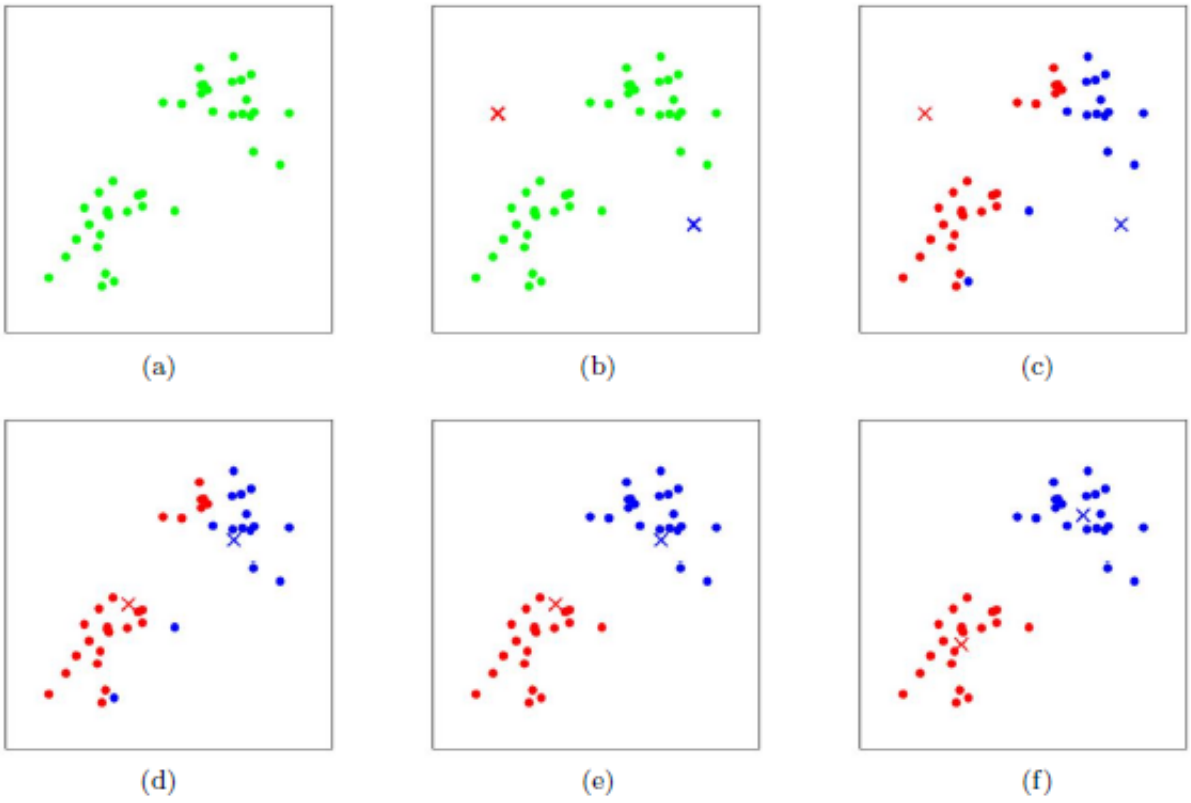
Unsupervised Learning: K-Means Clustering

Introduction

Unsupervised learning is a major method of machine learning which can detect patterns from data without given labels. It can help find the detailed commonality of data and use this to analysis new data. Unsupervised learning contains two main methods, clustering analysis and dimensionality reduction.[1] [2]

This report will focus on K-Means clustering, a centroid based algorithm used in clustering analysis. K-Means clustering is an automatic classification method that groups observations into k clusters, in which items has nearest mean. First, it randomly select k points as the centroid of k clusters. Then calculate the distance from each point to the center of the cluster, and select the closest point to re-center the cluster, and iterate. The goal of K-Means clustering is to minimize the difference in each group and maximize the difference among groups.



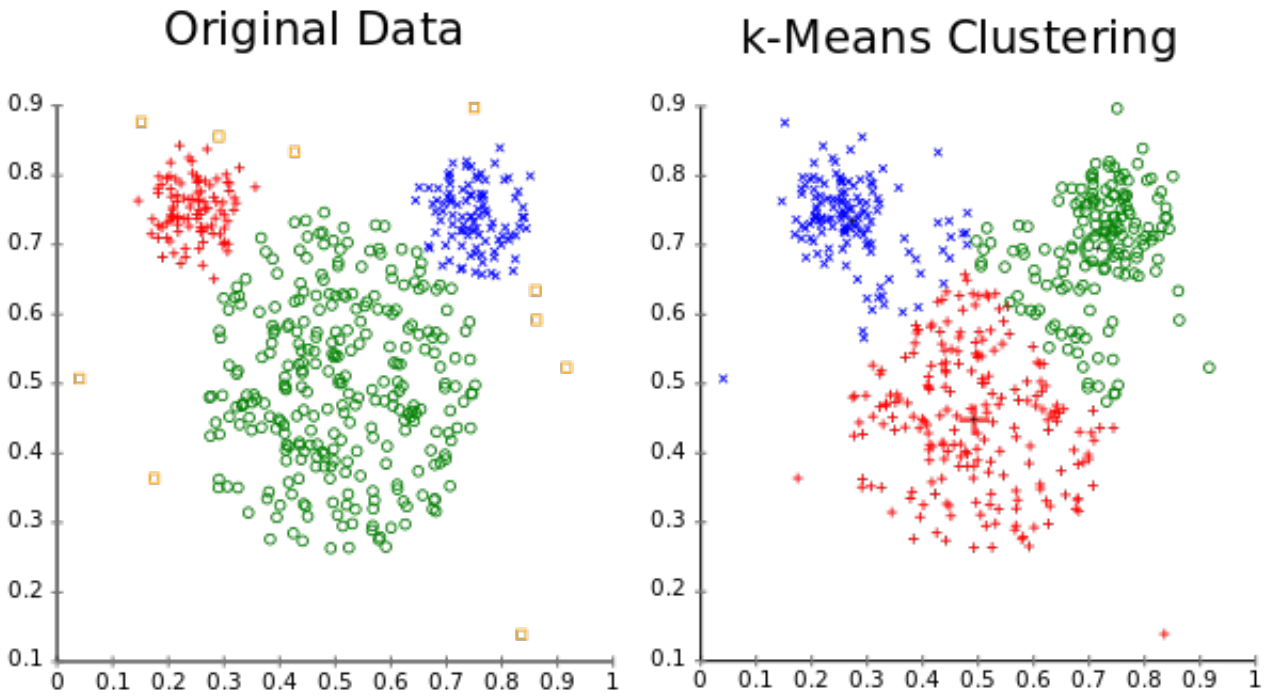
Typically, the K-Means algorithm is applied to crime analysis[3], user behavior classification[4] and traffic information analysis with many digitized data[5].

Analysis

- An inappropriate choice of cluster number k may yield poor results.
- Convergence to local minimum may happen, which can lead to counterintuitive result.[6]
- Easy to perform. The user can easily build up environment for K-Means clustering by:

    - Python[7]
    - Ruby[8]
    - JavaScript[9]
    - C[10]
    - ...
- K-Means is based on the concept that each cluster has similar size. This may limit the performance of K-Means when dealing with data sets with different size of clusters. (Shown by the following figure.)

e.g. Bad performance on a "mouse" dataset



Some Tips

- When performing K-Means clustering, try to find a suitable k value, instead of picking an arbitrary one. Remember to run diagnostic checks to get the proper value of k in the data set. Or work when the k value is known.
- Use K-Means to deal with dense and spherical datasets.
- Use K-Means to deal with datasets with similar size of clusters.

Conclusion

K-Means is a mainstream algorithm widely used in unsupervised learning, which makes it easier and more efficient to process spherical datasets without preset labels.

---

1. A. Kuh, M. S. Uddin and P. Ng, "Online unsupervised kernel learning algorithms," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 1019-1025. ↩

2. L. Wang, "Research on Distributed Parallel Dimensionality Reduction Algorithm Based on PCA

Algorithm," 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 2019, pp. 1363-1367. ↩

3. A. Joshi, A. S. Sabitha and T. Choudhury, "Crime Analysis Using K-Means Clustering," 2017 3rd International Conference on Computational Intelligence and Networks (CINE), Odisha, 2017, pp. 33-39. ↩

4. L. Xue and W. Luan, "Improved K-means Algorithm in User Behavior Analysis," 2015 Ninth International Conference on Frontier of Computer Science and Technology, Dalian, 2015, pp. 339-342. ↩

5. M. A. Mondal and Z. Rehena, "Identifying Traffic Congestion Pattern using K-means Clustering Technique," 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), Ghaziabad, India, 2019, pp. 1-5. ↩

6. H. Chu and C. Wang, "Using K-means algorithm for the road junction time period analysis," 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), Taichung, 2017, pp. 441-446. ↩

7. https://github.com/lars76/kmeans-anchor-boxes ↩

8. https://github.com/reddavis/K-Means ↩

9. https://github.com/harthur/clustering ↩

10. https://github.com/serban/kmeans ↩