

Course. Introduction to Machine Learning

Work 1. Clustering Exercise

Session 2

Course 2020-2021

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,
Faculty of Mathematics and Informatics,
University of Barcelona

1. Introduction (session 1)
2. Preprocess the data (session 1)
3. DBSCAN with sklearn (session 1)
4. K-Means (your own code) (session 2)
5. Bisecting K-Means (your own code) (session 2)
6. K-medians, K-means++ or k-harmonic means (your own code) (session 2)
7. Fuzzy clustering (your own code) (session 3)
8. Validation techniques (using sklearn validation metrics) (session 3)



UNIVERSITAT DE BARCELONA



K-Means

Implement your own code

- 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 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.

- 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

- 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



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.



Note all the documents with this icon are in a zip file in campus virtual



UNIVERSITAT DE BARCELONA



Bisecting K-Means

- Bisecting k-Means is an extension of k-Means towards hierarchical clustering
- It starts with all objects in a single cluster
- It splits one cluster in 2 sub-clusters until the k number of clusters is reached
- Different ways to choose the cluster which is to be split:
(a) largest cluster, (b) the most heterogeneous cluster, (c) the largest cluster which has a predetermined degree of heterogeneity,

Basic Bisecting K-means Algorithm for finding K clusters

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic k-Means algorithm (*Bisecting step*)
3. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.

Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

- The critical part is which cluster to choose for splitting. And there are different ways to proceed, for example, you can choose the biggest cluster or the cluster with the worst quality or a combination of both.



Source: "A comparison of document clustering techniques", M. Steinbach, G. Karypis and V. Kumar. Workshop on Text Mining, KDD, 2000.



UNIVERSITAT DE BARCELONA



k-Medians

- Medians are less sensitive to outliers than means
 - Think of the median salary vs. mean salary of a large firm when adding a few top executives!
- K-Medians: Instead of taking the mean value of the object in a cluster as a reference point, medians are used (L_1 -norm as the distance measure)
- The criterion function for the K-Medians algorithm:

$$S = \sum_{k=1}^K \sum_{x_i \in C_k} |x_{ij} - med_{kj}|$$



UNIVERSITAT DE BARCELONA



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.






UNIVERSITAT DE BARCELONA



K-Harmonic means

- K-Harmonic means is a center-based partition clustering and it is similar to k-Means
 - It selects k initial centroids in the beginning
 - KHM uses harmonic averages of the distances from each data point to the centers as components of its performance function
- There is no hard cluster membership $(0,1)$ of each data point
 - A data point may have membership to each cluster
- It converges faster than k-Means
 - When the initialization is far from the local optimal

-  Dolfing, H. “**K-Harmonic Means Clustering Algorithm**”, Seminar Actual Developments in Visual Data Mining
-  Zhi X., Fan J. (2010) “**Some Notes on K-Harmonic Means Clustering Algorithm**”. In: Quantitative Logic and Soft Computing 2010. Advances in Intelligent and Soft Computing, vol 82. Springer, Berlin, Heidelberg.
https://doi.org/10.1007/978-3-642-15660-1_36
-  Ahmad, A., Hashmi, S. (2016) “**K-Harmonic means type clustering algorithm for mixed datasets**”, Applied Soft Computing. <https://doi.org/10.1016/j.asoc.2016.06.019>