K-means Clustering

Based on module COMP5122M Data Science
Lectured by Duygu Sarikaya

1 Introduction

Clustering is an unsupervised learning algorithm which separates data points into several clusters. As an unsupervised model, the data is not labelled, only considering features. Therefore, the model is trained to cluster the data points with similar features. In other words, the data points in the same cluster are similar, while the data points in different clusters are much different. Clustering is useful when labels are unknown, aiming to find meaningful insights from the data. There are several types of clustering, such as centroid-based, density-based, distribution-based, and hierarchical clustering, depending on the methods used to group the data points.

This report will focus on the most popular clustering algorithm, k-means clustering, and will include further experiments on this algorithm in Python.

2 K-means Clustering

K-means clustering is one of the centroid-based clustering algorithms where clusters are determined by assigning data points to the closest centroids. Here, the centroids represent the centres of each cluster. K-means clustering is the most commonly used algorithm, as it is simple and easy to compute.

The k-means clustering algorithm is as follows: first, k centroids are randomly selected, where k denotes the number of clusters specified manually. Next, each data point is assigned to the closest centroid by calculating its squared distance from the centroids. The centroid for each cluster is then recomputed by taking the average of all data points in each cluster. As the centroids are updated, the data points are reassigned to the closest centroids again. This process iterates until the centroids no longer change, indicating that the algorithm is converged. This way, the data points are assigned to clusters, minimising the variance.

K-means clustering has several advantages. One of its most significant advantages is its simple algorithm, which makes it easy and fast to compute for any dataset, even large datasets. Furthermore, as the k-means clustering algorithm continues until there is no change in the centroids, it guarantees convergence. Additionally, adapting unseen data points to the model is straightforward. Despite the advantages, there are some disadvantages that should be considered. Firstly, the k-means clustering algorithm requires the number of clusters k to be specified manually. In most real-world applications, however, k is unknown, making it challenging to find the appropriate value for k. Moreover, the algorithm starts with selecting random k centroids. As the algorithm iterates, it assigns data points to these centroids, therefore the algorithm might get stuck in local optima, resulting in a bad clustering model. This shows that the k-means algorithm is sensitive to the initial centroids. Lastly, the algorithm makes a restrictive assumption that clusters are spherical and equally sized. Therefore, the algorithm cannot cluster the data points well in cases with complex shape clusters, even though it seems obvious when visualised.

K-means clustering is an unsupervised algorithm that cannot calculate accuracy like regression or classification models. Instead, it is evaluated using loss functions. For instance, models with the same number of clusters but different initial centroids might have different clustering results. In such scenarios, the loss functions help to evaluate the best model. There are two commonly used loss functions: inertia and distortion. Inertia measures the sum of squared distances from each data point to its centroid, while distortion measures the average sum of squared distances from each data point to its centroid.

Another way to evaluate the k-means clustering model is by finding the optimal number for k, the number of clusters. As mentioned previously, in most cases, the number of clusters is unknown. Therefore, the elbow method can be used to find the optimal k. The elbow method involves plotting inertia values for each k and finding the point where the graph flattens. This point indicates the optimal k for the model. Here, the inertia represents how well the data points are clustered.

3 Experiments

Some experiments are conducted to evaluate the performance of the k-means clustering algorithm and related techniques with two different datasets. The first dataset is labelled, which allows assessing how well the k-means clustering algorithm works, using labels only for evaluation purposes. The wine dataset, consisting of 178 instances with 13 numerical features from the results of a chemical analysis of each wine, is used. Since the number of clusters is unknown, the elbow method is used by plotting a graph of inertia for each value k from 1 to 10 to find the optimal k. The graph in graph 1 clearly shows that the inertia starts to flatten at $k = 3$. Therefore, the model is trained using the k-means clustering algorithm with $k = 3$. To visualise the clustering model in 2 dimensions, PCA is applied to plot a scatter graph. The graph in 2a shows that the data points are well separated into three clusters. To compare the labels clustered by the trained model with the real labels of the data points, another scatter graph is plotted in graph 2b. As clearly shown, the labels clustered by the model and the real labels are relatively similar, indicating a well-trained model.

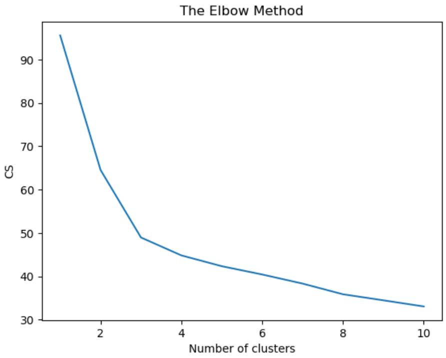
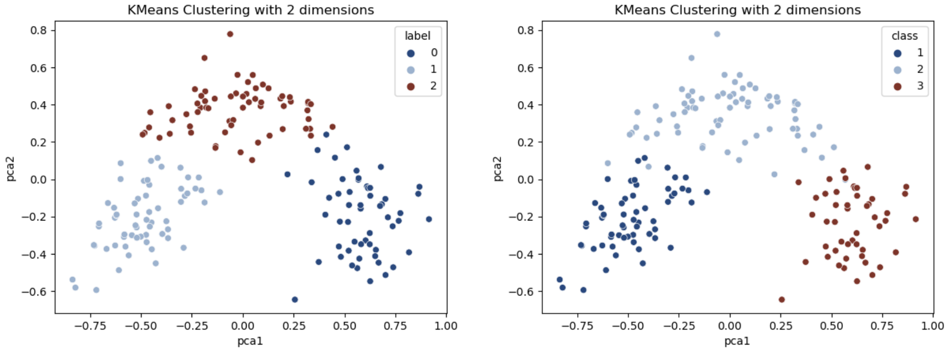


Figure 1: The Elbow Method with the Wine Dataset



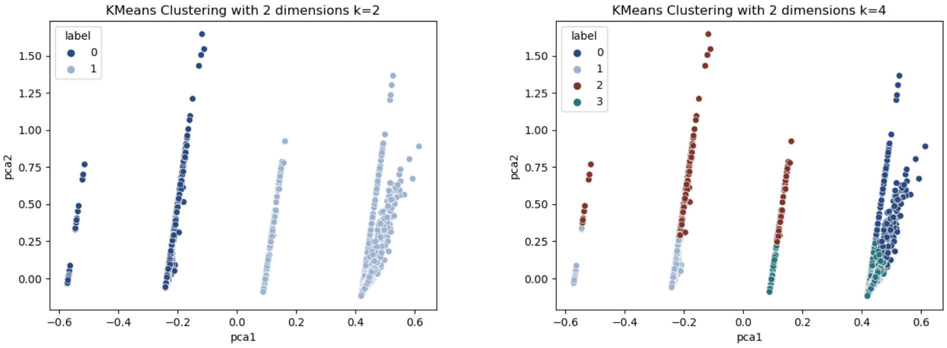
(a) Clusters of the Trained Model

(b) Clusters of Real Labels

Figure 2: K-means Clustering with the Wine Dataset

However, the k-means clustering algorithm assumes that clusters are spherical and equally sized. To evaluate the case with non-spherical clusters, a model is trained using the Facebook Live Sellers dataset, which consists of 11 features of engagement metrics and 7,050 instances. The exact process is applied again. As the elbow method indicates that the optimal k might be 2 by plotting a scatter graph of inertia, the k-means clustering algorithm is applied with $k = 2$. However, graph 3a clearly shows four different non-spherical clusters. Training the model with $k = 4$ shows that the k-means clustering does not work well, as shown in graph 3b. This indicates that the k-means clustering does not work with clusters of complex shapes, even though it seems obvious when visualised.

In summary, training two different models shows that the k-means clustering algorithm only works for spherical and equally sized clusters, not complex shapes.



(a) k=2

(b) k=4

Figure 3: K-means Clustering with the Facebook Live Sellers Dataset

4 Conclusion

To conclude, clustering is an algorithm that groups similar data points into clusters. The most commonly used clustering algorithm is k-means clustering due to its simplicity and speed. However, it requires k to be assigned manually before training a model. One of the ways to find the optimal k is using the elbow method by calculating the inertia of each model with different values of k. Also, it is sensitive to initial clusters and assumes that the clusters are spherical and equally sized. By training different models with two datasets, the k-clustering algorithm does not work when clusters are not spherical and equally sized.

References

[1] Guido, S., Muller, A. 2016. *Unsuprvied Learning and Preprocessing.* In: Guido, S., Muller, A. *Introduction to Machine Learning with Python.* Sebastopol, CA: O'Reilly, pp.131-209.

[2] Maimon, O., Rokach, L. 2005. *Clustering Methods.* In: Maimon, O., Rokach, L. (eds) *The Data Mining and Knowledge Discovery Handbook.* Boston, MA: Springer, pp.321-352

[3] Aeberhand, S., Forina M. 1991. *Wine.* UCI Machine Learning Repository. [Online]. [Accessed 20 February 2024]. Available from: `https://archive.ics.uci.edu/dataset/109/wine`

[4] Nassim, D. 2019. *Facebook Live Sellers in Thailand.* UCI Machine Learning Repository. [Online]. [Accessed 20 February 2024].

Available from: `https://archive.ics.uci.edu/dataset/488/facebook+live+sellers+in+thailand`