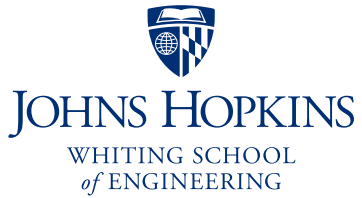
Johns Hopkins Engineering

Applied Machine Learning for Mechanical Engineers

Unsupervised Machine Learning Techniques, Part 1, A

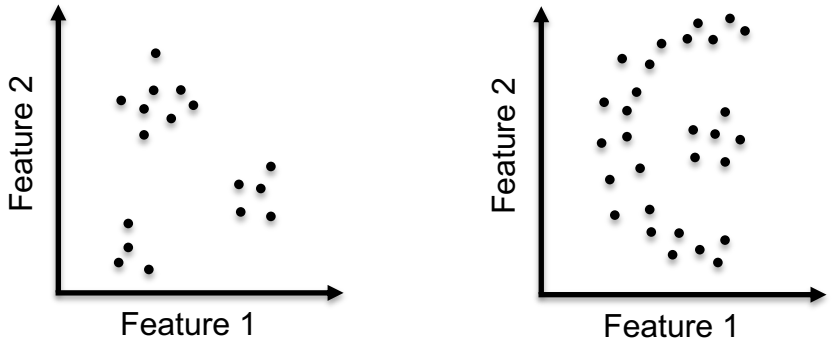


Clustering Algorithms

- By the end of this lecture you will be able to:
  - Describe clustering in general
  - Describe K-mean clustering algorithm
  - Describe density-based spatial clustering of applications with noise (DBSCAN)

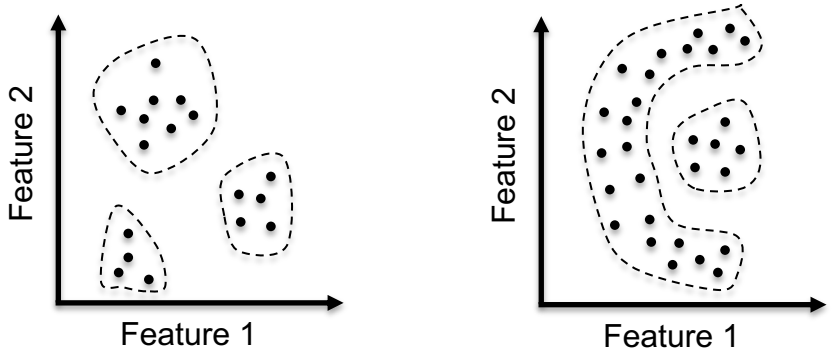
Clustering Algorithms

- Clustering
  - Assign a category to a number of datapoints with similar patterns



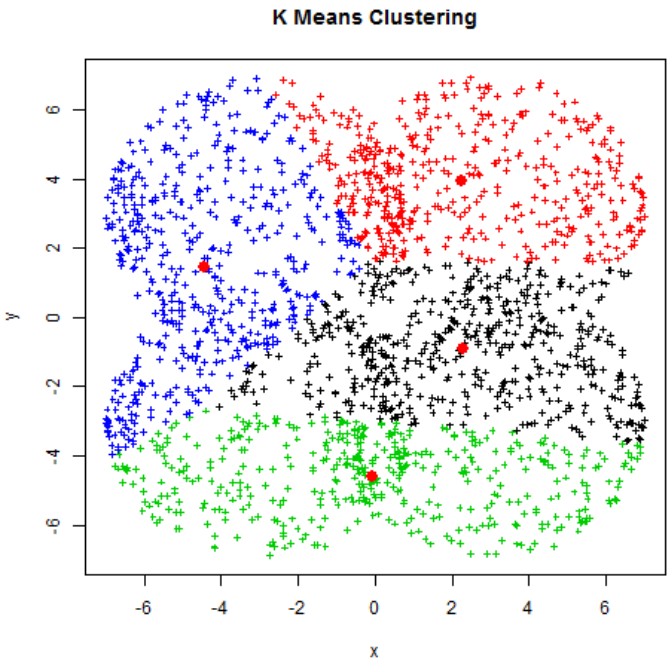
Clustering Algorithms

- Clustering
  - Assign a category to a number of datapoints with similar patterns



Clustering Algorithms

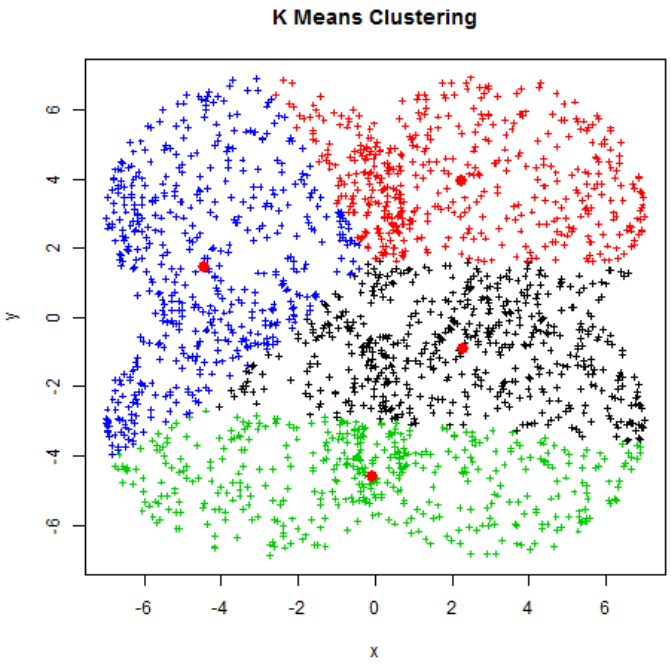
- K-Mean
  - Cluster the data into K clusters (iterative)
    - Step 1: assign K random coordinates (i.e., datapoints), referred to as centroids, representing the centroid of the clusters within the feature space.
    - Step 2: for every centroid, compute the distance between every datapoint and that centroid.
    - Step 3: assign a datapoint to the cluster of the centroid with minimum distance from.
    - Step 4: compute the new centroids of each cluster and repeat steps 2 to 4.
    - Stopping criteria: insignificant changes to the location of centroids or reaching to the total number of iterations.



https://miro.medium.com/max/960/0*FBFDk2FPSMaXAV6x.gif

Clustering Algorithms

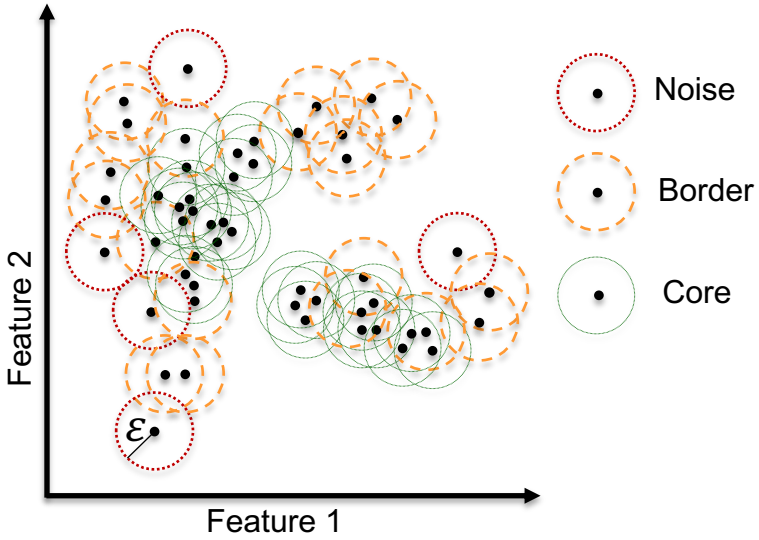
- K-Mean
  - Simple implementation
  - Convergence is guaranteed
  - K is not being automatically identified
  - Does not exclude outliers
  - Relies on distances rather than density



https://miro.medium.com/max/960/0*FBFDk2FPSMaXAV6x.gif

Clustering Algorithms

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - Jargon
    - Radius or epsilon (ε): the radius of a hypersphere with centroid to be a datapoint.
    - Minimum Points threshold (M): minimum number of datapoints within a hypersphere (excluding the centroid) to call that datapoint a "core point."
    - Border point: the datapoint which at least one but less than M other datapoints fall within its hypersphere.
    - Noise datapoint: a datapoints with no other datapoints within its hypersphere.

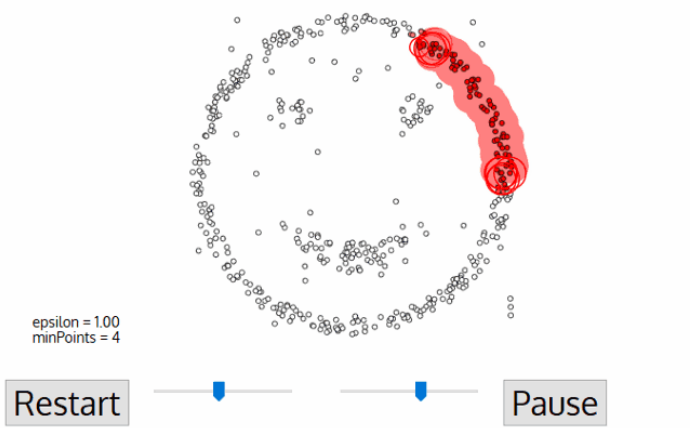


Example with $M = 3$

http://primo.ai/index.php?title=Density-Based_Spatial_Clustering_of_Applications_with_Noise_(DBSCAN)

Clustering Algorithms

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - Core and border points with overlaps are clustered together
  - Great for high density clustering
  - Discover outliers (noises are not clustered)
  - Computationally intensive



http://primo.ai/index.php?title=Density-Based_Spatial_Clustering_of_Applications_with_Noise_(DBSCAN)

Clustering Algorithms

- In this lecture, you learned about:
  - Clustering in general
  - K-mean clustering algorithm
  - Density-based spatial clustering of applications with noise (DBSCAN)
- In the next lecture, we will talk about Deep Belief Networks and Auto-Encoders