

Partitional Clustering

K-Means Algorithms

Unsupervised Learning

Cluster Analysis

Machine Learning

How do we apply k-means clustering algorithm for mixed data-numeric and categorical?

Ad by DatadogHQ.com

End-to-end visibility with the addition of Datadog Synthetics.

Fully integrated with Datadog's infrastructure metrics, logs, and APM.

[Learn More](#)

4 Answers



Shehroz Khan, ML Researcher, Postdoc @U of Toronto

Answered Jan 12, 2016 · Author has **1.4k** answers and **4.2m** answer views

Originally Answered: How do we apply k-means clustering algorithm for mixed data-numeric and categorical ?

K-means cannot be directly used for data with both numerical and categorical values because of the cost function it uses. K-means uses Euclidean distance, which is not defined for categorical data. Therefore, to use K-means type or partitional clustering algorithm on mixed data you have to change the cost function s.t. it can capture distance or similarity between both the types of data.

Huang developed a simple method where Euclidean distance is used for finding similarity between numerical data and hamming distance for similarity between categorical data and combined both of them together with some weights as one cost function to handle mixed data. The paper is here <http://grid.cs.gsu.edu/~wkim/ind...>

You should read this more recent and highly cited paper on mixed data clustering using k-means type algorithm <http://edu.cs.uni-magdeburg.de/E...>

9.8k views · View 24 Upvoters · Answer requested by Tuhin Batra

Related Questions

More Answers Below

[How do I use a clustering algorithm on data that has both categorical and numeric value ?](#)

[Why does K-means clustering perform poorly on categorical data? The weakness of the K-means method is that it is applicable only when the mean...](#)

[How do I do clustering for categorical data?](#)

[What's a clustering algorithm for purely categorical data?](#)

[Can a binary categorical variable be used in K-means clustering?](#)

Related Questions

[How do I use a clustering algorithm on data that has both categorical and numeric value ?](#)

[Why does K-means clustering perform poorly on categorical data? The weakness of the K-means method is that it is applicable only when the mean...](#)

[How do I do clustering for categorical data?](#)

[What's a clustering algorithm for purely categorical data?](#)

[Can a binary categorical variable be used in K-means clustering?](#)

[What type of data is best suited for K-means clustering?](#)



Want to explore more

Try one of the questions or continue to log in and view more.

Continue



John Sanders, Retired AI researcher and software engineer (Mathematics, Geology degrees)

Answered May 26, 2017 · Author has **1.8k** answers and **339.4k** answer views

Originally Answered: How do we apply k-means clustering algorithm for mixed data-numeric and categorical ?

The problem of categoric data is that it is discrete and cannot sensibly be linearly interpolated.

Eg the mineral quartz can be {clear , white, pink , red , yellow, brown, black, purple } (categoric set)

This discrete set could be laid in continuous manner but it would reflect an arbitrary basis and so cannot be used. How can calculate the distance between pink sample and a yellow one and still make sense of the distance between a clear and black sample? It has no meaning - it also suggests that categories like this should be excluded. And yet all belong to the same class "mineral quartz " . In this and perhaps other situations let the other attributes dominate (providing they are independent) (use Principle component analysis which it find you a basis set - ie independent set)

And then use the categoric attributes to further subdivide the class.

{rock crystal, milky, rose, citrine (or ferruginous) , cairngorm, morion, amethyst} Quartz

(avoiding rutilated; crypto chrystalline - chalcedony, opal ; specials christobolite and tridymite)

If the categoric attribute is definitive eg number of legs then you split your data into that number of classes (one for each number) then K-means each set.

Also note that large multi-dimensional sets of attributes the range can dominate the importance so a distance measure such direction cosines (see T Kohonen, Self Organization and associative memory (Springer Verlag) Page 60)) can be a better option than Euclidean ,


3.8k views · View 2 Upvoters

Sponsored by MathWorks

Free guide to machine learning basics and advanced techniques.

Download the ebook and discover that you don’t need to be an expert to get started with machine learning.

 Download




Anoop Vasant Kumar, Data Scientist

Answered Mar 22, 2016 · Author has 123 answers and 511.6k answer views

Originally Answered: How do we apply k-means clustering algorithm for mixed data-numeric and categorical ?

The answer to this question has a couple of suggestions on how to get k-means clustering to work on a data mix(numeric and categorical) [Why does K-means clustering perform poorly on categorical data? The weakness of the K-means method is that it is applicable only when the mean is defined, one needs to specify K in advance, and it is unable to handle noisy data and outliers.](#)

3.3k views



Anonymous

Answered Jun 7, 2016

Originally Answered: How do we apply k-means clustering algorithm for mixed data-numeric and categorical ?

You can use MFA (multiple factor analysis) for getting the factor score for the observations. These factor score will be continuous values. Than you can apply k-means clustering.

1.7k views



Related Questions

How do I use a clustering algorithm on data that has both categorical and numeric value ?

Why does K-means clustering perform poorly on categorical data? The weakness of the K-means method is that it is applicable only when the mean...

How do I do clustering for categorical data?

What's a clustering algorithm for purely categorical data?

Can a binary categorical variable be used in K-means clustering?

What type of data is best suited for k-means clustering?

How do I do clustering of mixed data types in Python?

How can I use KNN for mixed data (categorical and numerical)?

How do you implement a K-means clustering algorithm on both numerical and categorical data?

Is an attribute numeric or categorical in a K-means algorithm?

How can I apply dimensionality reduction on mixed data (categorical and hybrid)?

What happens when you try clustering data with higher dimensions using k-means? For example, if the dimensionality of the data set is 1000, nu...

Why doesn't clustering algorithms go well with numeric (non-mixed) data?

How do we apply the k-means clustering algorithm for a mixed data numerical and text?

How can I apply an SVM for categorical data?