K-Mean: Getting The Optimal Number Of Clusters

50+ Exciting Industry Projects to become a Full-Stack Data Scientist 👨‍💻 Download Projects ×

Ankita Banerji — Published On May 18, 2021 and Last Modified On July 19th, 2022

Beginner Clustering Machine Learning Python Unsupervised

This article was published as a part of the Data Science Blogathon.

Introduction

K-means clustering is an unsupervised algorithm. In an unsupervised algorithm, we are not interested in making predictions (since we don't have a target/output variable). The objective is to discover interesting patterns in the data, e.g., are there any subgroups or 'clusters' among the bank's customers?

Clustering techniques use raw data to form clusters based on common factors among various data points. Customer segmentation for targeted marketing is one of the most vital applications of the clustering algorithm.

Image Source: https://www.marketing91.com/4-types-market-segmentation-segment/

Practical Application of Clustering

Customer Insight:

Let a retail chain with so many stores across locations wants to manage stores at best and increase the sales and performance. Cluster analysis can help the retail chain to get desired insights on customer demographics, purchase behaviour and demand patterns across locations.

This will help the retail chain with assortment planning, planning promotional activities and store benchmarking for better performance and higher returns.

Marketing:

We use cookies on Analytics Vidhya websites to deliver our services, analyze web traffic, and improve your experience on the site. By using Analytics Vidhya, you agree to our Privacy Policy and Terms of Use. Accept

customised marketing campaign for each of the group. You do not have any label in mind, such as good customer or bad customer. You want to just look at patterns in customer data and then try and find segments. This is where clustering techniques can help.

Social Media:

In the areas of social networking and social media, Cluster Analysis is used to identify similar communities within larger groups.

Medical:

Cluster Analysis has also been widely used in the field of biology and medical science like sequencing into gene families, human genetic clustering, building groups of genes, and clustering of organisms at species and so on.

Important Factors We Must consider While Using K-means Algorithm

Certain factors can impact the efficacy of the final clusters formed when using k-means clustering. So, we must keep in mind the following factors when solving business problems using the K-means clustering algorithm.

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.

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.

In this blog, we will discuss the most important parameter i.e., ***the ways by which we can select an optimal number of clusters (K)***. There are several methods to find the best value of K. We will discuss them one by one.

1. Elbow Curve Method

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

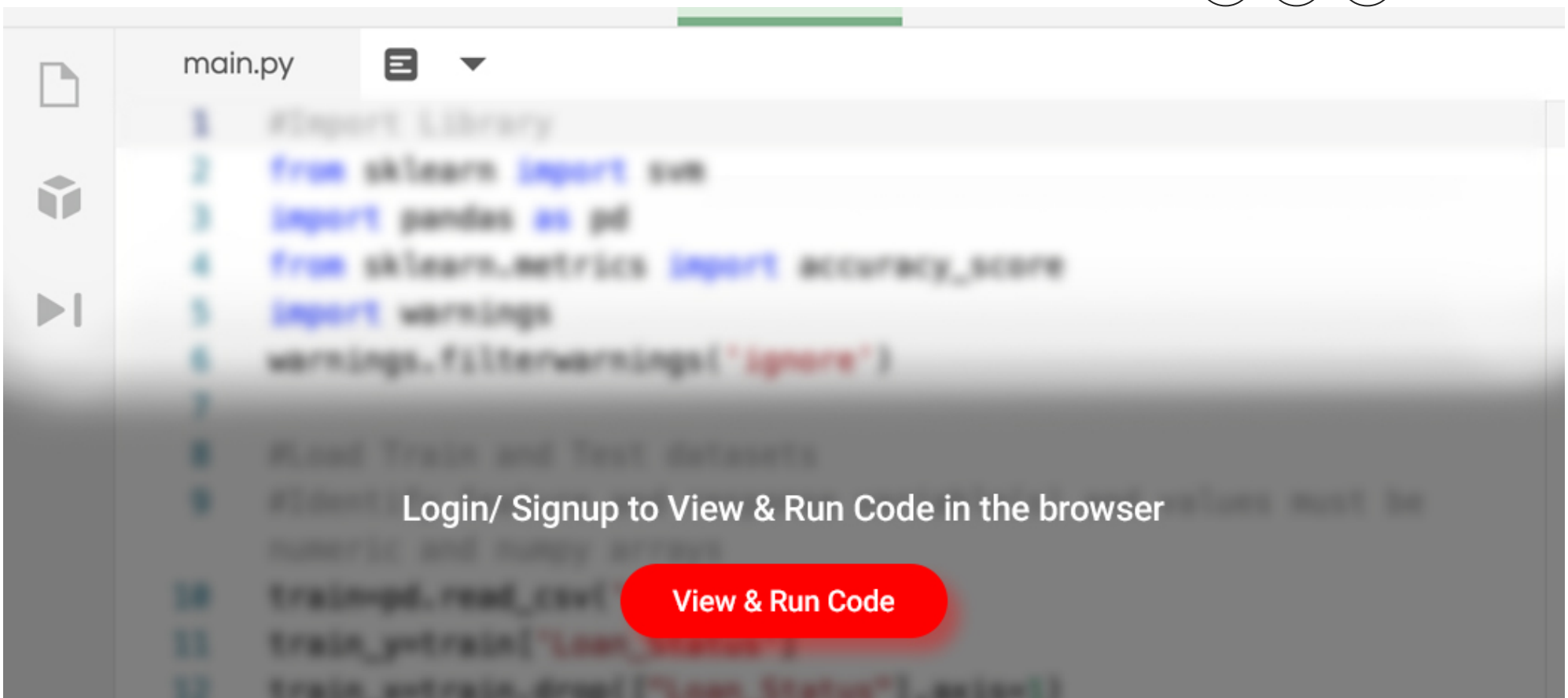
- Perform K-means clustering with all these different values of K. For each of the K values, we calculate average distances to the centroid across all data points.
- Plot these points and find the point where the average distance from the centroid falls suddenly (“Elbow”).

Let us see the python code with the help of an example.

Python Code:

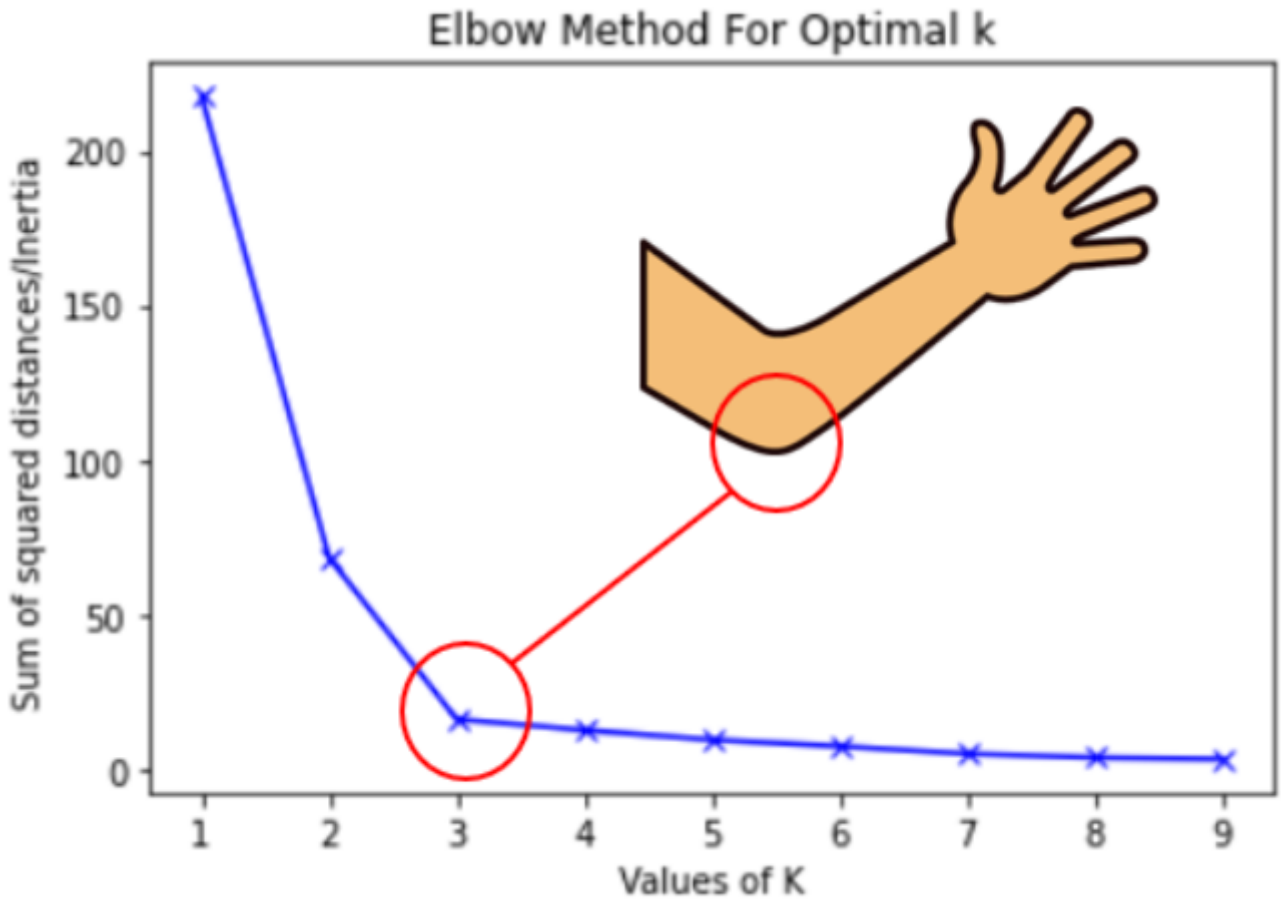
We use cookies on Analytics Vidhya websites to deliver our services, analyze web traffic, and improve your experience on the site. By using Analytics Vidhya, you agree to our Privacy Policy and Terms of Use. Accept

K-Mean: Getting The Optimal Number Of Clusters



Visually we can see that the optimal number of clusters should be around 3. But ***visualizing the data alone cannot always give the right answer.***

```
Sum_of_squared_distances = []
K = range(1,10)
for num_clusters in K :
 kmeans = KMeans(n_clusters=num_clusters)
 kmeans.fit(data_frame)
 Sum_of_squared_distances.append(kmeans.inertia_)
plt.plot(K,Sum_of_squared_distances,'bx-')
plt.xlabel('Values of K')
plt.ylabel('Sum of squared distances/Inertia')
plt.title('Elbow Method For Optimal k')
plt.show()
```



Line plot between K and inertia

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

We use cookies on Analytics Vidhya websites to deliver our services, analyze web traffic, and improve your experience on the site. By using Analytics Vidhya, you agree to our Privacy Policy and Terms of Use. Accept

K-Mean: Getting The Optimal Number Of Clusters

The silhouette coefficient is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation).

- Select a range of values of k (say 1 to 10).
- Plot Silhouette coefficient for each value of K.

The equation for calculating the silhouette coefficient for a particular data point:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- S(i) is the silhouette coefficient of the data point i.
- a(i) is the average distance between i and all the other data points in the cluster to which i belongs.
- b(i) is the average distance from i to all clusters to which i does not belong.

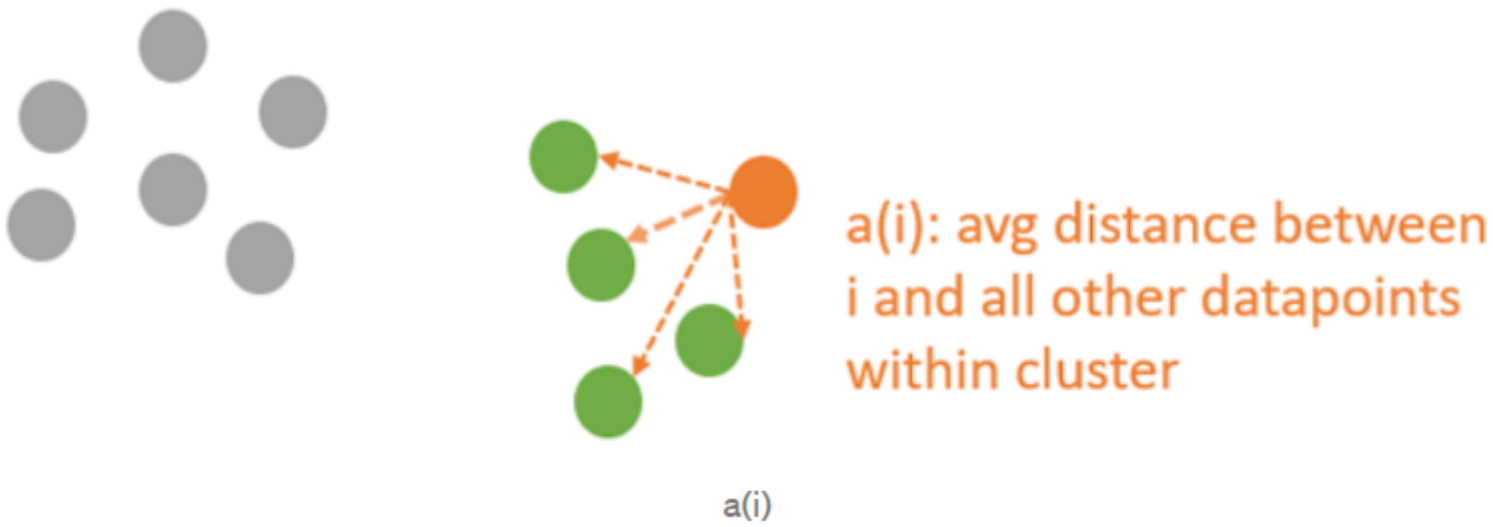


Image Source:https://ankitajhumu.medium.com/selecting-number-of-clusters-in-k-mean-clustering-d60a1f85d65b

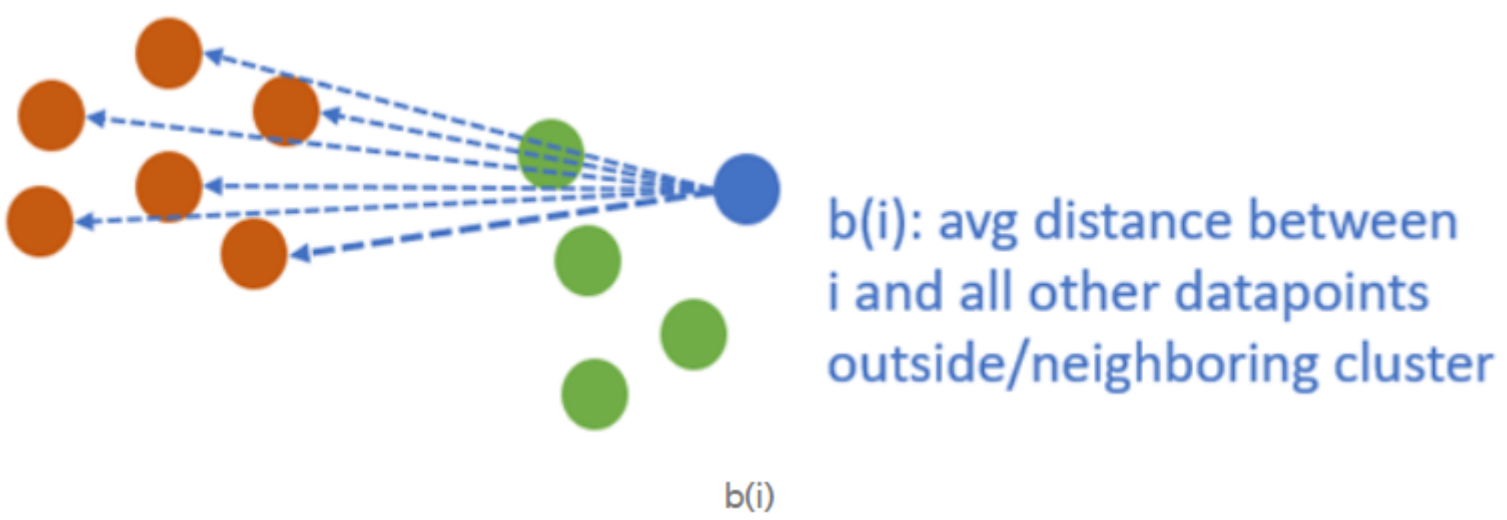


Image Source: https://ankitajhumu.medium.com/selecting-number-of-clusters-in-k-mean-clustering-d60a1f85d65b

We will then calculate the average_silhouette for every k.

$$\text{AverageSilhouette} = \text{mean}\{S(i)\}$$

Then plot the graph between average_silhouette and K.

Points to remember while calculating silhouette coefficient:

- The value of the silhouette coefficient is between [-1, 1].
- A score of 1 denotes the best meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
- The worst value is -1. Values near 0 denote overlapping clusters.

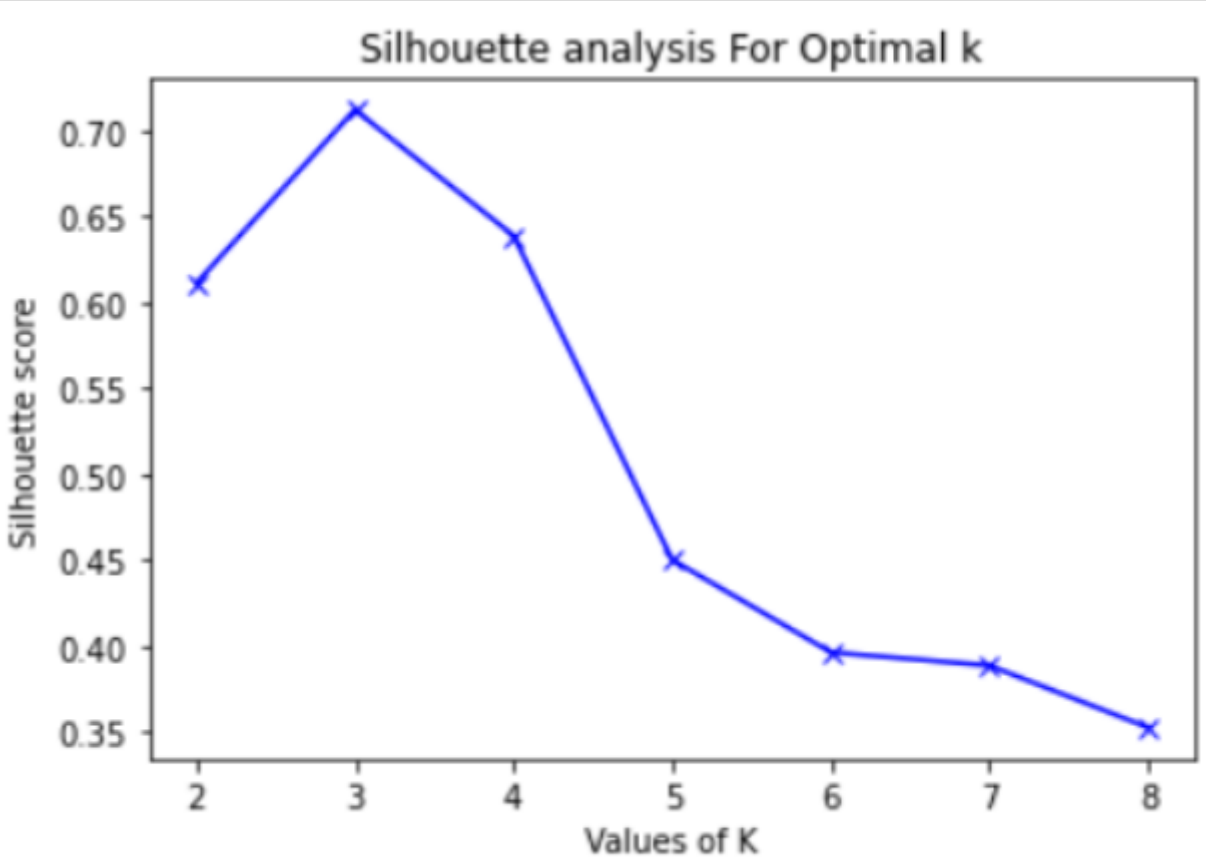
We use cookies on Analytics Vidhya websites to deliver our services, analyze web traffic, and improve your experience on the site. By using Analytics Vidhya, you agree to our Privacy Policy and Terms of Use. Accept

K-Mean: Getting The Optimal Number Of Clusters

```
silhouette_avg = []
for num_clusters in range_n_clusters:

 # initialise kmeans
 kmeans = KMeans(n_clusters=num_clusters)
 kmeans.fit(data_frame)
 cluster_labels = kmeans.labels_

 # silhouette score
 silhouette_avg.append(silhouette_score(data_frame,
cluster_labels))plt.plot(range_n_clusters,silhouette_avg,'bx-')
plt.xlabel('Values of K')
plt.ylabel('Silhouette score')
plt.title('Silhouette analysis For Optimal k')
plt.show()
```



Line plot between K and Silhouette score

We see that the silhouette score is maximized at k = 3. So, we will take 3 clusters.

NOTE: The silhouette Method is used in combination with the Elbow Method for a more confident decision.

In k-means clustering, the number of clusters that you want to divide your data points into i.e., the value of K has to be pre-determined whereas in Hierarchical clustering data is automatically formed into a tree shape form (dendrogram).

So how do we decide which clustering to select? We choose either of them depending on our problem statement and business requirement.

Hierarchical clustering gives you a deep insight into each step of converging different clusters and creates a dendrogram. It helps you to figure out which cluster combination makes more sense.

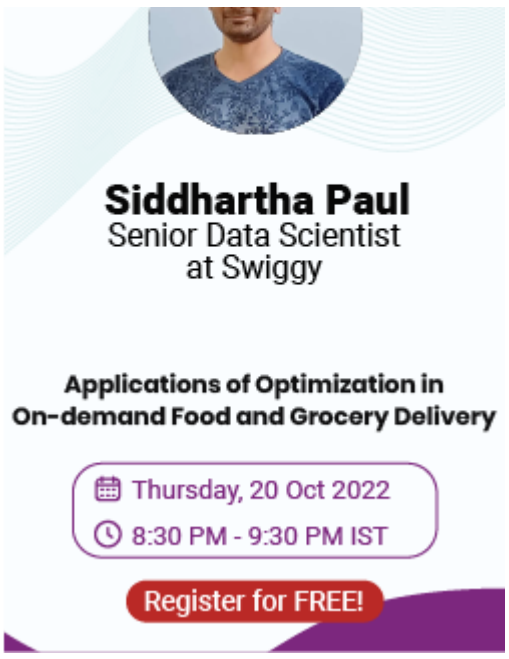
My medium page: https://ankitajhumu.medium.com/

The media shown in this article are not owned by Analytics Vidhya and is used at the Author's discretion.

blogathon clustring K Means Algorithm unsupervised learning

We use cookies on Analytics Vidhya websites to deliver our services, analyze web traffic, and improve your experience on the site. By using Analytics Vidhya, you agree to our Privacy Policy and Terms of Use. Accept

K-Mean: Getting The Optimal Number Of Clusters



About the Author

Ankita Banerji

Our Top Authors

view more

Download

Analytics Vidhya App for the Latest blog/Article

Previous Post

Multiclass Classification Using SVM

Next Post

Feature Scaling Techniques in Python – A Complete Guide

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

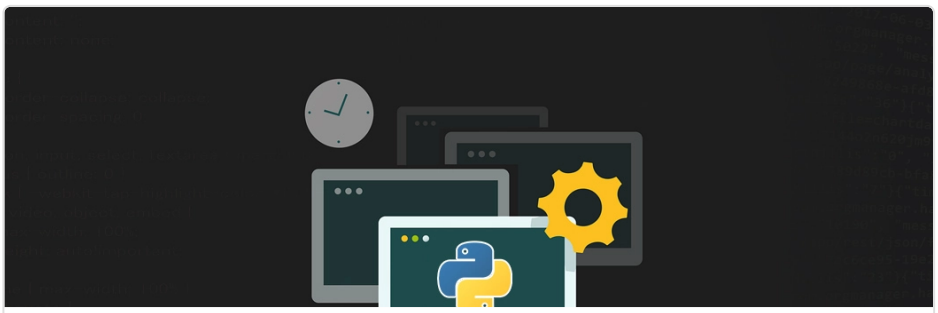
We use cookies on Analytics Vidhya websites to deliver our services, analyze web traffic, and improve your experience on the site. By using Analytics Vidhya, you agree to our Privacy Policy and Terms of Use. Accept

K-Mean: Getting The Optimal Number Of Clusters

Notify me of new posts by email.

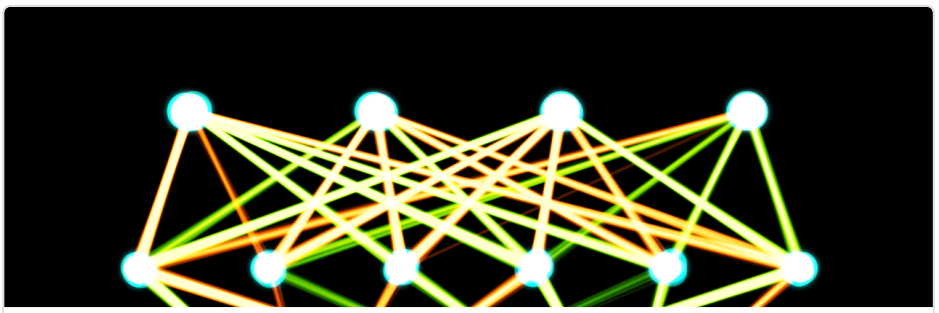
Submit

Top Resources



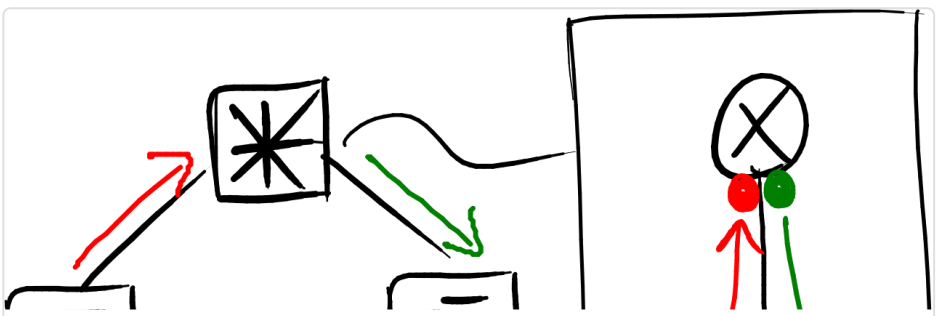
[Python Tutorial: Working with CSV file for Data Science]

Harika Bonthu - AUG 21, 2021



[Boost Model Accuracy of Imbalanced COVID-19 Mortality Prediction Using GAN-based..]

Bala Gangadhar Thilak Adiboina - OCT 07, 2020



[Introductory guide on Linear Programming for (aspiring) data scientists]

avcontentteam - FEB 28, 2017

[Understanding Random Forest]

Sruthi E R - JUN 17, 2021

Download App

Analytics Vidhya

About Us

Our Team

Careers

Contact us

Companies

Post Jobs

Trainings

Hiring Hackathons

Advertising

Data Scientists

Blog

Hackathon

Discussions

Apply Jobs

Visit us

We use cookies on Analytics Vidhya websites to deliver our services, analyze web traffic, and improve your experience on the site. By using Analytics Vidhya, you agree to our Privacy Policy and Terms of Use. Accept