Unsupervised

Machine Learning

Disclaimer: The content is curated for educational purposes only.

Agenda

- Unsupervised learning
- How to identify problem at hand is of Unsupervised type?
- Types of Unsupervised learning
- Clustering and Types of clustering
- Unsupervised ML algorithm: K-mean clustering
- Evaluation technique for k-mean clustering

Unsupervised Learning

Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision.[1]

How to identify problem at hand is of Unsupervised type?

❑ When the data is without labels , tags or classes

❑ When data is categorized based on their similarities and dissimilarities

Types of Unsupervised learning

❑ Unsupervised Transformations

❑ Clustering Algorithms[2]

Clustering

- Partitioning data into different groups called as clusters based on their similarities.
- The data points in the same group are similar in nature and different from the data points in other groups.[2]

Types of Clustering

❑ k-means clustering

❑ Agglomerative clustering

❑ DBSCAN

k-means Clustering

- ❑ Find cluster Centers that are representative of certain regions of the data
- ❑ It iterative works on two steps -
  1. assigning each data point to the closest cluster Center
  2. setting each cluster Center as the mean of the data points that are assigned to it
- ❑ Finished when the assignment of instances to clusters no longer changes[2]

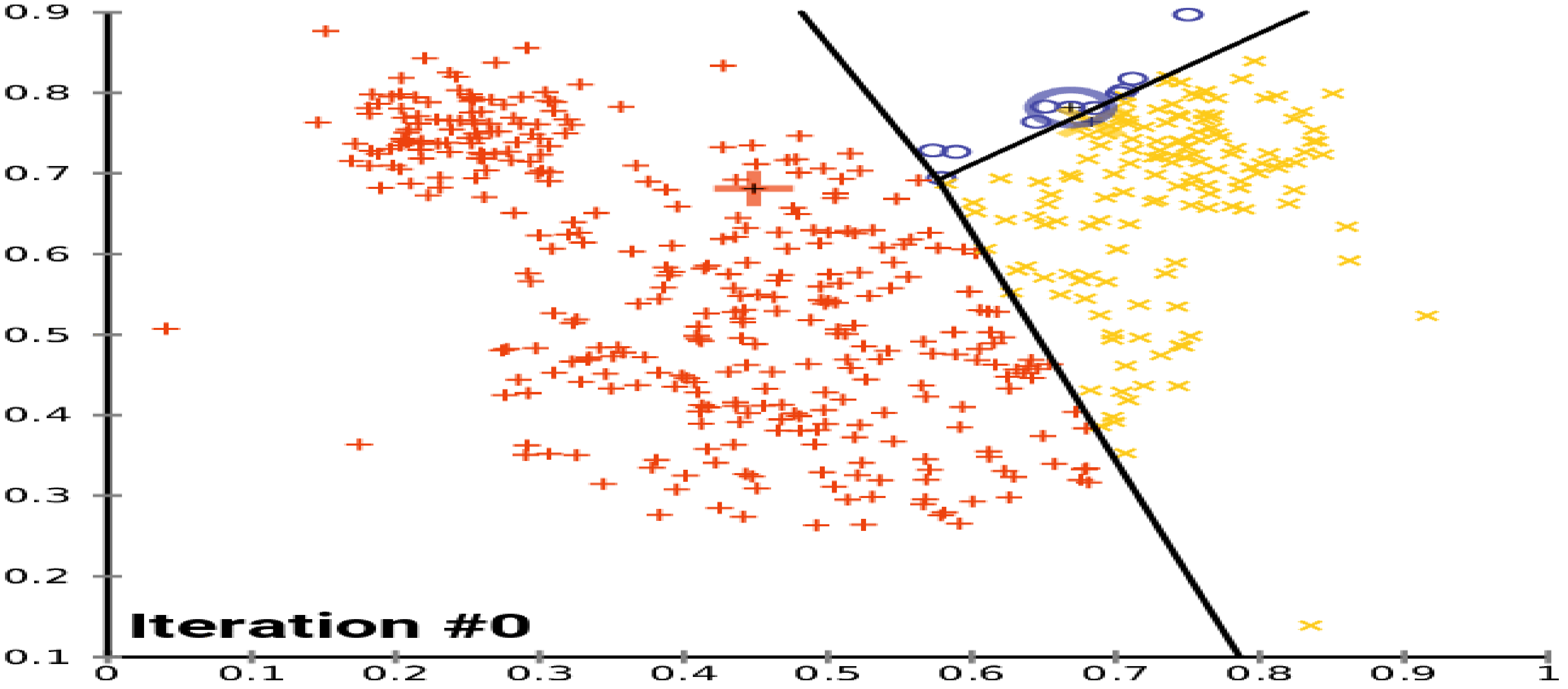
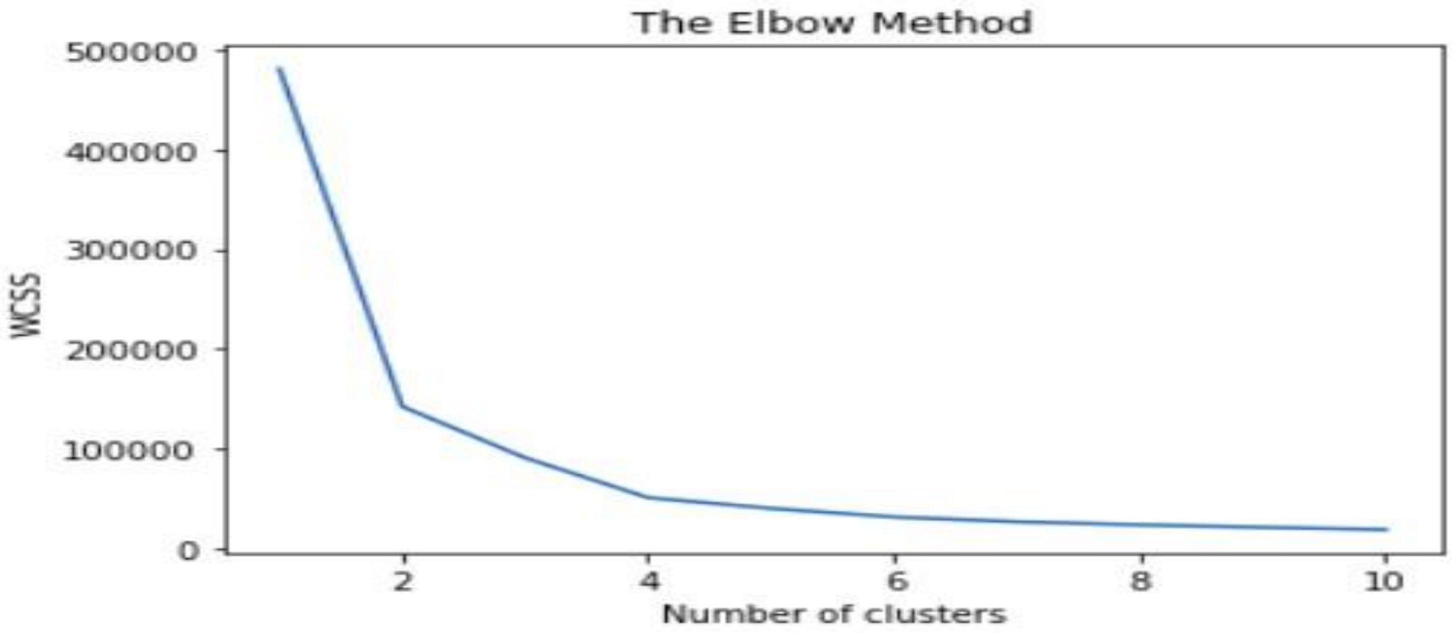


Image Source - https://commons.wikimedia.org/wiki/File:K-means_convergence.gif[3]

How to decide value of k?



Evaluation technique for k-mean clustering[4]

❑ **Adjusted Rand Index**

computes a similarity measure between two clustering

considers all pairs of samples that are assigned in the similar or different clusters in the predicted and true clustering the raw Rand Index score is ‘adjusted for chance’ into the Adjusted Rand Index score

***Adjusted_RI* = (*RI*−*Expected_RI*) / (*max*(*RI*)−*Expected_RI*)**

❑ **Mutual Information based score**

Normalized Mutual Information (NMI)

Adjusted Mutual Information (AMI)

❑ Silhouette Coefficient

The Silhouette function will compute the mean Silhouette Coefficient of all samples using the mean intra-cluster distance and the mean nearest-cluster distance for each sample.

$$S=(b-a)/max(a,b)$$

Here, a is intra-cluster distance and, b is mean nearest-cluster distance.

REFERENCES

1. https://en.wikipedia.org/wiki/Unsupervised_learning
2. Andreas C. Müller and Sarah Guido, *Introduction to Machine Learning with Python A GUIDE FOR DATA SCIENTISTS,* 1st ed., CA:O'Reilly Media, Inc.,2016. [E-book] Available: Safari ebook.
3. https://commons.wikimedia.org/wiki/File:K-means_convergence.gif
4. https://www.tutorialspoint.com/scikit_learn/scikit_learn_clustering_performance_evaluation.htm

THANK YOU