January 26, 2024 | 4:30-6:10pm

Virtual Workshop

Machine Learning with R: K-Means Clustering

u.mcmaster.ca/scds-events

Data Analysis Support Hub

SCDS

Library

McMaster University

Machine Learning with R: K-means Clustering

Seyed Amirreza Mousavi
Master’s student at McMaster University

DASH: Data Analysis Support Hub Workshop Series
26/01/2024

Lewis & Ruth Sherman Centre for Digital Scholarship
scds.ca

McMaster University | Library

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Laslovarga, “Webster Falls in Winter, Waterdown, Hamilton, Ontario, Canada - Spencer Gorge / Webster's Falls Conservation Area,” 23 January 2011, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Waterdawn_Webster_Falls_in_Winter8.jpg

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information: scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Attendance that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

Verify your participation at a session: https://u.mcmaster.ca/verification

At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

DASH: Data Analysis Support Hub Workshops

Register for upcoming DASH events: u.mcmaster.ca/scds-events

Feb 14: “Introduction to Python” – Jadon Vivek

Feb 27: “Multivariable Analysis with R” – Humayun Kabir

Mar 28: “Intermediate Python programming"– Amirreza Mousavi

Apr 30: “Survival Analysis with R” – Humayun Kabir

Etc

Book an Appointment with the DASH Team

Receive help from a member of the DASH team! DASH can assist with the following topics:

- □ Creating data visualizations, including charts, graphs, and scatter plots
- □ Figuring out which statistical tests to run (e.g., t-test, chi-square, etc.).
- □ Analyzing data with software including SPSS, Python, R, SAS, ArcGIS, MATLAB, and Excel
- □ Choosing which software package to use, including free and open-source software
- □ Troubleshooting problems related to file formats, data retrieval, and download
- □ Selecting methodology and type of data analysis to use in a thesis project

Book an appointment: https://library.mcmaster.ca/services/dash

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences. In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Supervised Vs Unsupervised Learning

What is supervised learning?

Supervised learning is a machine learning approach defined by its use of labelled datasets. These datasets are designed to train or “supervise” algorithms to classify data or predict outcomes accurately.

Supervised learning can be separated into two types of problems:

Classification problems use an algorithm to accurately assign test data into specific categories, such as separating apples from oranges

Regression is another type of supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables

Supervised Vs Unsupervised Learning

What is unsupervised learning?

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are "unsupervised").

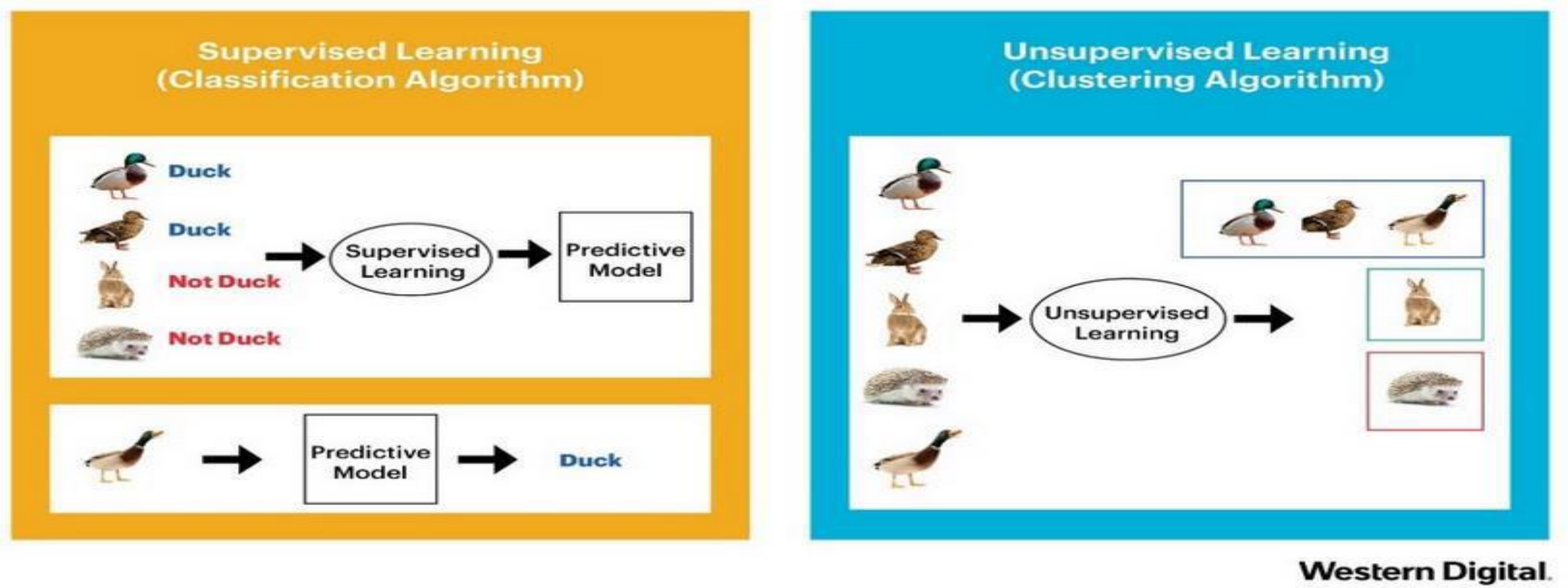
Unsupervised learning models are used for three main tasks:

Clustering is a data mining technique for grouping unlabeled data based on their similarities or differences.

Association is another type of unsupervised learning method that uses different rules to find relationships between variables in a given dataset.

Dimensionality reduction is a learning technique used when the number of features (or dimensions) in a given dataset is too high

Supervised Vs Unsupervised Learning



Source: Sanjaya, H. “Supervised vs Unsupervised Learning”. March 17, 2020. accessed on September 22, 2022. https://medium.com/hengky-sanjaya-blog/supervised-vs-unsupervised-learning-aae0eb8c4878

K-means Clustering

K-means clustering is one of the simplest and most popular unsupervised machine learning algorithms.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* (≤ *n*) sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ so as to minimize the within-cluster sum of squares. Formally, the objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \qquad \boldsymbol{\mu_i} = \frac{1}{|S_i|} \sum_{\mathbf{x} \in S_i} \mathbf{x},$$

This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{\mathbf{x}, \mathbf{y} \in S_i} \left\| \mathbf{x} - \mathbf{y} \right\|^2$$

Algorithm

The most common algorithm uses an iterative refinement technique. It is sometimes also referred to as "naïve *k*-means", because there exist much faster alternatives.

Given an initial set of *k* means $m_1^{(1)}$, ..., $m_k^{(1)}$ (see below), the algorithm proceeds by alternating between two steps:

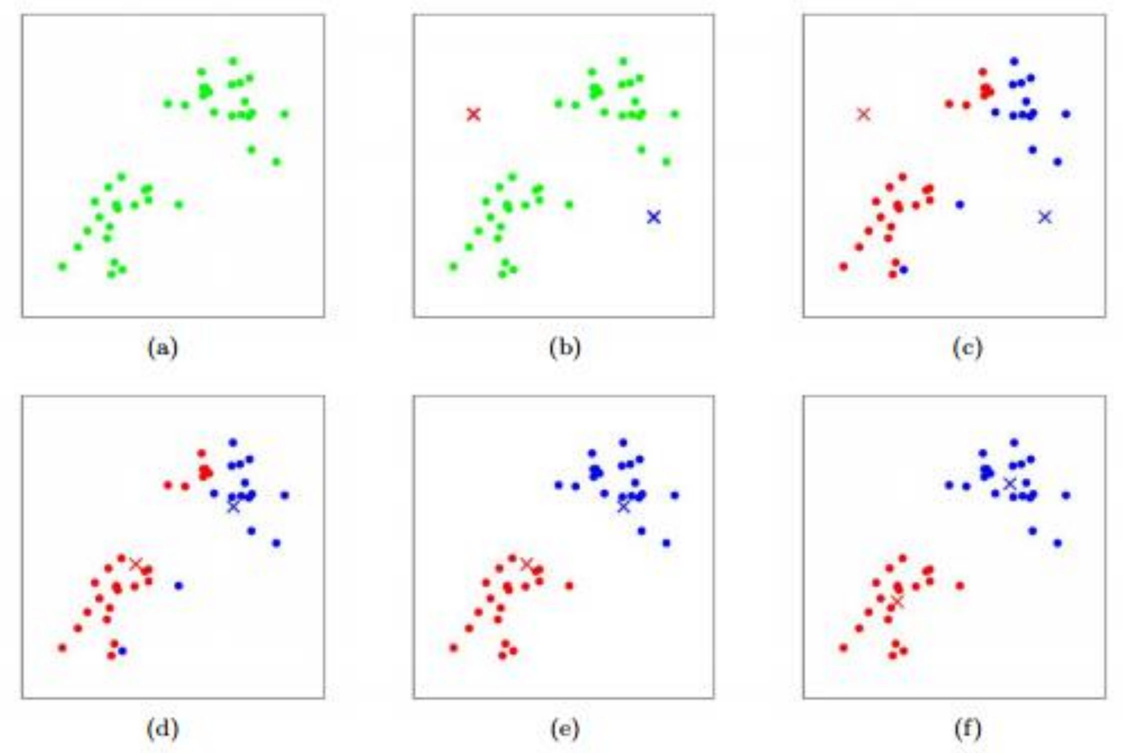
Assignment step: Assign each observation to the cluster with the nearest mean, that with the least squared Euclidean distance:

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \le \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \le j \le k \right\},$$

Update step: Recalculate means (centroids) for observations assigned to each cluster.

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j$$

Algorithm

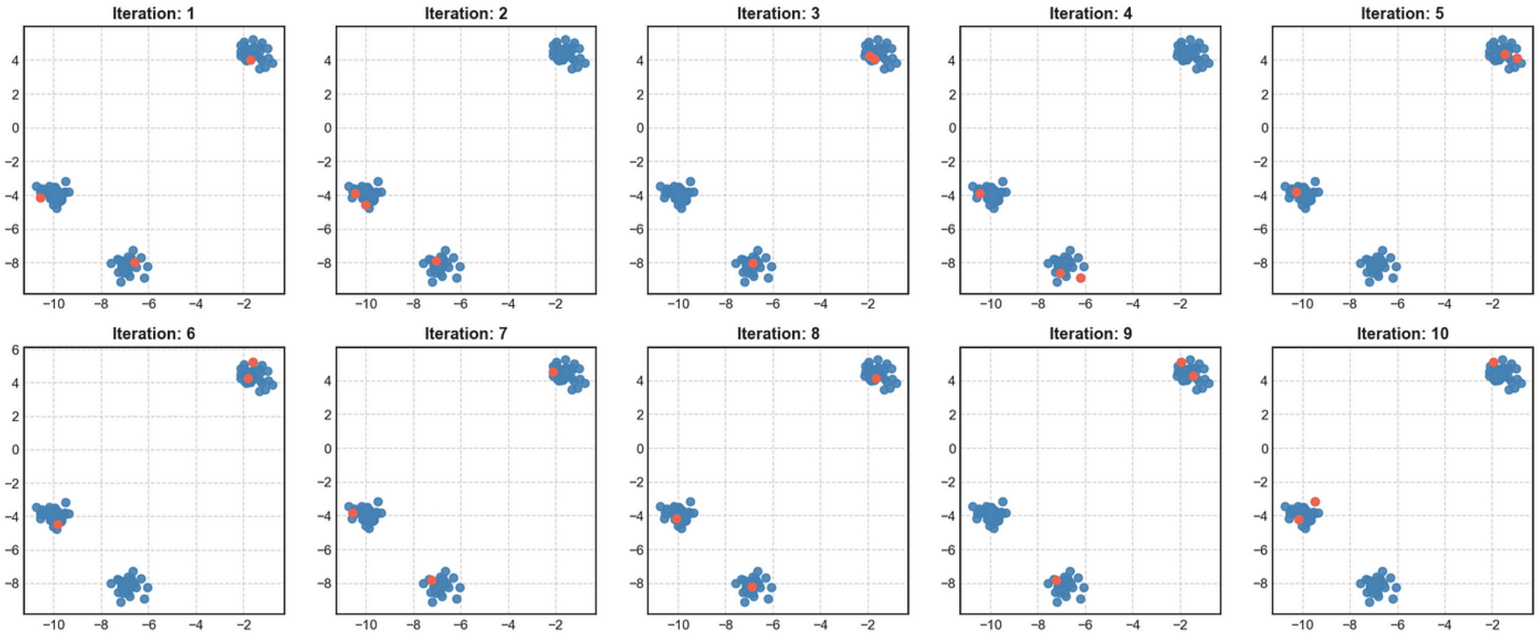


K-means algorithm: Training examples are shown as dots, and cluster centroids are shown as crosses.
(a)Original dataset.
(b) Random initial cluster centroids.
(c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (shown by "painting" the training examples the same color as the cluster centroid to which is assigned); then we move each cluster centroid to the mean of the points assigned to it.

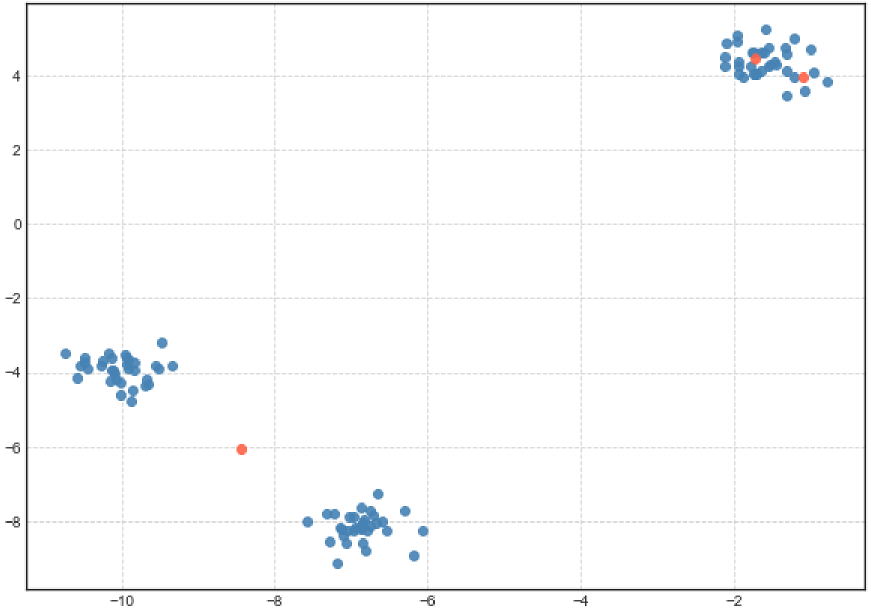
Source: Piech, C. and Ng, A. “K Means”. https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Initialization

1) Forgy: Forgy initialization involves randomly selecting K data points from the dataset as the initial centroids. The data points are chosen independently, without replacement.



10 Initial configurations chosen by Forgy's Method

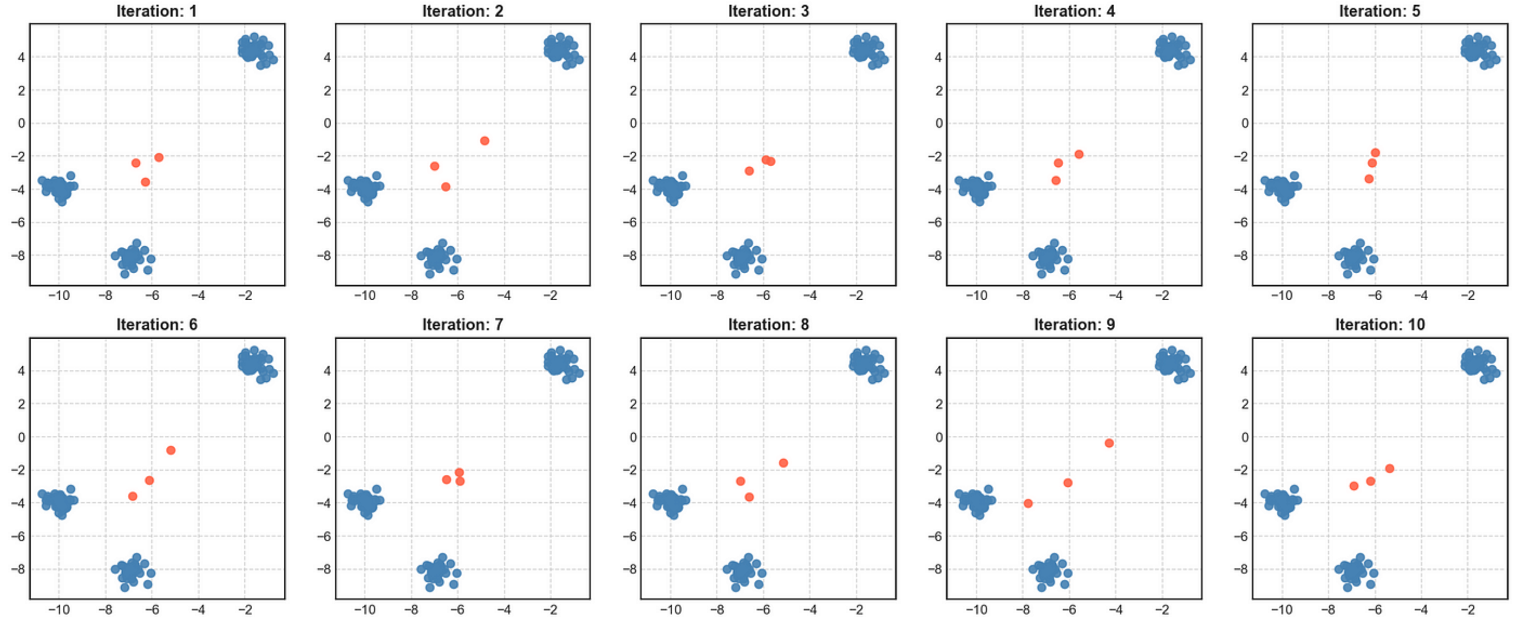


Result of running k-Means from improper initial points

Source:k-Means Clustering: Comparison of Initialization strategies. https://medium.com/analytics-vidhya/comparison-of-initialization-strategies-for-k-means-d5ddd8b0350e

Initialization

2) Random Partition: In this method, we randomly assign each point in the data to a random cluster ID. Then, we group the points by their cluster ID and take the average (per cluster ID) to yield the initial points. Random Partition method is known to yield initial points close to the mean of the Data.

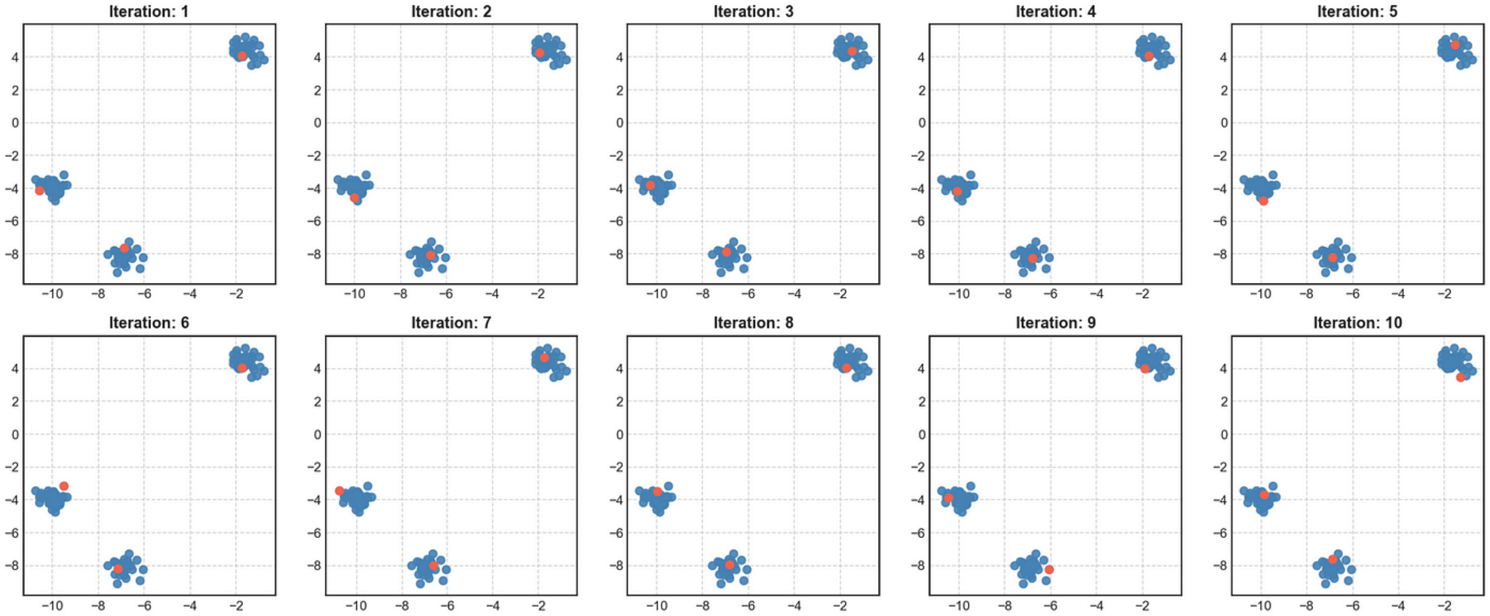


Initial Points chosen by Random Partition Method

Source: k-Means Clustering: Comparison of Initialization strategies. https://medium.com/analytics-vidhya/comparison-of-initialization-strategies-for-k-means-d5ddd8b0350e

Initialization

3) Kmeans++: K-Means++ is an improvement over random initialization. It ensures that the initial centroids are spread out across the dataset, making it more likely for K-means to converge to a better final solution. The first centroid is randomly chosen from the data points, and subsequent centroids are selected with a probability proportional to the square of the distance from the point to the nearest existing centroid.



Initial Points chosen by kmeans++ Method

Source:k-Means Clustering: Comparison of Initialization strategies. https://medium.com/analytics-vidhya/comparison-of-initialization-strategies-for-k-means-d5ddd8b0350e

Applications of Clustering in real-world scenarios

- Customer Segmentation
- Document Clustering
- Image Segmentation
- Recommendation Engines
- …

Important Factors

1. Number of clusters (K): The number of clusters you want to group your data points into, has to be predefined.
2. Initial Values/ Seeds: The choice of the initial cluster centres can have an impact on the final cluster formation. The K-means algorithm is non-deterministic. This means that the outcome of clustering can be different each time the algorithm is run even on the same data set.
3. Outliers: Cluster formation is very sensitive to the presence of outliers. Outliers pull the cluster towards itself, thus affecting optimal cluster formation.

Source: Banerji, A. “K-Mean: Getting The Optimal Number Of Clusters”. https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

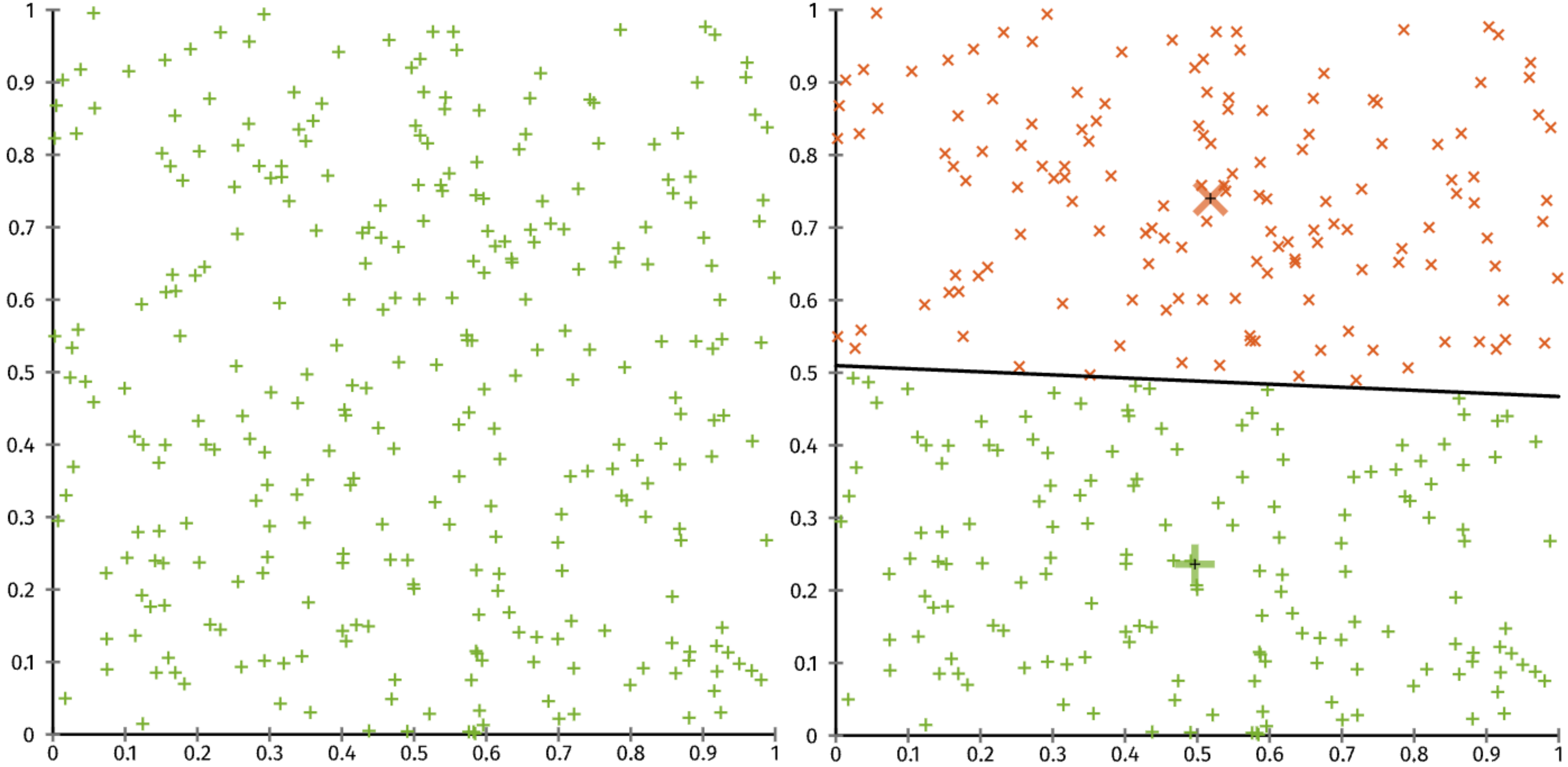
Important Factors

4. Distance Measures: Using different distance measures (used to calculate the distance between a data point and cluster centre) might yield different clusters.
5. The K-Means algorithm does not work with categorical data.
6. The process may not converge in the given number of iterations. You should always check for convergence.

Source: Banerji, A. “K-Mean: Getting The Optimal Number Of Clusters”. https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

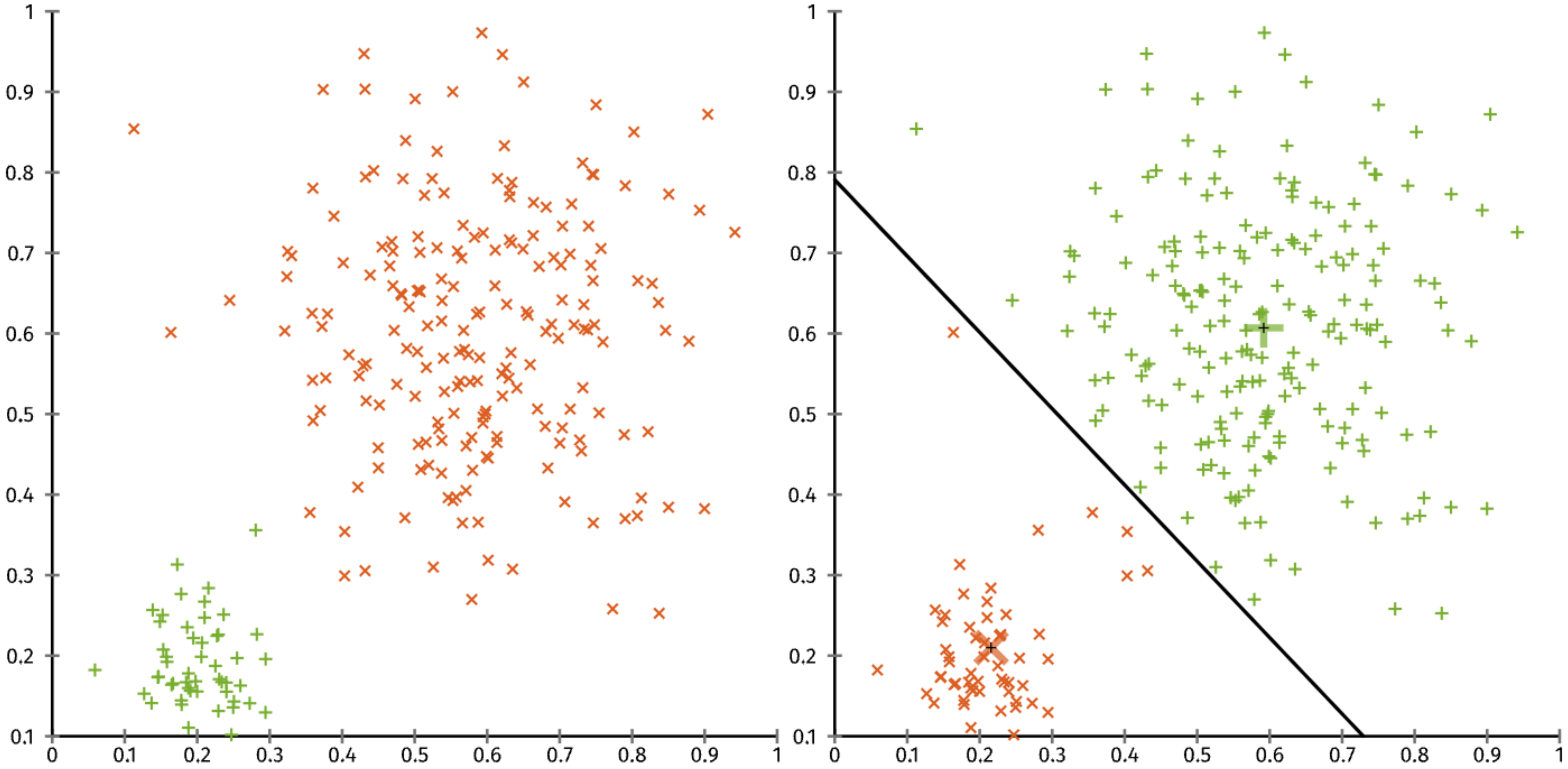
Limitations of K-means

Example – badly chosen *k*:



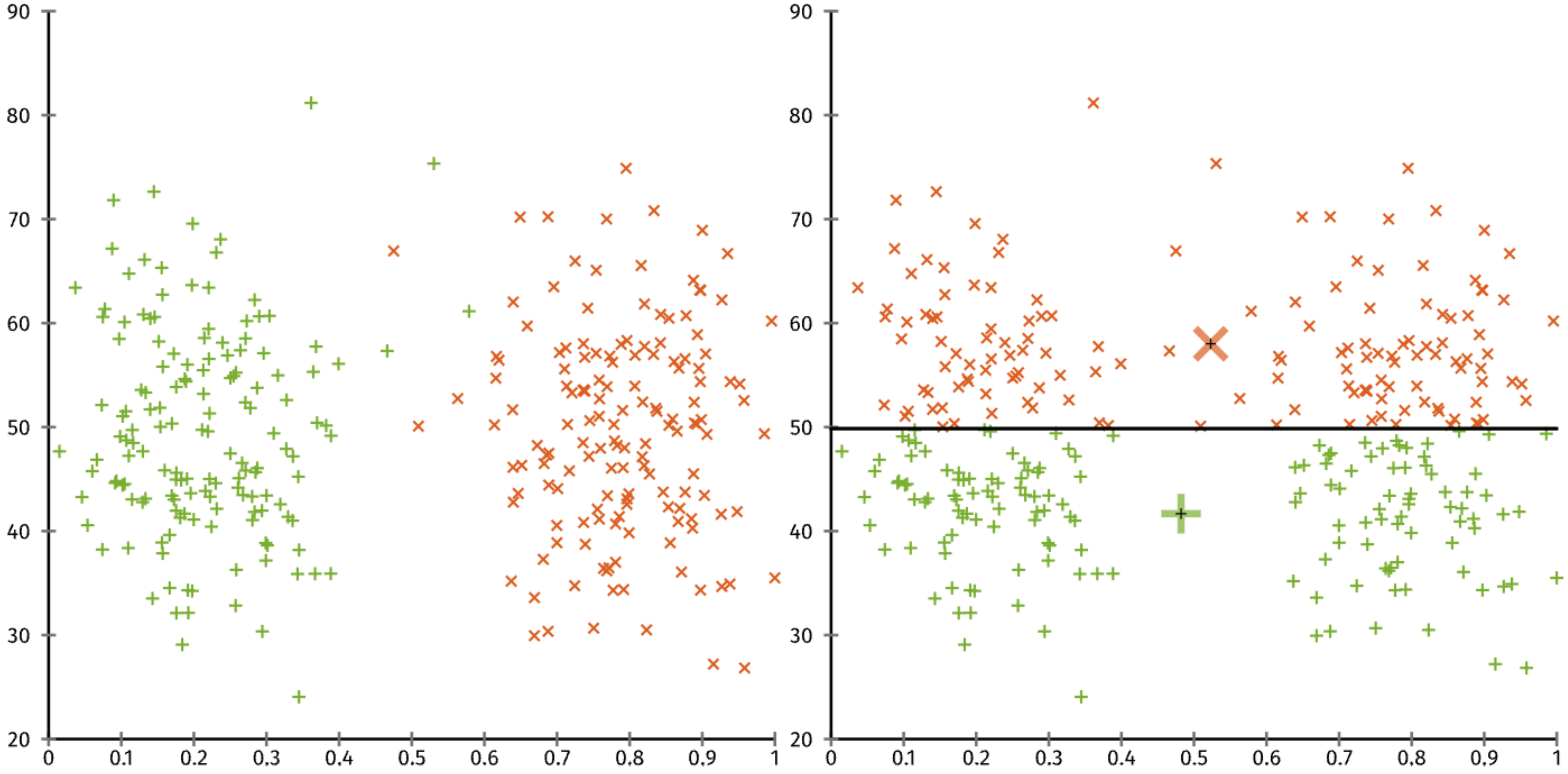
Limitations of K-means

Example – different diameter:



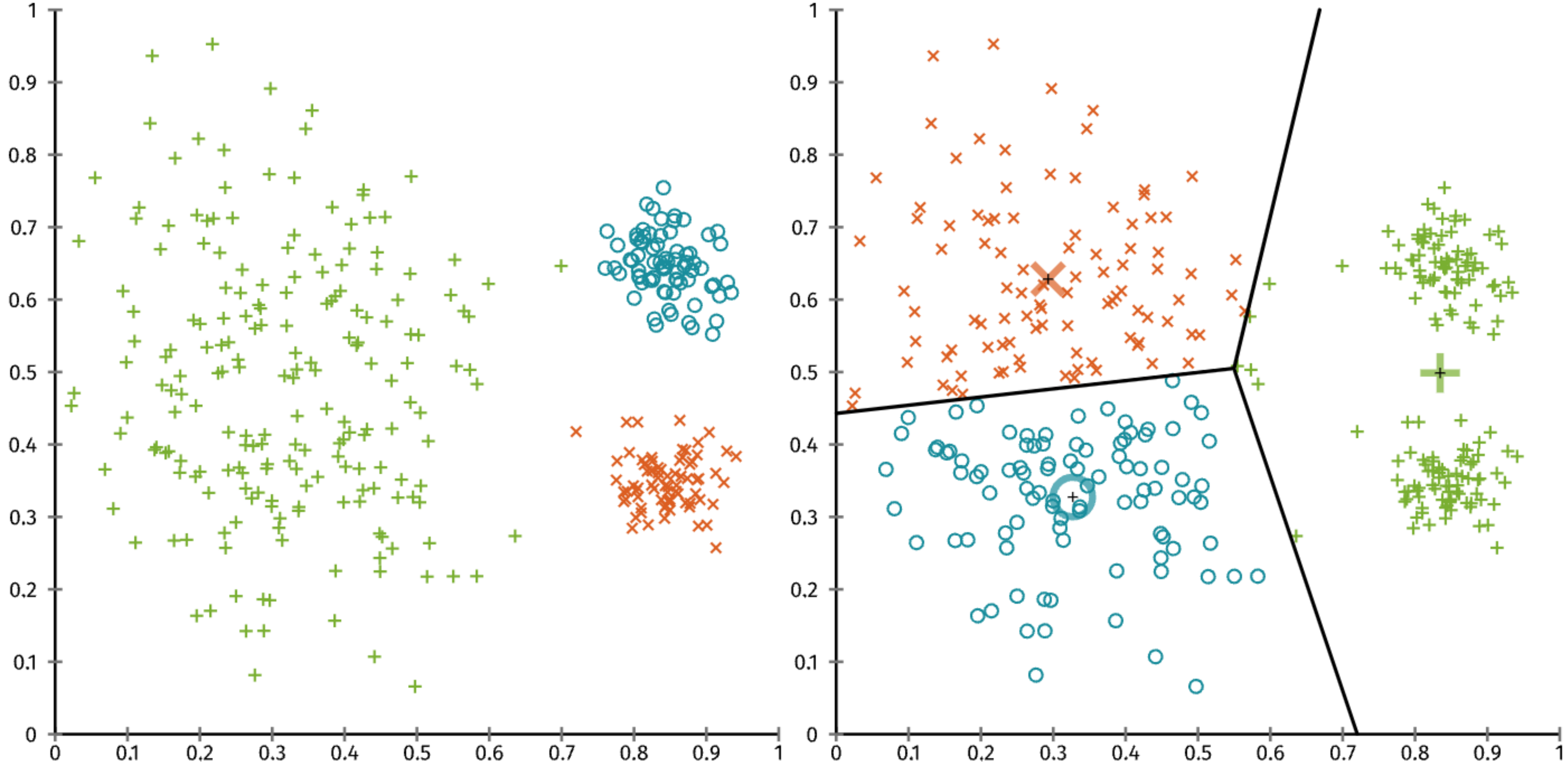
Limitations of K-means

Example – scaling:



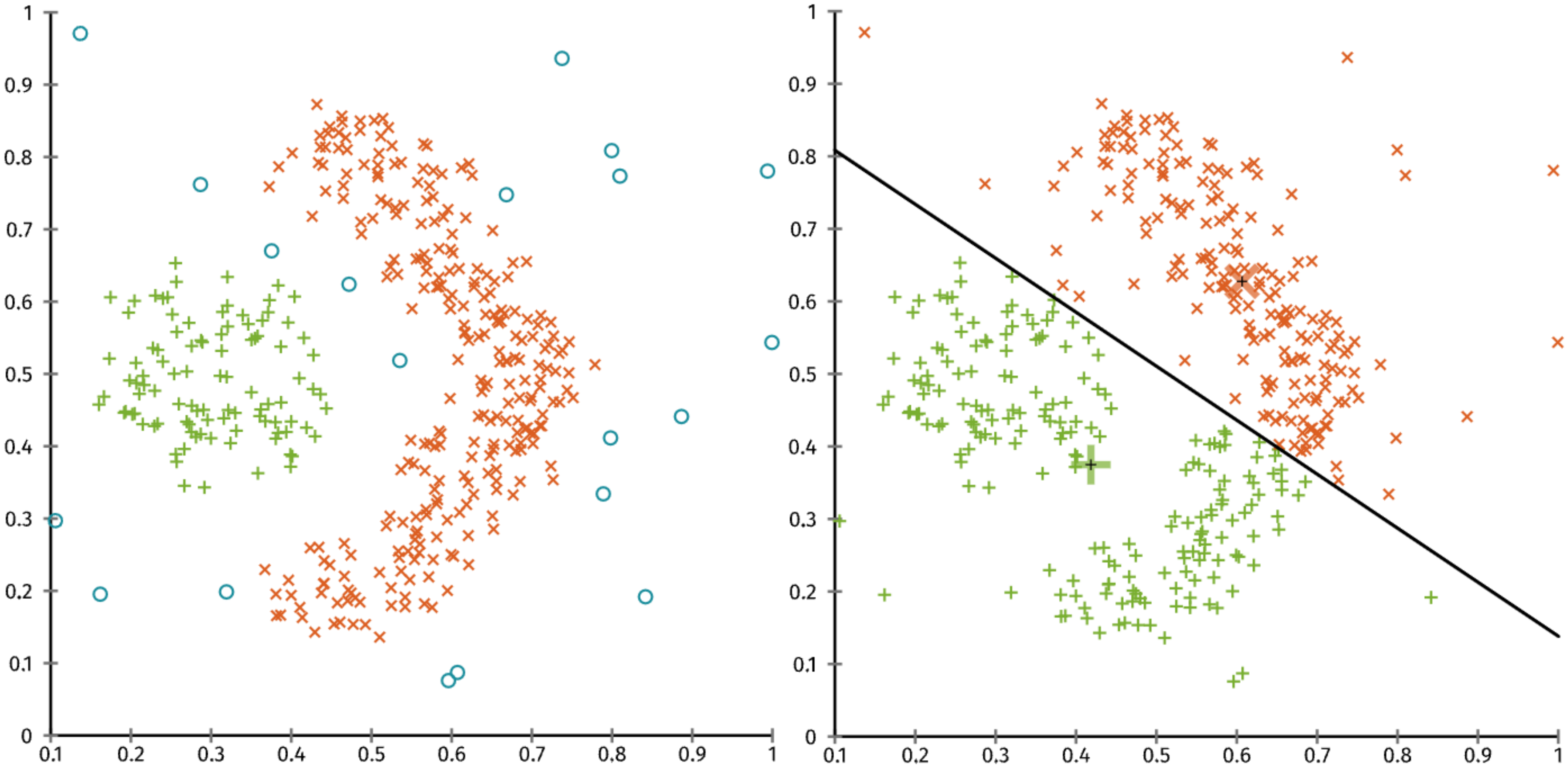
Limitations of K-means

Example – different densities:



Limitations of K-means

Example – cluster shapes:

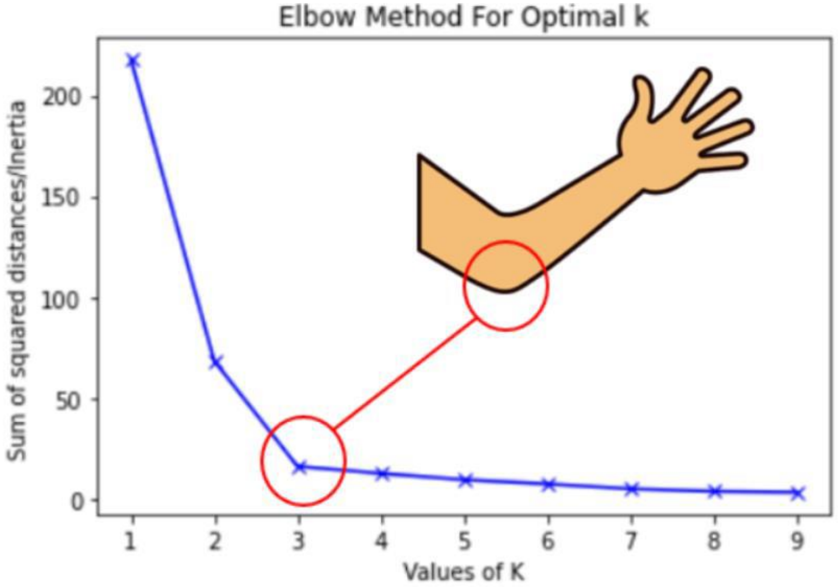


Selecting the optimal number of Clusters (K)

Elbow curve method:

The elbow method runs k-means clustering on the dataset for a range of values of k (say 1 to 10).

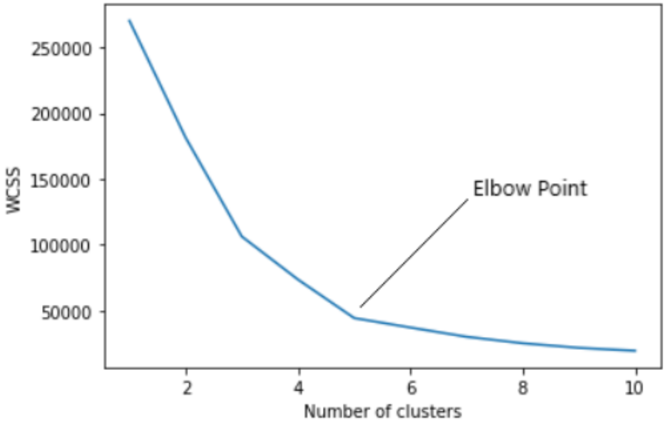
- Perform K-means clustering with all these different values of K. For each of the K values, we calculate average distances to the centroid across all data points.
- Plot these points and find the point where the average distance from the centroid falls suddenly (“Elbow”).



“The curve looks like an elbow. In the above plot, the elbow is at k=3 (i.e. Sum of squared distances falls suddenly) indicating the optimal k for this dataset is 3.”

Selecting the optimal number of Clusters (K)

- “For each value of K, we are calculating WCSS (Within-Cluster Sum of Square).
- WCSS is the sum of squared distance between each point and the centroid in a cluster.
- When we plot the WCSS with the K value, the plot looks like an Elbow.
- As the number of clusters increases, the WCSS value will start to decrease.
- WCSS value is largest when K = 1.
- When we analyze the graph we can see that the graph will rapidly change at a point and thus creating an elbow shape.
- From this point, the graph starts to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an



Source: Saji, B. “In-depth Intuition of K-Means Clustering Algorithm in Machine Learning”. https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

Thank you!

Contact

mousas27@mcmaster.ca

Book an appointment with DASH: https://library.mcmaster.ca/services/dash

Contact DASH: Data Analysis Support Hub: libdash@mcmaster.ca