Unsupervised learning: Clustering

Machine Learning and Deep Learning

Aniello Panariello, Emanuele Frascaroli, Lorenzo Bonicelli, Matteo Boschini

October 28th, 2022

University of Modena and Reggio Emilia

K-means

K-means

- Is a partitional clustering model
  - splits data $\{x_i\}_1^n$ into k disjoint sets
  - the number of sets k has to be provided as input
- solves the following optimization problem:

$$\underset{\{c_1,\dots,c_k\}}{\arg\min} = \sum_{j=1}^{k}\sum_{i=1}^{n} \mathbf{I}(i,j)||x_i - c_j||^2$$

$$\mathbf{I}(i,j) = \begin{cases} 1, & x_i \text{ belongs to cluster } j \\ 0, & \text{otherwise} \end{cases}$$

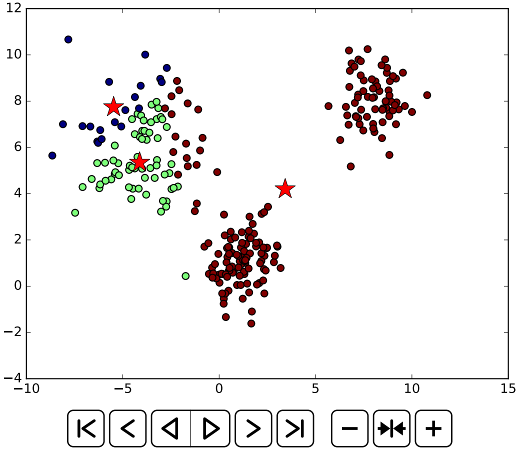
K-means: algorithm

The problem is NP-hard. A simple heuristic algorithm can be employed to converge to a *local* minimum:

- Initialize k centers randomly
- Repeat until convergence:
  - assign each example to the closest center (i.e. lower euclidean distance)
  - re-estimate centers as the mean of their clusters

Try to implement it from scratch!

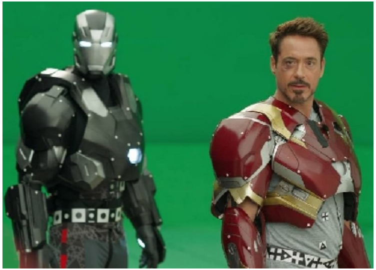
K-means: iterations



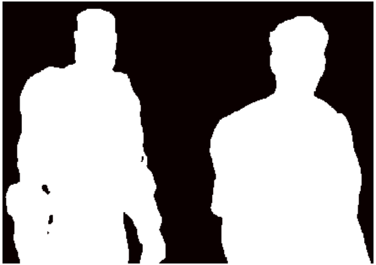
Application: color segmentation

K-means can be employed for image segmentation, simply by grouping pixels in the color space. You can also add coordinates to each pixel to obtain a smooth output.

Image



Segmentation

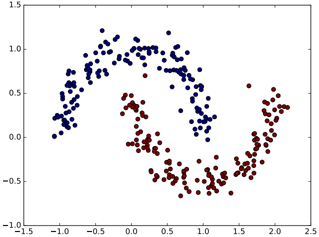


Try it!

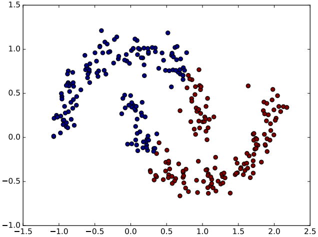
Kmeans: limitations

- it can get stuck into bad local minima
  - OPTIONAL: run the algorithm many times and choose the most recurrent solution
- can only be employed in spaces where the mean operation is defined
- due to its cost function, it can only cope with compact ball-shaped clusters

GT



Result


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5