Pattern Recognition

Lecture 19. Clustering

Dr. Shanshan ZHAO

shanshan.zhao@xjtlu.edu.cn

School of AI and Advanced Computing

Xi'an Jiaotong-Liverpool University

Academic Year 2021-2022

Table of Contents

PR/ML algorithms category

Supervised Learning

labeled data, task-driven

- **Classification**
- Regression
  - *e.g., Population growth prediction, Market prediction, etc.*

Unsupervised Learning

unlabeled data, data-driven

- **Dimensionality Reduction**
- **Clustering**

Reinforcement Learning

learn from error

- *e.g., Real-time Decision, Game AI*

Clustering

- Unsupervised learning , Requires data, but no labels
- Detect patterns, e.g.
  - Group emails or search results
  - Regions of images
- Useful when don't know what you're looking for

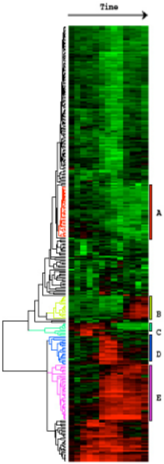
Applications

Image Segmentation

[Slide from James Hayes]

Applications

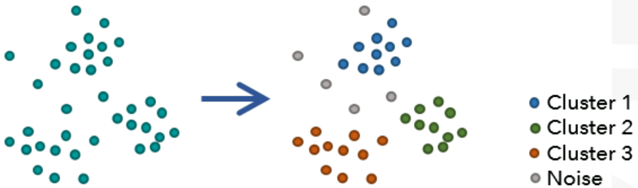
Clustering gene expression data



Eisen et al, PNAS 1998

Clustering

- The organization of unlabeled data **into** similarity groups called clusters.
- A **cluster** is a collection of data items which are “similar” between them, and “dissimilar” to data items in other clusters.



Similarities?

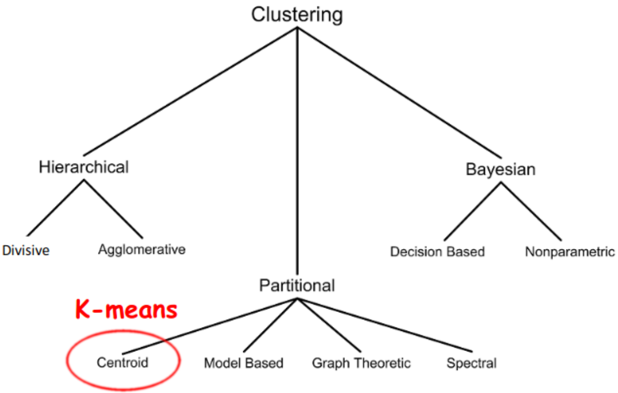
Defining Distance Measures

Definition: Let O1 and O2 be two objects from the universe of possible objects. The distance (dissimilarity) between O1 and O2 is a real number denoted by D(O1,O2) .

e.g.

- Euclidean distance
- Correlation coefficient
- etc.

Clustering


Clustering
Hierarchical
Bayesian
Divisive
Agglomerative
Decision Based
Nonparametric
Partitional
K-means
Centroid
Model Based
Graph Theoretic
Spectral

Clustering

- **Hierarchical** algorithms find successive clusters using previously estabilished clusters. These algorighms can either agglometrative (*bottom-up*) or divisive (*top-down*):
  - **Agglometrative algorithms** begin with each elements as a separate cluster and merge them into successive larger clusters;
  - **Divisive algorithms** begin with the whole set and proceed to divide it into successively smaller clusters.
- **Partitional** algorithems typically determine all clusters at once, but can also be used as divisive algorithms in the hierarichical clustering.
- **Bayesian** algorithms try to generate a *posteriori distribution* over the collection of all parittions of the data.

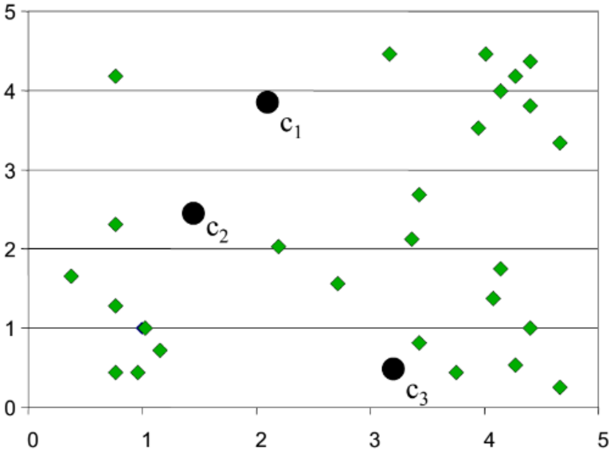
K-means

An iterative clustering algorithm

- Initialize: Pick K random points as cluster centers
- Alternate:
  - 1. Assign data points to closest cluster center
  - 2. Change the cluster center to the average of its assigned points
- Stop when no points' assignments change

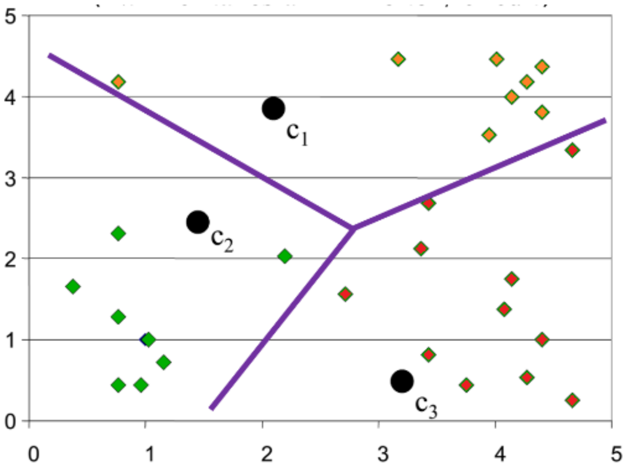
K-Means

Step 1. Randomly initialize cluster centers



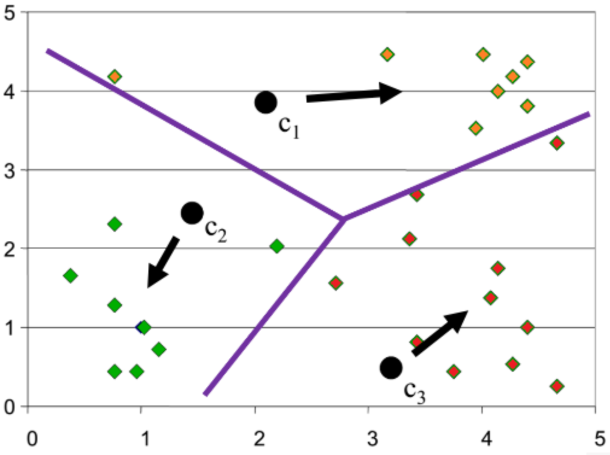
K-Means

Step 2. Determine cluster membership for each input



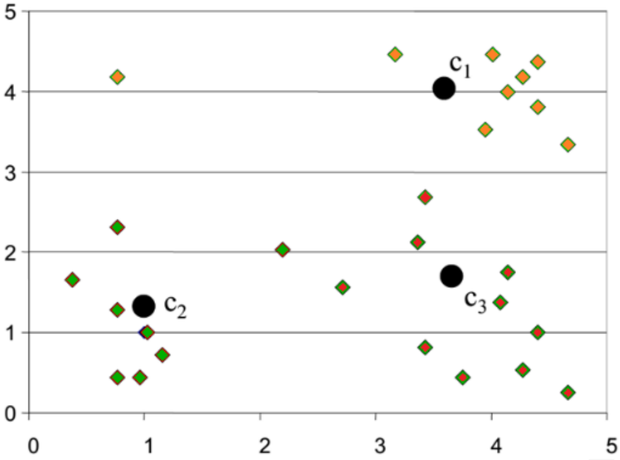
K-Means

Step 3. Re-estimate cluster centers



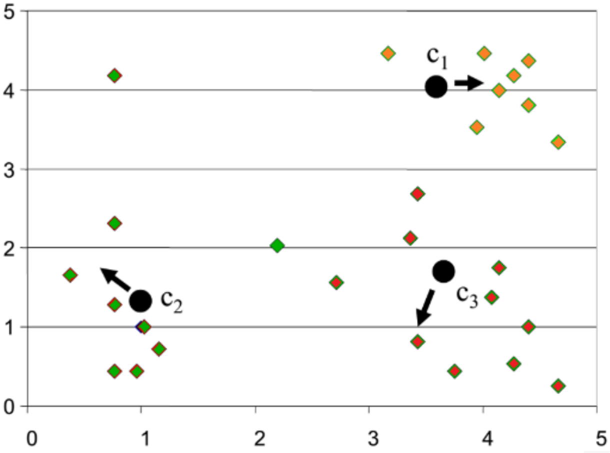
K-Means

Result of the first iteration



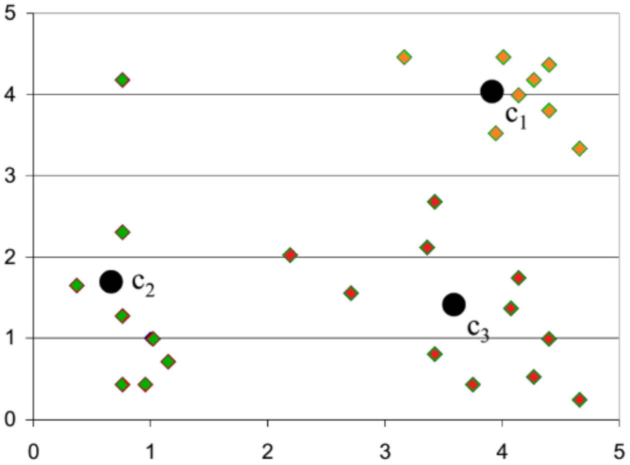
K-Means

Second Iteration



K-Means

Result of the second iteration



K-means convergence (stopping) criterion

- no (or minimum) re-assignments of data points to different clusters
- no (or minimum) change of centroids
- minimum decrease in the sum of squared error (SSE)

$$SSE = \sum_{j=1}^{k} \sum_{\mathbf{x} \in C_j} d(\mathbf{x}, \mathbf{m}_j)^2$$

- C_j is the jth cluster,
- $\mathbf{m}_j$ is the centroid of cluster C_j (the mean vector of all the data points in C_j),
- $d(\mathbf{x}, \mathbf{m}_j)$ is the (Eucledian) distance between data point $\mathbf{x}$ and centroid $\mathbf{m}_j$.

Summary: K-Means

Strength

- Simple, easy to implement and debug
- Intuitive objective function: optimizes intra-cluster similarity
- Relatively efficient.

Weakness

- Applicable only when mean is defined, what about categorical data?
- Often terminates at a local optimum. Initialization is important.
- Need to specify K, the number of clusters, in advance.
- Unable to handle noisy data and outliers.
- Not suitable to discover clusters with non-convex shapes

Appendix

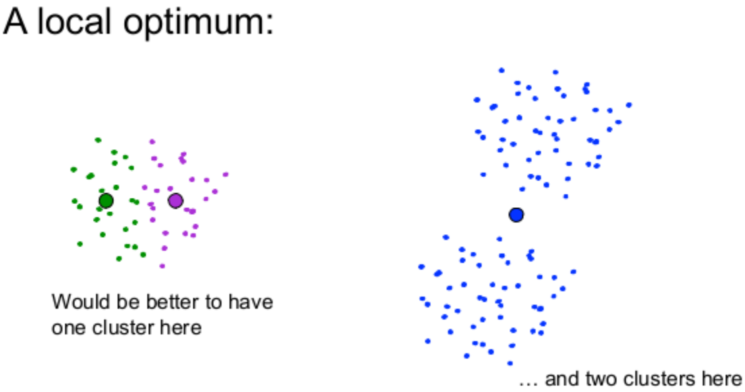
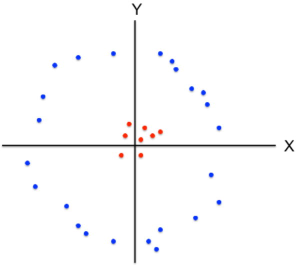
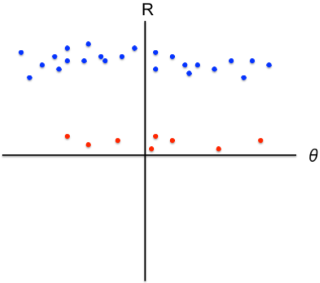


Figure: K-Means Getting Stuck

Appendix



(a) K-means not able to properly cluster



(b) Changing the features (distance function) can help

Reference I

Thank You !

$\mathcal{Q}$ & $\mathcal{A}$