Standardization vs. normalization ?

Sign in to follow this Followers 1

By Saurabh Jain, February 21 in Data Management - Data Cleaning & Data Plumbing

Saurabh Jain
Advanced Member

Administrators
5
38 posts

Posted February 21 Report post

In the overall knowledge discovery process, before data mining itself, data preprocessing plays a crucial role. One of the first steps concerns the normalization of the data. This step is very important when dealing with parameters of different units and scales. For example, some data mining techniques use the Euclidean distance. Therefore, all parameters should have the same scale for a fair comparison between them.

Two methods are usually well known for rescaling data. Normalization, which scales all numeric variables in the range [0,1]. One possible formula is given below:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

On the other hand, you can use standardization on your data set. It will then transform it to have zero mean and unit variance, for example using the equation below:

$$x_{new} = \frac{x - \mu}{\sigma}$$

Both of these techniques have their drawbacks. If you have outliers in your data set, normalizing your data will certainly scale the “normal” data to a very small interval. And generally, most of data sets have outliers. When using standardization, your new data aren’t bounded (unlike normalization).

So my question is what do you usually use when mining your data and why?

AdChoices
Normalization Examples
Database Normalization
Big Data and Data Mining
Analytic Big Data VS

TOPICS

B Bach of business information systems student. 0
By BRandon
Started September 15

 Mr 0
By Arjun Regmi
Started August 24

 BI on prem 0
By John Wittenbrook
Started August 11

 Web focus 0
By John Wittenbrook
Started August 11