Machine Learning with R: K-means Clustering

Shaila Jamal

Data Analysis Support Assistant, DASH, McMaster Library

Ph.D. Candidate, School of Earth, Environment and Society, McMaster University

October 03, 2022

McMaster University | DASH

McMaster University | Library

McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the "Dish With One Spoon" wampum agreement.

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.

Code of Conduct

The DASH program 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/

Supervised vs Unsupervised Learning

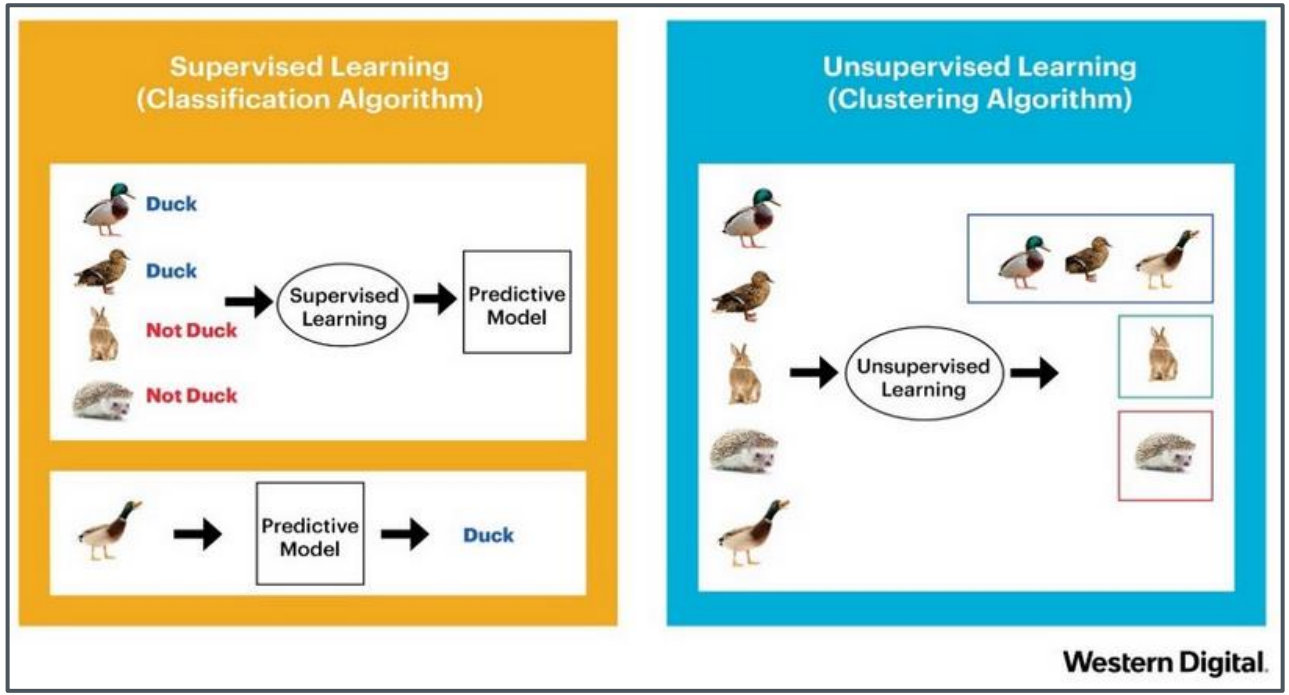
"The main difference is that supervised learning uses **labeled data** to help predict outcomes, while unsupervised learning does not."

- "**Supervised learning** is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to **train or "supervise" algorithms** into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time."

- **"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")."

Source: Delua, J. "Supervised vs. Unsupervised Learning: What's the Difference?" March 12, 2021, accessed on September 29, 2022. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

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

Unsupervised Learning: Clustering

An unsupervised algorithm will not make predictions since it does not have a target/output variable.

Item	Calories	Fat (g)	Carb. (g)	Fiber (g)	Protein	Sodium
Cool Lime Starbucks Refreshersâ„¢ Beverage	45	0	11	0	0	10
Strawberry Acai Starbucks Refreshersâ„¢ Beverage	80	0	18	1	0	10
Very Berry Hibiscus Starbucks Refreshersâ„¢ Beverage	60	0	14	1	0	10
Evolution Freshâ„¢ Organic Ginger Limeade	110	0	28	0	0	5
Iced Coffee	0	0	0	0	0	0
Iced Espresso Classics - Vanilla Latte	130	2.5	21	0	5	65
Iced Espresso Classics - Caffe Mocha	140	2.5	23	0	5	90
Iced Espresso Classics - Caramel Macchiato	130	2.5	21	0	5	65
Shaken Sweet Tea	80	0	19	0	0	10
TazoÂ® Bottled Berry Blossom White	60	0	15	0	0	10
TazoÂ® Bottled Black Mango	150	0	38	0	0	15
TazoÂ® Bottled Black with Lemon	140	0	35	0	0	10
TazoÂ® Bottled White Cranberry	140	0	35	0	0	10
TeavanaÂ® Shaken Iced Black Tea	30	0	8	0	0	5
TeavanaÂ® Shaken Iced Black Tea Lemonade	70	0	17	0	0	0

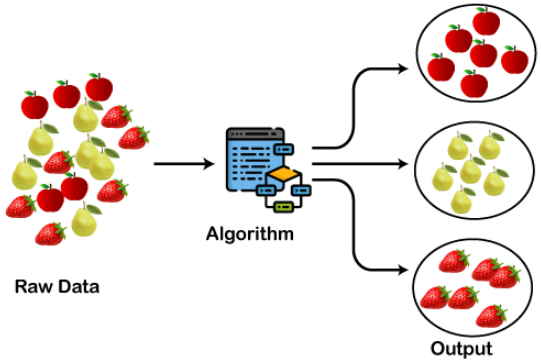


Image Source: https://static.javatpoint.com/tutorial/machine-learning/images/clustering-in-machine-learning.png

Data Source: https://www.kaggle.com/datasets/starbucks/starbucks-menu

K-means Clustering

Most popular clustering method

“The algorithm tries to find groups by minimizing the distance between the observations”

“K-Means finds the best centroids by alternating between (1) assigning data points to clusters based on the current centroids (2) choosing centroids (points which are the center of a cluster) based on the current assignment of data points to clusters.

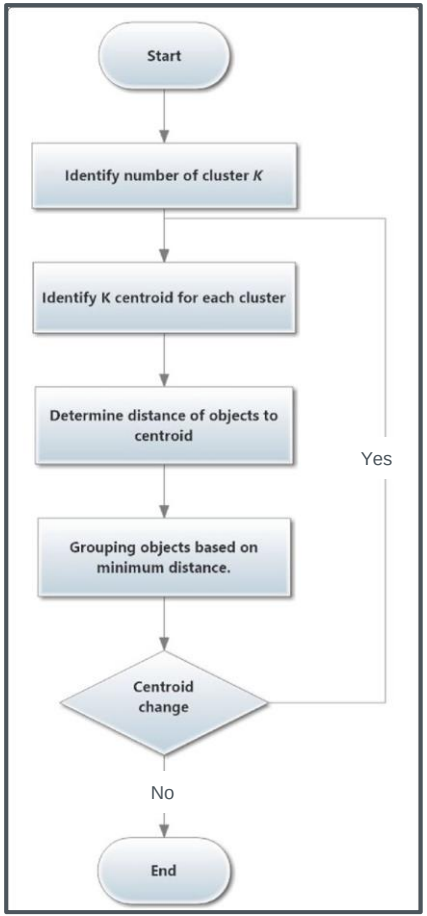
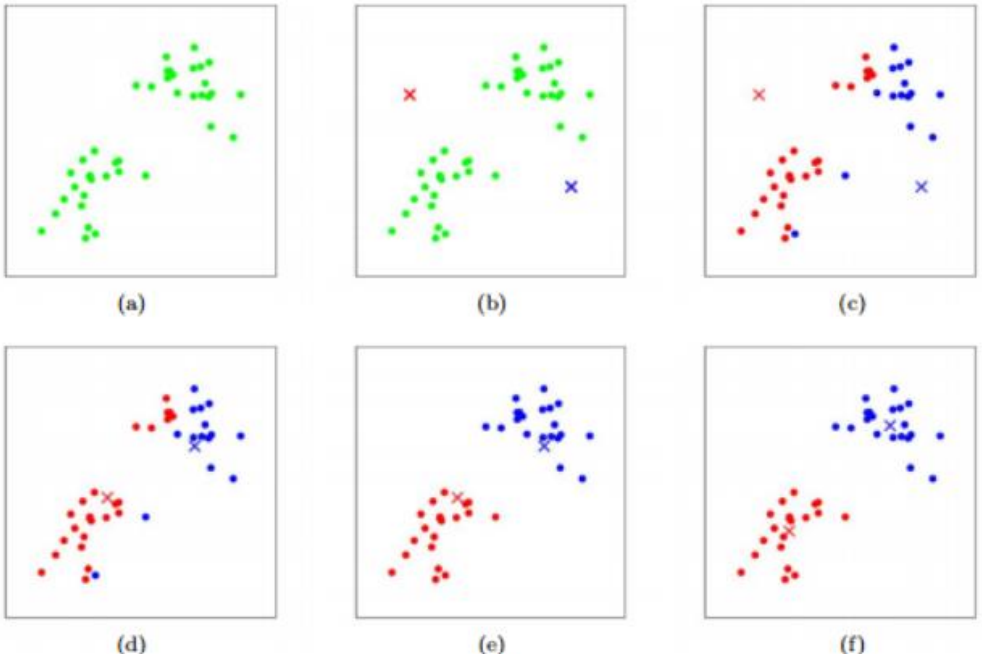


Image Source: https://www.researchgate.net/publication/271915066_Content-based_image_retrieval_using_PSO_and_k-means_clustering_algorithm/figures?lo=1

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

K-means Clustering



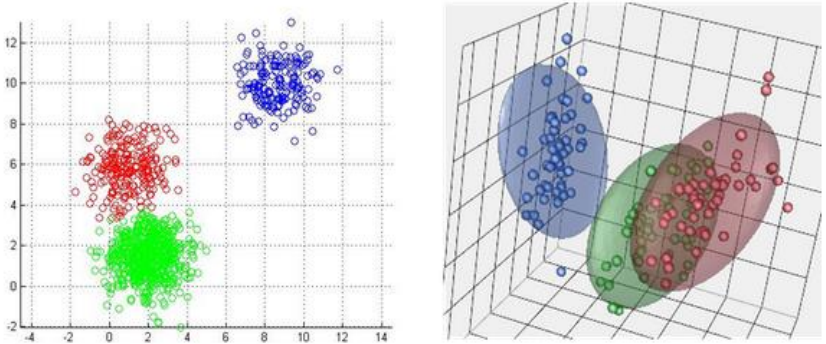
“K-means algorithm: Training examples are shown as dots, and cluster centroids are shown as crosses.
(a) Original dataset.
(b) Random initial cluster centroids.
(c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (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.”

“Images courtesy of Michael Jordan”

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

K-means Clustering

Previous figure “shows k-means with a 2-dimensional feature vector (each point has two dimensions, an x and a y). In your applications, will probably be working with data that has a lot of features. In fact each data-point may be hundreds of dimensions. We can visualize clusters in up to 3 dimensions (see figure 3) but beyond that you have to rely on a more mathematical understanding.”



KMeans in other dimensions. (left) K-means in 2d. (right) K-means in 3d. You have to imagine k-means in 4d.

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

K-means Clustering

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: 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/

K-means Clustering

Important Factors:

“4. Distance Measures: Using different distance measures (used to calculate 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/

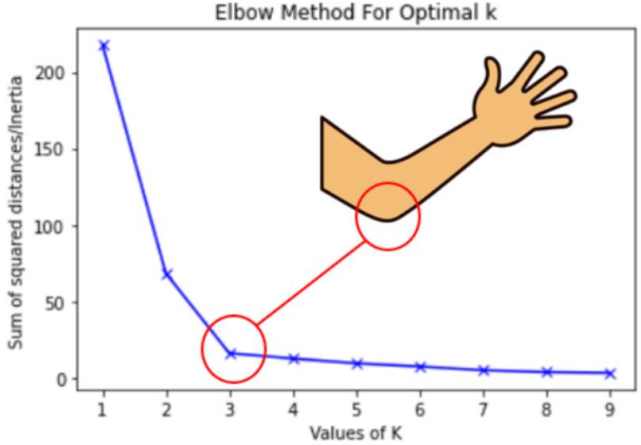
K-means Clustering

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”).”



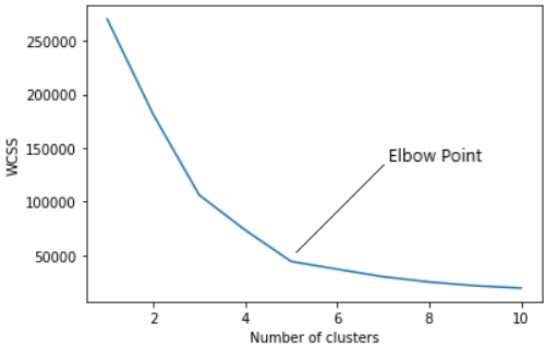
“The curve looks like an elbow. In the above plot, the elbow is at k=3 (i.e. Sum of squared distances falls suddenly) indicating the optimal k for this dataset is 3.”

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/

K-means Clustering

Elbow curve method:

- “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 optimal number of clusters.”



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!