UNIVERSITAT DE BARCELONA

Introduction to Machine Learning

Master in Artificial Intelligence

UPC, UB, URV

Course. Introduction to Machine Learning

Work 3. Clustering exercise

Session 2

Dr. Maria Salamó Llorente

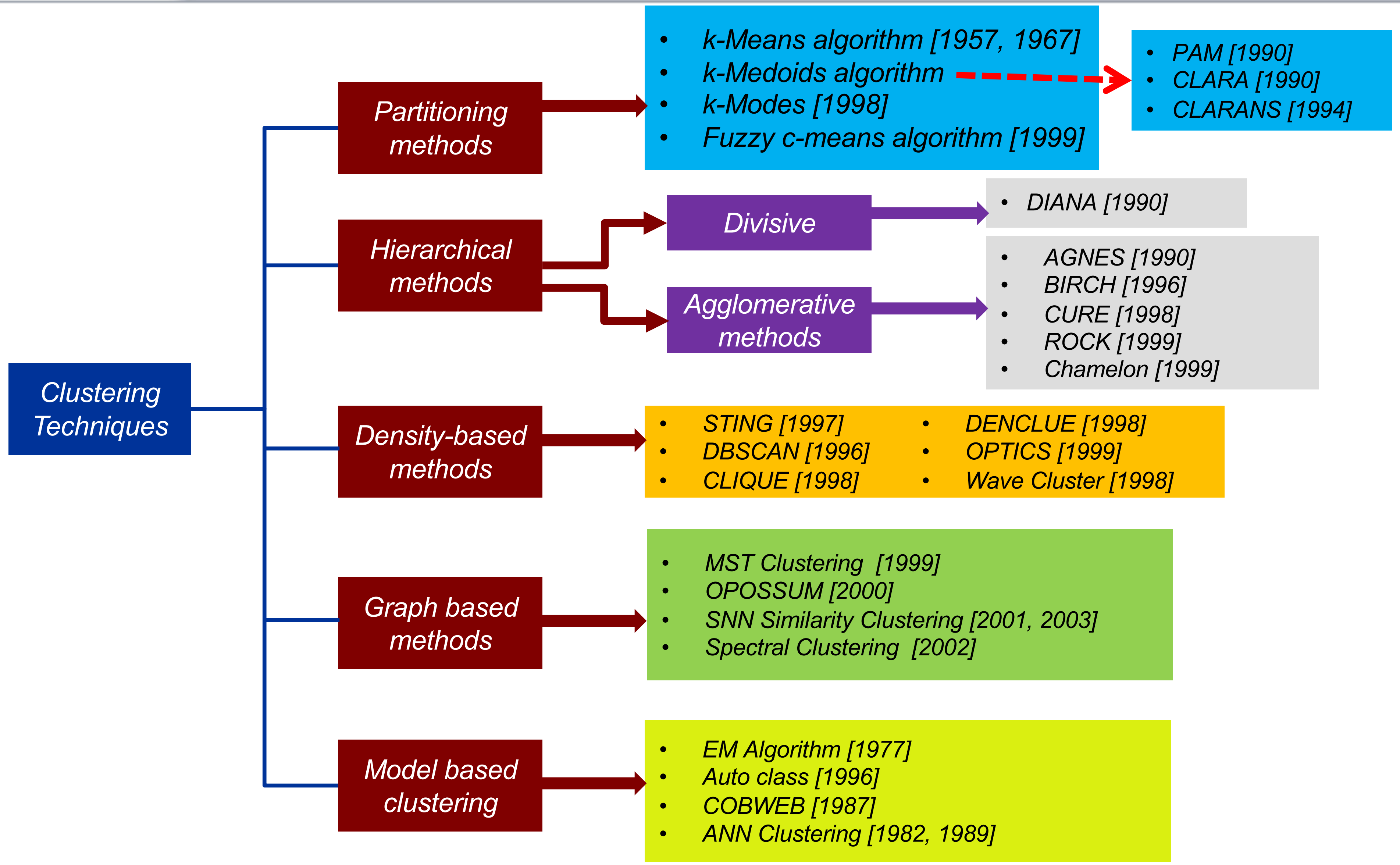
Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona

UNIVERSITAT DE BARCELONA

Contents

Taxonomy of Clustering Algorithms



4. K-MEANS AND IMPROVED K-MEANS

IMPLEMENT YOUR OWN CODE

K-Means basis

- It is a partitional algorithm that …
  - Assumes instances are **real-valued vectors**
  - Clusters based on ***centroids, center of gravity***, or **mean of points** in a cluster, ***c***:

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

  - Reassignment of instances to clusters is **based on distance** to the current cluster centroids
    - Manhattan distance (L_1 norm), Euclidean distance (L_2 norm), Cosine similarity

K-Means basis

Algorithm	Basic K-means algorithm.
1:	Select K points as initial centroids.
2:	**repeat**
3:	Form K clusters by assigning each point to its closest centroid.
4:	Recompute the centroid of each cluster.
5:	**until** Centroids do not change.

Discussion on the K-Means method

- K-Means clustering often **terminates at a local optimal**
  - Initialization can be important to find high-quality clusters
- **Need to specify K**, the number of clusters, in advance
  - There are ways to automatically determine the "*best*" K
  - In practice, one often runs a range of values and selected the "*best*" K value
- **Sensitive to noisy data and outliers**
  - Variations: Using K-medians, K-medoids, etc.
- K-Means is applicable only to objects in a **continuous n-dimensional space**
  - Using the K-Modes for **categorical data**
- Non suitable to discover clusters with **non-convex shapes**
  - Using density-based clustering, kernel k-means, etc.

Variations of K-Means

- There are many variants of the K-Means methods, varying different aspects
  - Choosing better initial centroid estimates
    - K-Means++, Intelligent K-Means, Genetic K-Means
  - Choosing different representatives for the clusters
    - K-Medoids, K-Medians, K-Modes
  - Applying feature transformation techniques *(explained at the supervised part of the course)*
    - Weighted K-Means, Kernel K-Means

Initialization of K-Means

- Different initializations may generate rather different clustering results
- Original proposal (MacQueen,1967): selects the k seed randomly
  - Need to run the algorithm multiple times using different seeds
- There are many methods proposed for better initialization of K seeds
  - **K-Means++** (Arthur and Vassilvitskii,2007):
    - The first centroid is selected randomly
    - The next centroid selected is the one that is farthest from the currently selected (selection is based on a weighted probability score)
    - The selection continues until K centroids are obtained

k-Means++

- K-Means algorithm is sensitive to the initialization of the centroids or the mean points
- K-Means++ ensures a smarter initialization of the centroids and improves the quality of the clustering
  - The initialization is different
  - The remaining of the algorithm is the same as standard k-Means

Arthur, D.; Vassilvitskii, S. (2007). **K-means++: the advantages of careful seeding**. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

Improved K-Means algorithms

Area of improvement	Reference	Method
Automatic specification of k	Pelleg and Moore	Optimization of BIC or AIC
Automatic specification of k	Hamerly and Elkan	Statistical text on cluster centers
Improved initial centroids selection	Arthur and Vassilvitskii	Datapoint’s probability contribution
Improved initial centroids selection	Ismkhan	Cluster elimination and division
Improved initial centroids selection	Likas, Vlassis, and Verbeek	Dynamic adding of cluster center
Improved initial centroids selection	Fahim et al.	Selection rule

Some k-Means references

- MacQueen, J. B. (1967). **Some Methods for classification and Analysis of Multivariate Observations**. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

- Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013). **A comparative study of efficient initialization methods for the k-means clustering algorithm**. Expert Systems with Applications. 40 (1): 200–210.
- Arthur, D.; Vassilvitskii, S. (2007). **K-means++: the advantages of careful seeding**. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

Some k-Means references

- H. Ismkhan, **I-K-means-+:** An iterative clustering algorithm based on an enhanced version of the K-means, Pattern Recognition, 79 (2018) 402–413.
- A. Likas, N. Vlassis, J.J. Verbeek. The **global K-means** clustering algorithm, Pattern Recognition, 36 (2) (2003) 451–461.
- Pelleg, Dan & Moore, Andrew. (2002). **X-means**: Extending K-means with Efficient Estimation of the Number of Clusters. International Conference on Machine Learning.
- G. Hamerly, C. Elkan, **G-Means**: Learning the k in K-means, Adv. Neural Inf. Proces. Syst. 16 (2003).
- A.M. Fahim, A.M. Salem, F.A. Torkey, M. Ramadan. An efficient **enhanced K-means** clustering algorithm, Journal of Zhejiang University-Science A 7 (10) (2006) 1626–1633.

5. FUZZY CLUSTERING

IMPLEMENT YOUR OWN CODE

Fuzzy Clustering

- Data points are given partial **degree of membership** in multiple nearby clusters

- Central point in the fuzzy clustering is always no unique partitioning of the data in a collection of clusters

- In this membership value is assigned to each cluster. Sometimes this membership has been used to decide whether the data points belong to the cluster or not

Fuzzy C-Means Clustering (FCM)

- Several approximations
  - **FCM**: Fuzzy C-Means Clustering (Bezdek, 1981)
  - **PCM**: Possibililistic C-Means Clustering (Krishnapuram - Keller, 1993)
  - **FPCM**: Fuzzy Possibililistic C-Means (N. Pal - K. Pal - Bezdek, 1997)
  - **GIFP-FCM**: Generalized Fuzzy C-Means (Zhu et al., 2009)
  - **s-FCM**: Suppressed Fuzzy C-Means (Fan et al., 2003)

- The most well-known fuzzy clustering algorithm is FCM
- Bezdek introduced the idea of a fuzzification parameter (m) in the range [1, n]
  - When m = 1 the effect is a crisp clustering of points
  - When m >1 the degree of fuzziness among points i the decision space increases

Fuzzy C-Means Clustering

Iterative FCM algorithm

- Guess Initial Cluster Centers $V_0 = \left(V_{1,0}, \ldots, V_{c,0}\right) \in \mathcal{R}^{cp}$
- Alternating Optimization (AO)

t ← 0

REPEAT

t ← t +1

Compute matrix U_t *(Eq.1)*

Compute associated clusters centers V_t (Eq.2)

UNTIL (t = T or $\| V_t - V_{t-1} \| \leq \varepsilon$)

$(U, V) \leftarrow (U_t, V_t)$

UNIVERSITAT DE BARCELONA

References of Fuzzy Clustering

- Fuzzy C-Means (**FCM**) - Bezdek, J.C.: **Pattern recognition with fuzzy objective function algorithms**. Plenum, New York (1981)
- J. C. Bezdek, R. Ehrlich, W. Full (1984). **FCM: The fuzzy C-Means Algorithm**.
- James C. Bezdek, James Keller, Raghu Krishnapuram and Nikhil R. Pal (1999), *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, TA 1650.F89.
- Improved Fuzzy partition (**GIFP-FCM**). There are two alternatives: (i) Höppner, F., Klawonn, F.: **Improved fuzzy partition** for fuzzy regression models. Int. J. Approx. Reason. 5, 599–613 (2003); or (ii) Zhu, L., Chung, F.L., Wang, S.: **Generalized fuzzy c-means clustering** algorithm with improved fuzzy partition. IEEE Trans. Syst. Man Cybern. B. 39, 578–591 (2009)
- Suppressed FCM (**s-FCM**): Fan, J.L., Zhen, W.Z., Xie, W.X.: **Suppressed fuzzy c-means clustering algorithm**. Patt. Recogn. Lett. 24, 1607–1612 (2003) or (**gs-FCM**) Szilágyi, L., Szilágyi, S.M.: Generalization rules for the suppressed fuzzy c-means clustering algorithm. Neurocomput. 139, 298–309 (2014)

- R. Krishnapuram and J.M. Keller (1993). A possibilistic approach to clustering, IEEE Transactions on Fuzzy Systems, Vol 1. No. 2, pp. 98-110.
- N.R. Pal, K. Pal, and J.C. Bezdek (1997). A Mixed c-means clustering model. Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, Vol. 1, pp. 11-21.

Course. Introduction to Machine Learning

Work 3. Clustering exercise

Session 2

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona