Sign up for free to save articles, follow topics, and more. Sign in Sign up

An introduction to clustering algorithms

Raccess Rabi
Software Engineer

Published in Towards Data Science

Take a look at the image below. It's a collection of bugs and creepy crawlies at different shapes and sizes. Take a moment to categorize them by relatively like number of same type.

Did you categorize bugs with similar looking features together?

Image credit: Image by Free Vector at Vecteezy.com

Great! While doing this exercise, by "similar" I mean that, in your head, you implicitly assigned a "similarity" score between each pair of bugs. Based on this score, you then decided which bugs belong to the same group, and which bugs belong to different groups.

That wasn't too hard, eh? This is what this article is for, to work with non-screen bugs. Jokes aside, this is what we do in machine learning when we want to find groups of similar data points. This is called clustering.

For a machine through an algorithm to find similarities between data points, we first need to define what we mean by similarity. Similarity is a measure of how close two data points are to each other. (0.0 for different and maximum value for identical).

Clustering is a type of unsupervised learning. Unsupervised learning means that there is no outcome to predict, and the algorithm just tries to find patterns in the data.

There are many ways to do clustering, and there are many algorithms that perform clustering. In this article I will cover the following algorithms that fall under this category of unsupervised learning:

- K-means clustering
- Hierarchical clustering
- Density based clustering (DBSCAN)

K-means clustering

Concept

K-means clustering is the most simple and most intuitive clustering algorithm. The idea is that the closer data points are to each other, the more likely they are to belong to the same cluster. K-means clustering is based on this idea.

Method Example

Then, pick 3 random points to serve as the centroids of the 3 clusters. Place the centroids on the plot.

Plot randomly placed centroids, where k = 3

Step 2: For each point, assign it to the group with the closest mean. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Assign points to the closest centroid

Repeat: Move the centroids to the average of the points in the cluster. E.g. Place the centroids in the following image. Hint: The centroids are the same color as the points in the cluster.

Move centroids to the average of the points in the cluster

Repeat step 2 and 3 until the group does not change. E.g. Place the centroids in the following image. Hint: The centroids are the same color as the points in the cluster.

Final representation, where k = 3

K-means clustering

Pros and Cons

Pros & Cons

Pros: Simple, quick, scalable. Works well with large datasets. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Cons: Assumes spherical shaped clusters. Not suitable for non-spherical shapes. Sensitive to outliers. Requires specification of the number of clusters (k).

Pros details

Pros: Simple, quick, scalable. Works well with large datasets. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Simple, quick, scalable

Pros: Simple, quick, scalable. Works well with large datasets. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Works well with large datasets

Pros: Simple, quick, scalable. Works well with large datasets. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Works well with large datasets

Pros: Simple, quick, scalable. Works well with large datasets. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Works well with large datasets

Pros: Simple, quick, scalable. Works well with large datasets. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Works well with large datasets

Cons details

Cons: Assumes spherical shaped clusters. Not suitable for non-spherical shapes. Sensitive to outliers. Requires specification of the number of clusters (k).

Cons: Assumes spherical shaped clusters. Not suitable for non-spherical shapes. Sensitive to outliers. Requires specification of the number of clusters (k).

Cons: Assumes spherical shaped clusters. Not suitable for non-spherical shapes. Sensitive to outliers. Requires specification of the number of clusters (k).

Cons: Assumes spherical shaped clusters. Not suitable for non-spherical shapes. Sensitive to outliers. Requires specification of the number of clusters (k).

Cons: Assumes spherical shaped clusters. Not suitable for non-spherical shapes. Sensitive to outliers. Requires specification of the number of clusters (k).

Cons: Assumes spherical shaped clusters. Not suitable for non-spherical shapes. Sensitive to outliers. Requires specification of the number of clusters (k).

Hierarchical clustering

Concept

Hierarchical clustering is a clustering algorithm that creates a hierarchy of clusters.

Method Example

Hierarchical clustering starts by treating each data point as a cluster and calculate the distance between each pair of clusters and then merge the two closest clusters into one cluster. Repeat this process until only one cluster is left.

Hierarchical clustering

Step 1: Treat each data point as a cluster. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Treat each data point as a cluster

Repeat: Calculate the distance between each pair of clusters. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Calculate the distance between each pair of clusters

Repeat: Merge the two closest clusters. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Merge the two closest clusters

Repeat: Keep repeating the process until only one cluster is left. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Keep repeating the process until only one cluster is left

Hierarchical clustering

Pros and Cons

Pros & Cons

Pros: Hierarchical clustering provides a hierarchy of clusters. The dendrogram is easy to interpret. Flexible in terms of distance metrics and linkage methods. Can uncover nested clusters.

Cons: Computationally expensive for large datasets. Sensitive to noise and outliers. Difficult to determine the optimal number of clusters from the dendrogram. Lack of scalability for very large datasets.

Pros details

Pros: Hierarchical clustering provides a hierarchy of clusters. The dendrogram is easy to interpret. Flexible in terms of distance metrics and linkage methods. Can uncover nested clusters.

Hierarchical clustering provides a hierarchy of clusters

The dendrogram is easy to interpret

Flexible in terms of distance metrics and linkage methods

Can uncover nested clusters

Cons details

Cons: Computationally expensive for large datasets. Sensitive to noise and outliers. Difficult to determine the optimal number of clusters from the dendrogram. Lack of scalability for very large datasets.

Computationally expensive for large datasets

Sensitive to noise and outliers

Difficult to determine the optimal number of clusters from the dendrogram

Lack of scalability for very large datasets

Density based clustering (DBSCAN)

Concept

Density based clustering (DBSCAN) groups together points that are close to each other based on their density. This method can identify clusters of any shape and size, and it can also identify points that are considered noise or outliers.

Method Example

Pick a random point. Then, pick all points that are within a distance of 1.0 from that point. This distance is called epsilon (ε). E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Pick a random point

Step 1: Pick all points that are within a distance of 1.0 from that point. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Pick all points that are within a distance of 1.0 from that point

Step 2: For each point in step 1, count the distance points. For each point in step 2, count the distance points. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Count the distance points

Repeat: Keep repeating the process until all points have been visited.

DBSCAN clustering

Pros and Cons

Pros & Cons

Pros: Can cluster data of any shape and size. Can identify outliers (noise points). Does not require specifying the number of clusters in advance. Robust to noise and outliers.

Cons: Sensitive to the choice of epsilon (ε) and minimum points. Performance depends on the distance metric used. Can struggle with clusters of varying density. Computationally expensive for large datasets.

Pros details

Pros: Can cluster data of any shape and size. Can identify outliers (noise points). Does not require specifying the number of clusters in advance. Robust to noise and outliers.

Can cluster data of any shape and size

Can identify outliers (noise points)

Does not require specifying the number of clusters in advance

Robust to noise and outliers

Cons details

Cons: Sensitive to the choice of epsilon (ε) and minimum points. Performance depends on the distance metric used. Can struggle with clusters of varying density. Computationally expensive for large datasets.

Sensitive to the choice of epsilon (ε) and minimum points

Performance depends on the distance metric used

Can struggle with clusters of varying density

Computationally expensive for large datasets

Graph Community Detection

Use when

Use when the data can be represented as a network, or a graph.

How it works

Graph community detection algorithms attempt to cluster all nodes and edges into communities based on their connections. Nodes with more connections are more likely to belong to the same community.

Method example

Pick a random node. Then, pick all nodes that share a set number of connections with the random node. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Pick a random node

Perhaps the nodes share one edge connection.

Here, the nodes share two edge connections.

Finally, the nodes share three edge connections.

Step 1: Pick all nodes that share a set number of connections with the random node. E.g. Place the 3 points in the following image. Hint: The closest points are the same color.

Pick all nodes that share a set number of connections with the random node

Repeat: Pick a new random node, and repeat step 1 until all nodes have been visited.

Graph community detection

Pros and Cons

Pros & Cons

Pros: Effective for complex, real-world data represented as networks. Can uncover hierarchical community structures. Flexible in defining connections and similarity.

Cons: Performance depends on the choice of graph construction. Computationally expensive for large networks. Interpretation of results can be subjective.

Pros details

Pros: Effective for complex, real-world data represented as networks. Can uncover hierarchical community structures. Flexible in defining connections and similarity.

Effective for complex, real-world data represented as networks

Can uncover hierarchical community structures

Flexible in defining connections and similarity

Cons details

Cons: Performance depends on the choice of graph construction. Computationally expensive for large networks. Interpretation of results can be subjective.

Performance depends on the choice of graph construction

Computationally expensive for large networks

Interpretation of results can be subjective

Conclusion

Clustering is a powerful technique in machine learning and data science that helps us make sense of complex data. It's fascinating how it can reveal hidden patterns and relationships in the data, which can be crucial for various applications, from marketing to scientific research. Despite exercises involving bugs and crawlies, clustering can be a very useful tool.

Do you have any questions or comments?

Do you have any ideas for future topics?

If this article was helpful, Clap!

Leave a comment, follow me, and I'll let you know when I write a new article.

Join Medium with my referral link - Rabi Abdallah

Recommended from Medium

Transformers Explained Visually (Part 1): How LLMs Work, Step-by-Step
Intro to Machine Learning: Linear Regression
Machine Learning Mastery: The Complete Guide
Ultimate Guide to Machine Learning: Key Concepts, Algorithms, and Applications
Machine Learning for Beginners: Overview of Machine Learning
Intro to Machine Learning: Learning Curves
Intro to Machine Learning: Cross-Validation
Intro to Machine Learning: Bias-Variance Tradeoff
Intro to Machine Learning: Feature Engineering
Intro to Machine Learning: Feature Selection

See more recommendations

Help Status About Careers Press Blog Privacy Rules Terms Text to speech