

Re: how to compute distance metrics with multi dimensional data

Source: <http://sci.tech-archive.net/Archive/sci.nonlinear/2005-02/0019.html>

From: bluelagoon (bluelagoontrading_at_hotmail.com)

Date: 02/12/05

Date: 12 Feb 2005 05:27:06 -0800

Lou Pecora wrote:

> In article <1108143670.835637.54270@z14g2000cwz.googlegroups.com>,

> "bluelagoon" <bluelagoontrading@hotmail.com> wrote:

>

>> Lou,

>> no, it does not seem right. ok, i'll make it step by step real simple

>> i got 2d time series 1000 vectors, column one = cycle amplitude in

>> points,

>> colume two = cycle duration in seconds, points are not seconds, ie

>> different measurement units

>> 10, 8

>> 11, 5

>> 2, 3

>> 8, 4

>> 12, 1

>> 4, 24

>> 9, 14

>>

>> all the way till the 1000th row $t = 1000$

>>

>> now i embed with delay = 1 and dim = 4, in pairs! i'll show 4 rows

as

>> an example

>> (10,8)(11,5)(2,3) (8,4)

>> (11,5)(2,3) (8,4) (12,1)

>> (2,3) (8,4) (12,1)(4,24)

>> (8,4) (12,1)(4,24)(9,24)

>>

>> all the way till the 1000th row

>>

>> now, how do i compute the euclidean between the rows ???

considering

>> that "points" are not "seconds"

>> that's all i wanted to know.

>>

> > *thanks.*
>
> *We may be converging, but I think it is to what I originally suggested*
> -- *what you called z-score (zero mean, std=1). You're working in an 8D space. As you said you have to compare apples to apples. Once you do this then the Euclidean metric works as usual (a(t) and b(t) are the two columns of demeaned, rescaled data):*
>
>
> *vector norm=sqrt[a^2(t)+ b^2(t)+ a^2(t+1)+ b^2(t+1)+ a^2(t+2)+ b^2(t+2)+ a^2(t+3)+ b^2(t+3)]*
>
> *distance between time sequential points (as below)=*
> *sqrt[(a(t)-a(t+1))^2+(b(t)-b(t+1))^2+ (a(t+1)-a(t+2))^2+(b(t+1)-b(t+2))^2+ (a(t+2)-a(t+3))^2+(b(t+2)-b(t+3))^2+ (a(t+3)-a(t+4))^2+(b(t+3)-b(t+4))^2]*
>
> *It generalizes to any dimension.*
>

Lou, thanks.

although it's a little hard for me to believe that it's simply the same formula for multi dimensional time series, but i'll take it.

i think that misunderstanding came about the word "dimension" i used it with two meanings:

1. time series is 2-d if time series data is represented by 2 vars, ie 2d input data or i should've called it multivariate input data instead of typical univariate input data 1d.
2. dimension as in embedding dimension.

in any case, i think the with the above approach the neighbor distance comparing may have problems if data has outliers or somehow variables in multivar data have different importance. outlier cutting might be detrimental... so i think that ultimately it's up the user to define distance metric... a custom metric might be the solution

thanks a million anyway, and thanks for being patient with me.

ps. by the way,i am about to post another question on types of mapping functions from selected nearest neighbors to forecasts, more interested in non linear mapping functions and want to know what others tried and found to be the best.