Clustering K-Means Algorithms Cluster Analysis +3

How can we choose a "good" K for K-means clustering?

This question previously had details. They are now in a comment. More

Answer Follow · 60 Request 1

13 Answers

Andrea Trevino, works at DataScience
Answered Dec 31, 2016

(Excerpt from Introduction to K-means Clustering)

In general, there is no method for determining the exact value of *K*, but an accurate estimate can be obtained using the following techniques.

One of the metrics that is commonly used to compare results across different values of *K* is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing *K* will *always* decrease this metric, to the extreme of reaching zero when *K* is the same as the number of data points. Thus, this metric cannot be used ... (more)

Upvote · 20 Share

Add a comment... Recommended All

Daniel Martín, Software Engineer at PSPDFKit
Answered Dec 6, 2013 · Upvoted by Avinash Mishra, MSc Computer Science & Machine Learning, Technical University of Munich (2017) and Manohar Kuse, PhD Candidate researching computer vision and machine learning in robotics.

You can choose the number of clusters by visually inspecting your data points, but you will soon realize that there is a lot of ambiguity in this process for all except the simplest data sets. This is not always bad, because you are doing unsupervised learning and there's some inherent subjectivity in the labeling process. Here, having previous experience with that particular problem or something similar will help you choose the right value.

If you want some hint about the number of clusters that you should use, you can apply the *Elbow method*:

First of all, compute the sum of squared erro... (more)

Upvote · 247 Share · 1

Add a comment... Recommended All

Kaushik Kasi, (Data Science && Bitcoin) Enthusiast
Answered Mar 24, 2015

A quick (and rough) method is to take the **square root of the number of data points divided by two,** and set that as the number of clusters. The elbow

Related Questions

How can I choose the best K in KNN (K nearest neighbour) classification?

How is the k-nearest neighbor algorithm different from k-means clustering?

What is a good way to choose initial points of k clusters in k-means clustering?

What are the advantages of K-Means clustering?

What is the k-Means algorithm and how does it work?

What are the proper settings for document clustering with K-means?

How can I choose the optimal number of clusters in K-mode clustering?

How do you interpret k-means clustering results?

Why initial seed selection is important in K-means Clustering? Seed selection algorithm like-SPSS

What is the difference between k-means and hierarchical clustering?

+ Ask New Question

More Related Questions

For example, if I was trying to separate a population into 3 shirt sizes, I would be optimizing for the best location of the 3 clusters, rather than optimizing for the number of clusters that best segment the data.

$$k \approx \sqrt{n/2}$$

From: Determining the number of clusters in a data set
Paper on using a kernel matrix to find the optimal number of clu... (more)

Upvote · 25 Share

Add a comment... Recommended All

Soriz Nabil, phd Machine Learning
Answered May 12, 2018

You may use **G-means** (Gaussian-means algorithm). It discovers the number of clusters automatically using a statistical test to decide whether to split a k-means center into two. This algorithm takes a hierarchical approach to detect the number of clusters, based on a statistical test for the hypothesis that a subset of data follows a Gaussian distribution (continuous function which approximates the exact binomial distribution of events), and if not it splits the cluster. It starts with a small number of centers, say one cluster only (k=1), then the algorithm splits it into two centers (k=2) ... (more)

Upvote · 1 Share

Add a comment... Recommended All