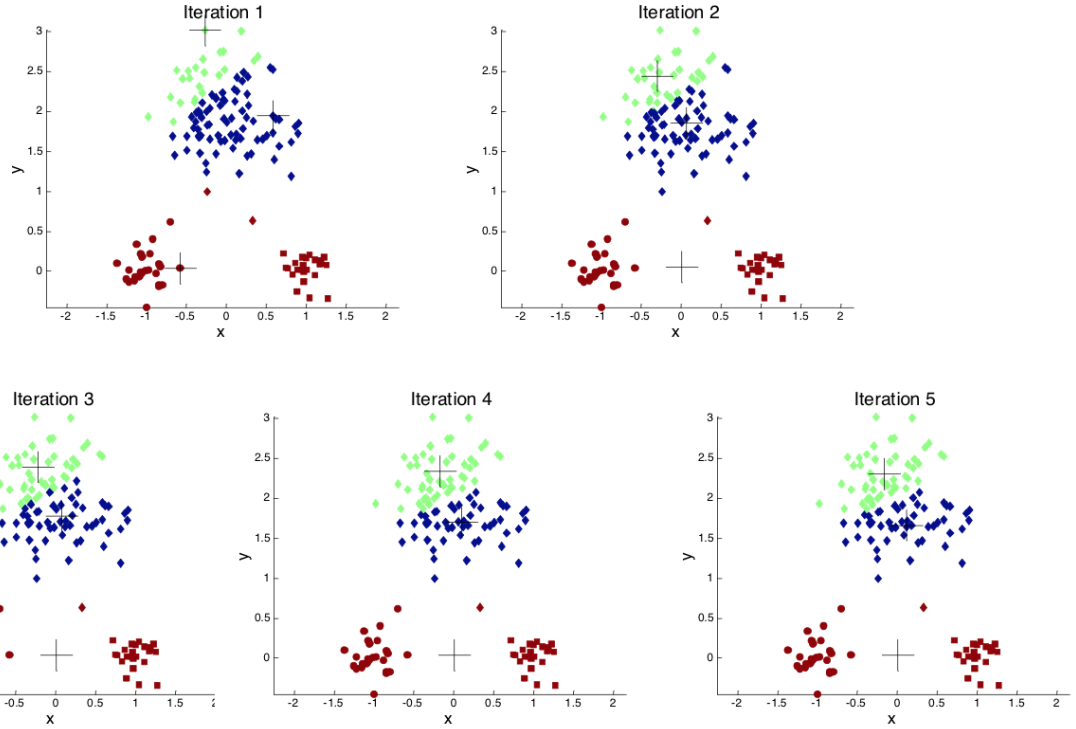


Run 2 – Iteration 5

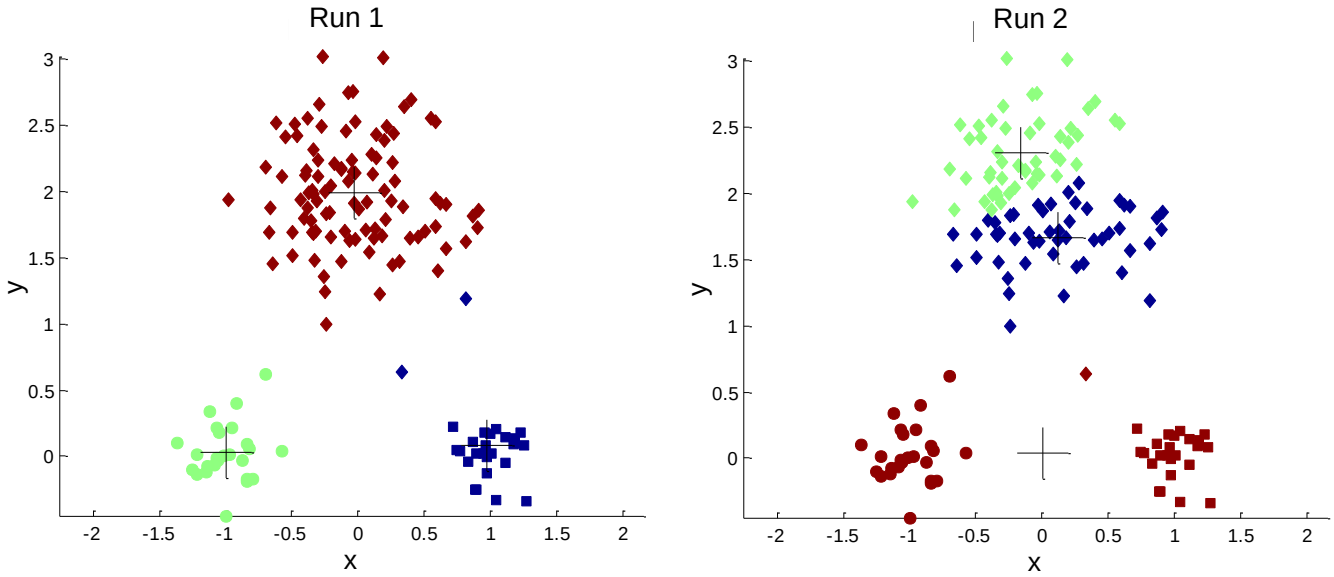


The process ends after no change in cluster membership.
Poor initial centroid choices result in poor clustering.

All Iterations of Run 2



Clustering Comparison



We can qualitatively “see” that Run 1 produced a better result than Run 2, but how do we quantify this?

Evaluating Clusters

- Most common measure is Sum of Squared Error
  - For each data instance, its **error** is the distance to the nearest centroid
  - To get SSE, we square all errors and sum them

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(m_i, x)^2$$

  - x is a data instance in cluster C_i and m_i is the centroid of cluster C_i
  - Given multiple runs of K-means, we typically choose the run with the smallest error

Solutions to Initial Centroid Problem

- Multiple runs of K-means
  - Improves your chances of creating a high quality clustering
- Create more than **K** initial centroids and then select among these initial centroids
  - Select the set of most widely separated centroids
  - *This assumes that you can supply initial centroid locations to a given clustering algorithm*

Pre-processing and Post-processing

- Pre-processing
  - Normalize the data
  - Eliminate outliers
- Post-processing
  - Eliminate small clusters that may represent outliers
  - Split “loose” clusters that have relatively high SSE and merge the points with the closest cluster
  - Merge “close” clusters that have relatively low SSE

Assignment 9