Clustering

Data Science with Python
CS677

Farshid Alizadeh-Shabdiz, PhD, MBA
Alizadeh@bu.edu
Fall 2021

Unsupervised Learning

- Why unsupervised learning?
- What is unsupervised Learning?

Unsupervised Learning

- More subjective
- No simple goal
- No easy assessment of the outcome
- But very important field on machine learning
  - Let data talk
  - Need unlabeled data, which is mostly easily available
- Example algorithm
  - Principle component analysis
  - Clustering algorithms to discover groups within a dataset

Clustering

- Group data to similar sets/clusters
- Why Cluster analysis?
  - Helps partition massive data to groups with similar features
  - Get insight into the data
  - Needed for the next step of analysis
    - Pattern discovery
    - Classification
    - Outlier analysis
- Clustering is unsupervised learning
  - It is different than classification, which is a supervised learning

Clustering Application

- Datamining
- Recommendation systems
- Customer segmentation
- Data summarization
- Detecting patterns and trends
- Gene sequencing

Clustering factors

- Single level or hierarchical partitioning
- Exclusive or non-exclusive (e.g. one article might belong to two classes)
- Similarity measure
  - Distance based, like Euclidean or Manhattan
  - Connectivity based, like density or contiguity
- Full space clustering or sub-space clustering

Clustering Challenges

- Discover clusters with different shapes
- Be able to detect a cluster in presence of noise
- Be able to deal with different data types
- Deal with large data set
- Deal with data with high dimensionality

Distance Functions

- Euclidean or L2-norm
- Manhattan distance or L1-norm

$$d(i,j) = \sum_{i=1}^{l} |x_{1i} - x_{2i}|$$

Note: When data is binary this is called Hamming distance.

- Minkowski distance: distance between two $l - \mathbf{dimensional}$ data points

$$d(i,j) = \sqrt[p]{\sum_{i=1}^{l} |x_{1i} - x_{2i}|^p}$$

Distance Functions - cont

- Correlation
- Gaussian Kernel function

$$\exp(\frac{||X_i - Xj||^2}{\sigma^2})$$

- Cosine similarity:

$$\cos(d_1, d_2) = \frac{d_1.d_2}{||d1||||d2||}$$

- A measure depending on your problem

Distance for Categorical Variables

- Option 1: number of mismatches over total number of variables (T) (total number of matches is M)
- $d(i, j) = \frac{T - M}{T}$
- Option 2: Using dummy variables to present categorical variables

K - Means

Clustering Algorithms

K – Means:

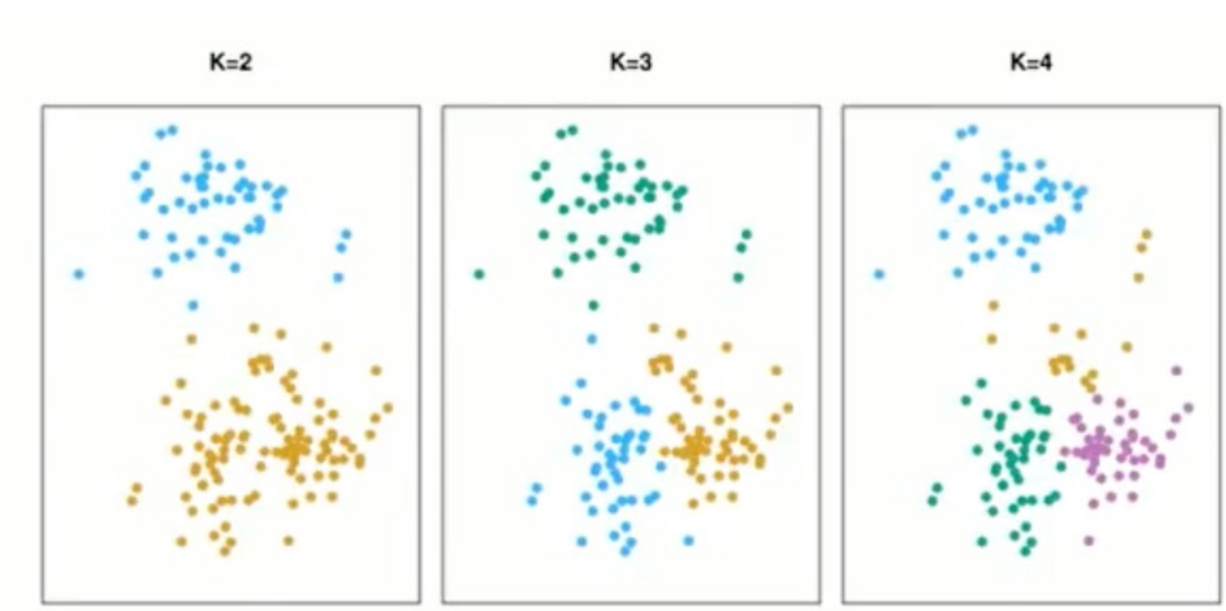
- Finding K different groups in the dataset
  - We have to define K - what is the K?

Hierarchical clustering

- Finding a tree like clusters of dataset, which is called *dendrogram*.
- This provide answer for any possible number of clusters
- There is no need to specify K

K – means Example

- Example below show k=2, 3, and 4 results



Reference: Tibshirani and Hastie – Intro to statistical learning

K – Means definition

Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

For instance, if the ith observation is in the kth cluster, then $i \in C_k$.

Reference: Tibshirani and Hastie – Intro to statistical learning

K - means

- The goal is finding K clusters, in which within the cluster variation is small
- Therefore, we need distance definitions. E.g. Euclidean.
- Therefore, K-means is trying to find clusters in such a way that

$$Minimize \sum_{i=1}^{k} (variation\ within\ class\ i)$$

- One simple measure is using Euclidean distance & minimize overall variation of pairwise points within a cluster *CN*

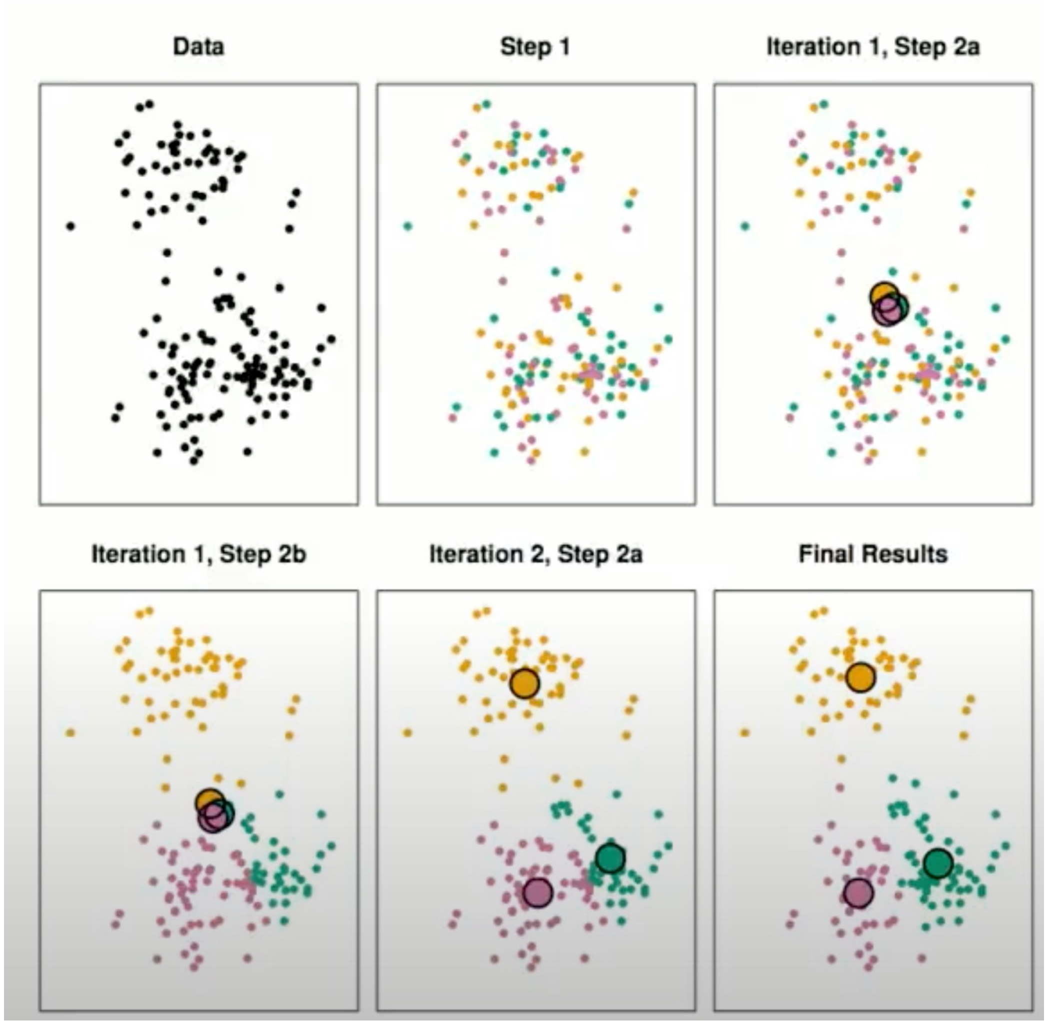
$$Minimize \sum_{i=1}^{k} \left(\frac{1}{|CN_i|} \sum_{j} \sum_{p} (x_{ij} - x_{ip})^2\right)$$

K – Means Clustering Algorithm

1. Randomly choose K random observations as initial cluster assignments
2. Iterate following steps
    1. Find cluster centroids – mean value of all the observations in all dimensions
    2. Assign observations to the cluster with closest cluster center (distance can be Euclidean distance)
    3. Stop if
        1. centroid of observations don't change
        2. After some iterations
        3. When few number of data points change cluster

Note: k-Means is also called Lloyd algorithm.

K - Means example



Reference: Tibshirani and Hastie – Intro to statistical learning

Why K-Means algorithm gets us there?

- Why the prescribed algorithm provides a solution to the K-means criteria
- It is because of

$$\sum_{i=1}^{k}\left(\frac{1}{|CNi|}\sum_{j}\sum_{p}(x_{ij}-x_{ip})^2\right)$$

$$= 2\sum_{i}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ip})^2 = 2$$

How to Choose parameters?

- Try with more than one initial points
- Try a range of K values to choose the best K

K – Means with different Initial points

Reference: Tibshirani and Hastie – Intro to statistical learning

K- means Close Family

- K Median
  - Uses median as the center of a cluster
  - Not sensitive to outliers
  - Less computation
  - Basically, minimized error in L1-norm metric
- K-Medoids
  - The center point has to be one of the data points, while in K-means that constraint is not enforced
  - Helps to better interpret the results
- K-modes
  - Handles categorical variables
- "Self Organized Maps" (SOM) is a special case of K-means and gets implemented by neural networks

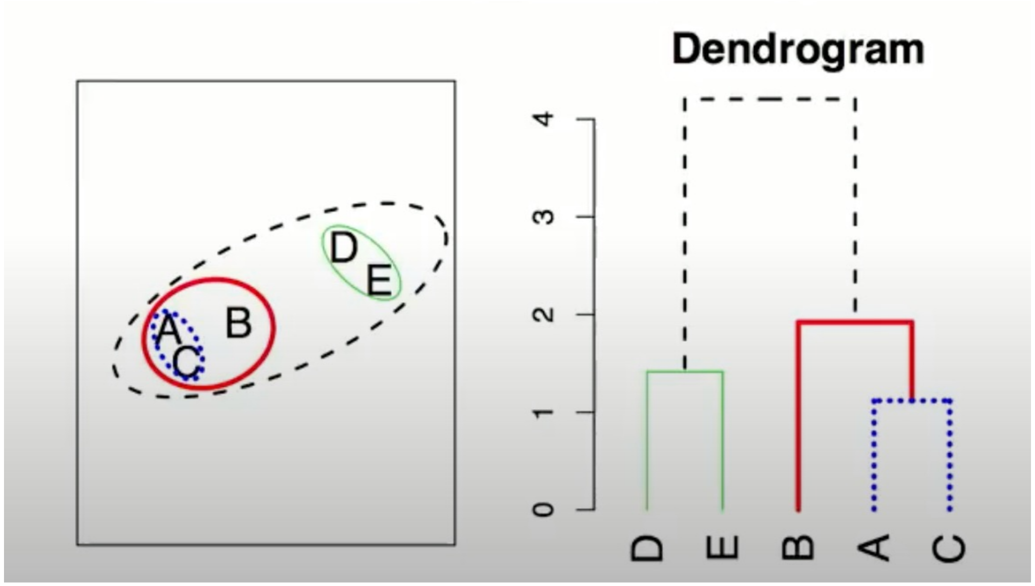
SKYHOOK®

Hierarchical Clustering

Hierarchical Clustering

- Attractive approach since no need to decide on K in advance.
- Bottom-up or agglomerative Hierarchical clustering

Example

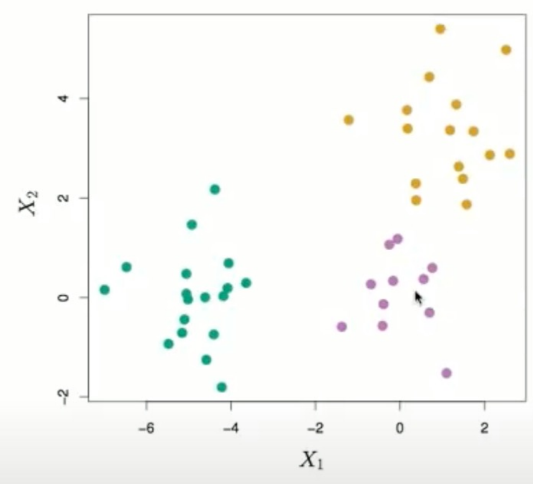


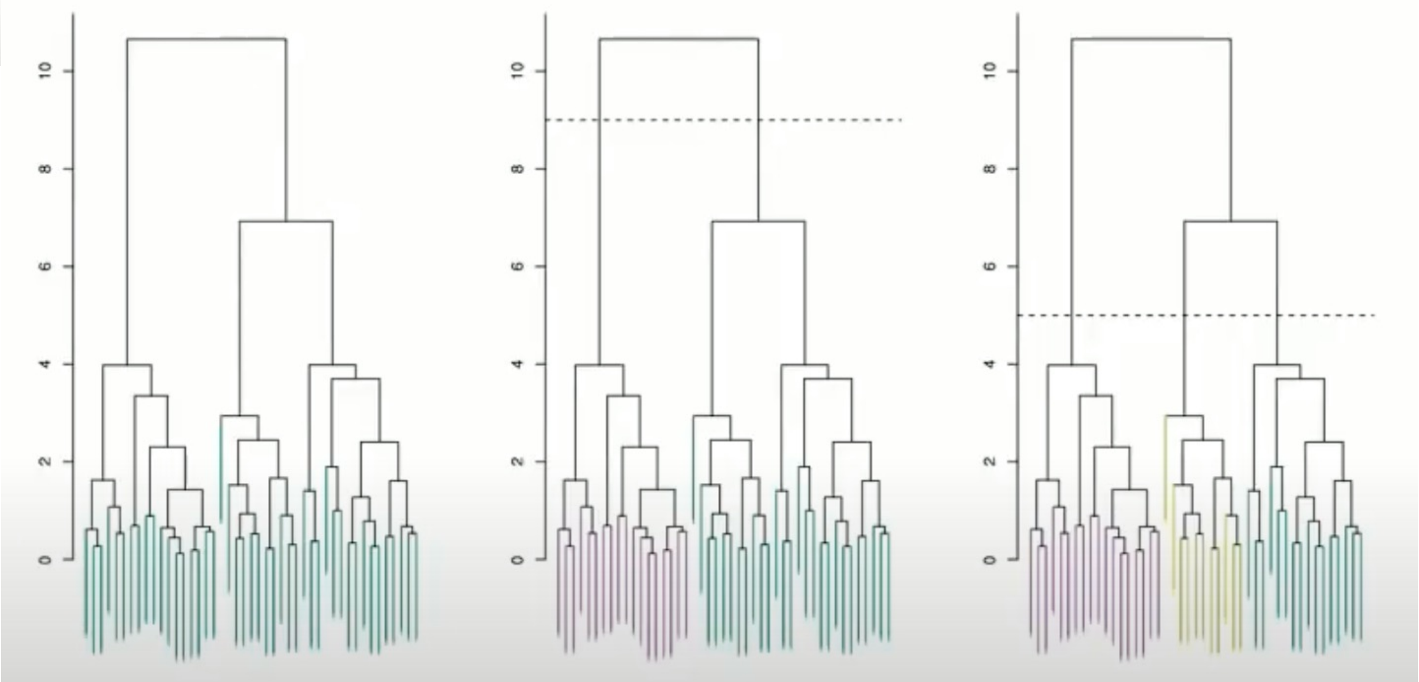
Reference: Tibshirani and Hastie – Intro to statistical learning

Hierarchical Clustering Algorithm

1. Find closest data points and link them
2. Continue with finding closest clusters and points and link them
3. Repeat until all the points are included
4. Draw dendrogram, and choose number of clusters

Hierarchical Clustering - example





Reference: Tibshirani and Hastie – Intro to statistical learning

Hierarchical Clustering – Linkage types

- Complete:
  Maximum dissimilarity between pair of samples of two cluster
- Single
  - Minimum dissimilarity between pair of samples of two cluster
- Average
  - Mean inter-cluster dissimilarity or mean value of pairwise distances between two clusters
- Centroid
  - Distance between centroid of the clusters

Practical Issues with These Clustering Techniques

- Scaling – scale of variables of observations
  - Scaling and standardization is suggested
- Distance function is important
- For K means, K value is important
- Hierarchical clustering can help to subjectively decide on number of Ks
- What features should be use for clustering

DB Scan

Density Based Spatial Clustering with Noise - DBSCAN

- Proposed in 1996
- The most common clustering algorithm
- Different than the other clustering algorithms

DBSCAN fundamentals

<u>Goal:</u>

- Finding continues region of high density

<u>Noise:</u>

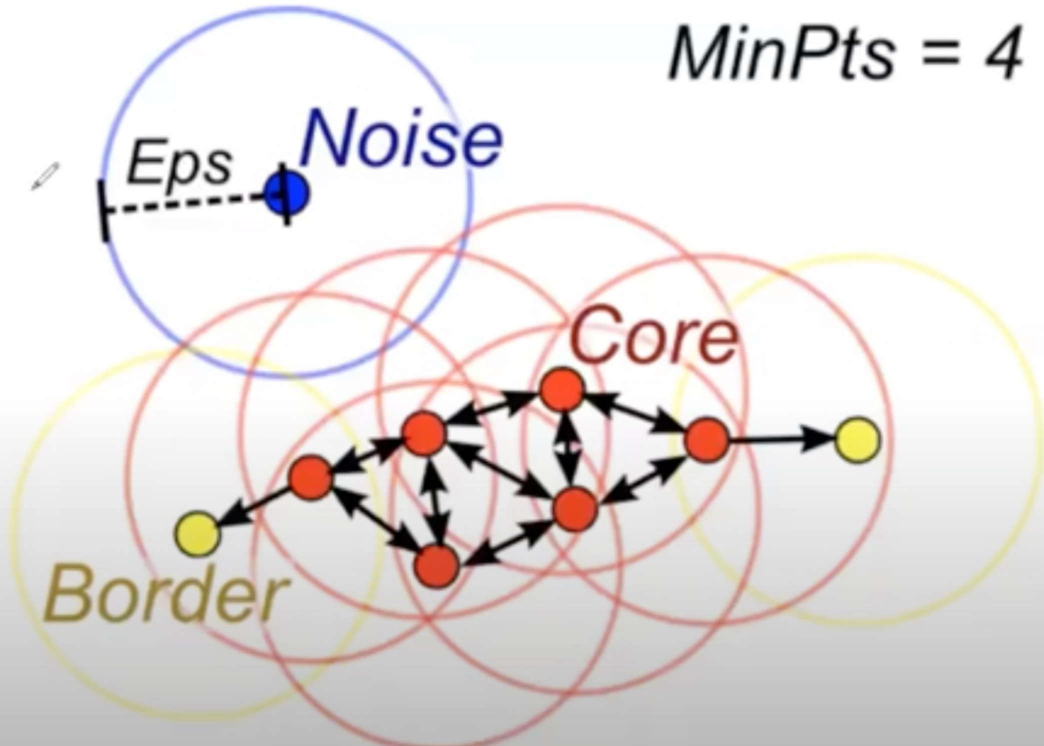
- Any point which is not close to a high density region

DBSCAN Basic Terms

- <u>Epsilon:</u> maximum radius to find neighbors
- Starting point: A point with minimum number of neighbors
- <u>M:</u> minimum number of neighbors to become a starting point
- Core points: neighbors of a starting point
- Border points: cluster points which don’t neighbor a staring point

Example

Density-Based Spatial Clustering of Applications with Noise(DBSCAN)



Red: Core Points

Yellow: Border points. Still part of the cluster because it's within epsilon of a core point, but not does not meet the min_points criteria

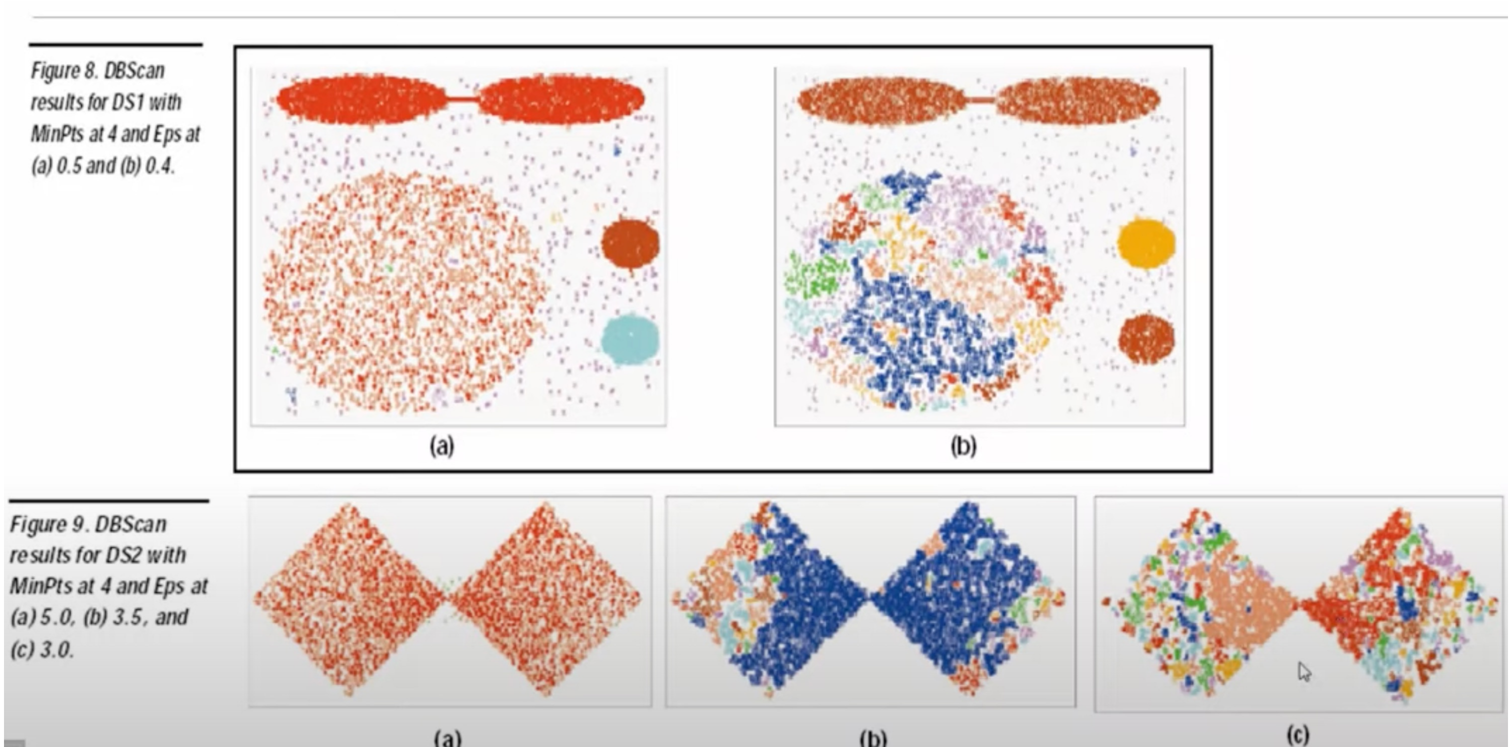
Blue: Noise point. Not assigned to a cluster

DBSCAN Algorithm

- Start with an arbitrary point P
- Find all neighbors of point P
- If number of neighbors at least = min points => P is a core point and a cluster gets formed
- If P is a border point, DBSCAN visits the next border point, and if no border point left, DBSCAN selects another point
- Continue until all the points have been processed
- Any point alone or smaller than M samples is considered noise.

DBSCAN Discussions

- Not sensitive to outliers – Detects outlier
- Need density
- Don't need to specify number of clusters
- Density across all clusters are the same
- It handles noise well
- Only check local area



Ack. Figures from G. Karypis, E.-H. Han, and V. Kumar, *COMPUTER*, 32(8), 1999

Spectral Clustering

Spectral Clustering

- It is a systematic way to find K cluster
- Have to specify K
- It can be used by using neural networks also

Steps of Spectral Clustering

1. First step is having a weighted graph or similarity graph and build similarity matrix in which

   W(i,j) = distance between node *i* and *j*

   Example of distance functions are Euclidean distance or Gaussian Kernel function $\mathbf{exp}(\frac{\|X_i - X_j\|^2}{\sigma^2})$

2. Divide samples to K groups

3. Goal is to minimize cost function = $\sum_{j=1}^{k}(W_j - \overline{W_j})$

   Note: $\overline{W_j}$ complement set of W_j means rest of the points. This is the cost of disconnecting W_j from rest of the points.

4. Solve approximation of the solution

SKYHOOK®

Approximating Spectral Clustering Solution

1. Degree matrix : $\mathbf{D} = \begin{bmatrix} d1 & & 0 \\ & \cdots & \\ 0 & & dn \end{bmatrix}$ which is a diagonal matrix, and each element is sum of the corresponding row of W (weight matrix)
2. Find $D^{-1/2}$which is easy to find : $\begin{bmatrix} {d_1}^{-1/2} & & 0 \\ & \cdots & \\ 0 & & {d_n}^{-1/2} \end{bmatrix}$
3. Find "*Normalized Laplacian Matrix*"
   $L = I_n - D^{-1/2} W D^{-1/2}$
4. Find $\mathbf{min}\, Trace(U^T L U)\, , Subject\ to\ U^T U = I_k$
5. Apply K-means clustering to rows of U

Note U has n rows (data points) and k columns (# of features)

How to find U

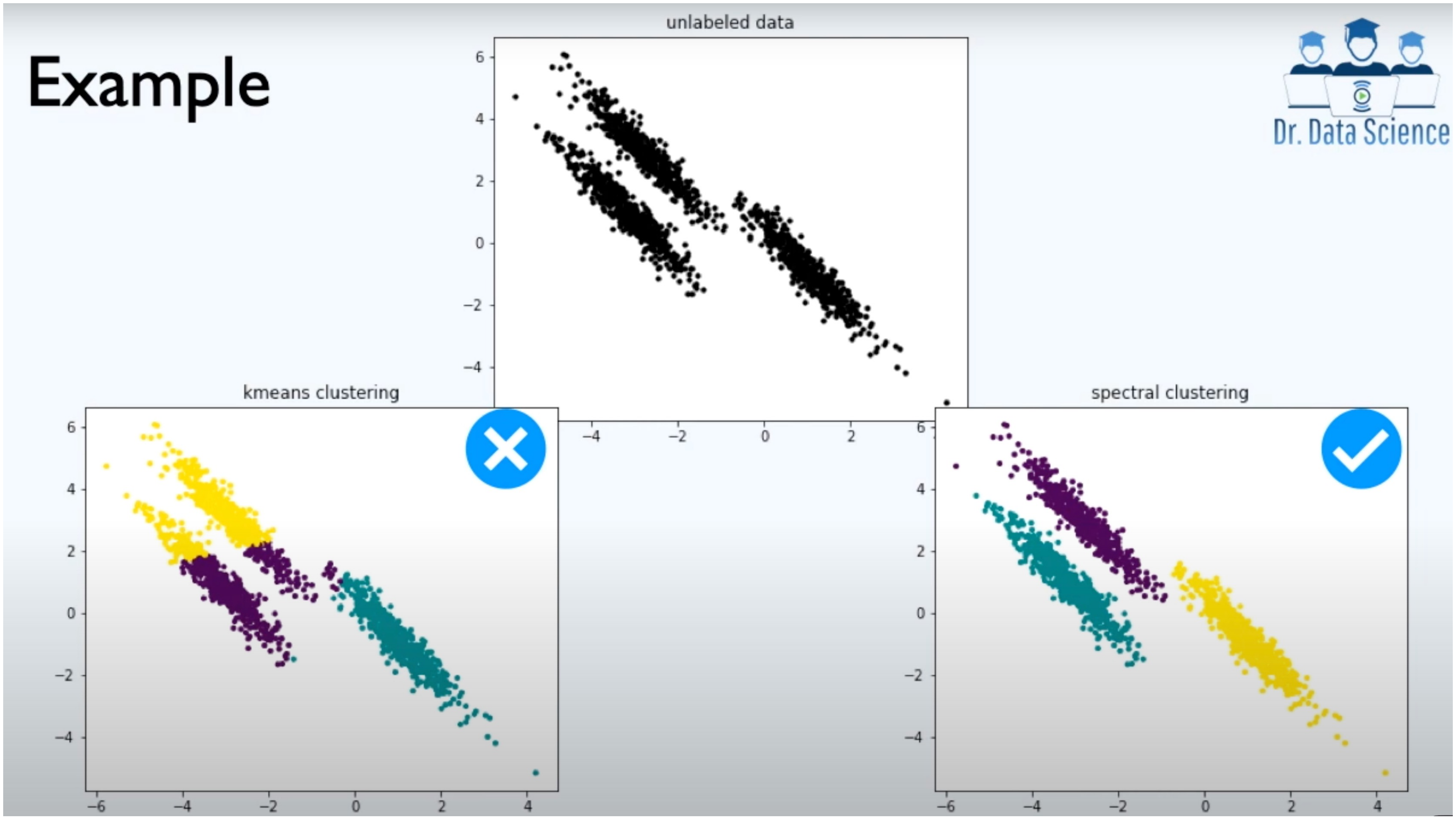
Note that this $D^{-1/2}WD^{-1/2}$ is normalized similarity matrix, and

- Applying eigen value decomposition to normalized similarity matrix gives us

$$D^{-1/2}WD^{-1/2} = V\Lambda V^T$$

- And Λ is matrix of eigen values.
- U is composition of K eigen vectors associated to the highest eigen values of normalized similarity matrix

Example of K-means and Spectral Clustering



Reference: Dr. Data Science channel on youtube

Clustering Assessment

Desirable Outcome

1. High inter-class separation – Between group variance

2. Low intra-class separation

- Main assumption is “all the clusters are the same”
- To check the assumption we apply F-Test

$$F - test = \frac{Mean\ sum\ of\ square\ between}{Mean\ sum\ of\ square\ within}$$

- This follows F distribution and

Note: this is also called CH index (Calinski-Harbusz index)

$$\text{Mean } SSB = \frac{\sum_{j=1}^{K} n_i (C_i - \bar{X}..)^2}{K - 1}$$

And

$$\text{Mean } SSW = \frac{\sum_{j=1}^{K} \sum_{i=1}^{nj} (X_i - C_j)^2}{n - K}$$

In which

- *K: is number of clusters*
- *ci*: is centroid of cluster *i*
- *ni*: is number of points in cluster *i*
- $\bar{X}..$: is average of all centroids
- X_i: is the data point

Many methods to Assess Clustering

- There are many equations combined SSB and SSW in different ways to assess clustering
- Elbow method

$$\sum_{j=1}^{K}\sum_{i=1}^{ni}(C_j + x_i)^2$$

- Hartigan Index

$$H = (\frac{SSW_K}{SSW_{K\ +\ 1}} - 1)(N - K - 1)$$

- Dunn Index

$$D = \frac{\min inter - cluster\ seperation}{\max intra - cluster\ seperation} = \frac{\min_{1 \le i \le j \le K} d(C_i, C_j)}{\max_{1 \le p \le K} diameter_p}$$