CSCI-B 565 DATA MINING
Homework 2
Morning class
Computer Science Core
Fall 2013
Indiana University

Ramakant Khandel
khandelr@indiana.edu

September 20, 2013

1. Definitions

1)a:

A metric on a set X is a function d:X*X–¿R such that 1)d(x,y)¿=0 for all x,y X 2)d(x,y)=0 iff x=y 3)d(x,y)=d(y,x) 4)d(x,z)¡=d(x,y)+d(y,z) for all x,y,z X

For the distance function to be metric it need to satisfy the above 4 properties d(x,y)=max—x-y— 1¡= i ¿=n

Solution:

Since d(x,y)=max—x-y— the distance between x,y will be greater than or equal to zero for any x,y

Since d(x,y)=max—x-y— has modulus the symmetry property is satisfied.

Since d(x,y)=max—x-y— and by triangular inequality theorem d(x,z)¡=d(x,y)+d(y,z) It is a distance metric.

1)b:

A metric on a set X is a function d:X*X–¿R such that 1)d(x,y)¿=0 for all x,y X 2)d(x,y)=0 iff x=y 3)d(x,y)=d(y,x) 4)d(x,z)¡=d(x,y)+d(y,z) for all x,y,z X

d(x,y)=sum c(x,y)/i 1¡= i ¿=n

c(x,y)=1 if x nt eq y =0 otherwise For the distance function to be metric it need to satisfy the above 4 properties

d(x,y)=0 for x=y since c(x,y)=0 for x=y d(x,y)¿=0 for x nt eq y since c(x,y)=1 for x nt eq y

Since d(x,y)=sum c(x,y)/i and by triangular inequality theorem d(x,z)¡=d(x,y)+d(y,z) It is a distance metric.

1)c-1:

A metric on a set X is a function d:X*X–¿R such that 1)d(x,y)¿=0 for all x,y X 2)d(x,y)=0 iff x=y 3)d(x,y)=d(y,x) 4)d(x,z)¡=d(x,y)+d(y,z) for all x,y,z X

d(x,y)=d0*d1

d0 and d1 are distance metrics— given For the distance function to be metric it need to satisfy the above 4 properties

d(x,y)=0 since d0 and d1 are equal to 0 for x=y

d(x,y)¿=0 since d0 and d1 are ¿= 0 for x,y

d(x,y)=d(y,x) since d0 and d1 are symmetric therefore the distance function will itself be symmetric

d(x,y)=d0*d1 and by triangular inequality theorem d(x,z)¡=d(x,y)+d(y,z) It is a distance metric.

1)c-2:

A metric on a set X is a function d:X*X–¿R such that 1)d(x,y)¿=0 for all x,y X 2)d(x,y)=0 iff x=y 3)d(x,y)=d(y,x) 4)d(x,z)¡=d(x,y)+d(y,z) for all x,y,z X

d(x,y)=(d0+d1)/d0d1

d0 and d1 are distance metrics— given For the distance function to be metric it need to satisfy the above 4 properties

d(x,y)=0/0 since d0 and d1 are equal to 0 for x=y It is not a distance metric.

1)c-3:

A metric on a set X is a function d:X*X–¿R such that 1)d(x,y)¿=0 for all x,y X 2)d(x,y)=0 iff x=y 3)d(x,y)=d(y,x) 4)d(x,z)¡=d(x,y)+d(y,z) for all x,y,z X

d(x,y)=maxd0,d1

d0 and d1 are distance metrics— given For the distance function to be metric it need to satisfy the above 4 properties

d(x,y)=0 since d0 and d1 are equal to 0 for x=y

d(x,y)¿=0 since d0 and d1 are ¿= 0 for x,y

d(x,y)=d(y,x) since d0 and d1 are symmetric therefore the distance function will itself be symmetric

d(x,y)=maxd0,d1 and by triangular inequality theorem d(x,z)¡=d(x,y)+d(y,z)

It is a distance metric.

1)c-4:

A metric on a set X is a function d:X*X–¿R such that 1)d(x,y)¿=0 for all x,y X 2)d(x,y)=0 iff x=y 3)d(x,y)=d(y,x) 4)d(x,z)¡=d(x,y)+d(y,z) for all x,y,z X

d0 and d1 are distance metrics— given For the distance function to be metric it need to satisfy the above 4 properties

d(x,y)¿0 since ——x intersection y—— ¿ 0 for x=y

It is not a distance metric.

2:

Foo
X Y Z A

1 1 Y abcd
3 255 N bcde
4 4 N bcd
2 1 Y acde
20 1 Y bdf
5 4 T fg
5 3 Y abf

The above data can be cleaned or transformed to

Foo
X Y Z A
1: 1 1 Y abcd
2: 3 2 N bcde
3: 4 4 N bcd
4: 2 1 Y acde
5: 2 1 Y bdf
6: 5 4 Y f
7: 5 3 Y abf

Euclidean Distance Z A(Jaccard)
0 1 4/6
1 0 4/6
0 0 3/6
1 1 4/6
1 1 3/6
1 1 1/6
2 0 3/6

Jaccard coefficient measures similarity between samplesets as it is defined as the size of intersection divided by the size of union of sets

Euclidean distance measures dissimilarity

Similar data
0 1 4/6 -1

Dissimilar data
1 1 1/6 -6

Misplaced Data
0 0 3/6 -3
1 1 4/6 -4
1 1 3/6 -5
2 0 3/6 -7

1 0 4/6 -2

5-2)a:

continuous ordinal

5-2)b:

continuous ratio

5-2)c:

discrete ordinal

5-2)d:

continuous ratio

5-2)e:

discrete ordinal

5-2)f:

continuous ordinal

5-2)g:

discrete ratio

5-2)h:

continuous ratio

5-2)i:

discrete ordinal

5-2)j:

discrete ordinal

5-2)k:

continuous ratio

5-2)l:

continuous ratio

5-2)m:

nominal discrete

5-3)a:

The boss is correct. Satisfaction= number of complaints /total number of sales for product

5-3)b:

The number of complaints for a product is directly proportional to the number of sales for the product. The satisfaction measure only considers number of complaints and not sales for the product. The attribute type of satisfaction attribute is ordinal.

5-6)a:

Association analysis is used to find the hidden relationship among data. It describes how data item are associated with each other. The educational psychologist can use the table which contains answers of the questions by the student to find pattern among the data sets. This can help them to classify on what questions majority of the students scored and not scored.

5-6)b:

discrete ordinal. Their are 100 attributes for each question.

5-12)a:

No

5-12)b:

Yes A data object that deviates significantly from the normal object as if it were generated by a different mechanism.

5-12)c:

No Outliers are different from the noise data.

5-12)d:

No A data object that deviates significantly from the normal object as if it were generated by a different mechanism.

5-12)e:

Yes

5-13)a:

The order of data objects and the distance between the data objects will contain duplicate entries.

5-13)b:

The problem can be fixed by considering one object for duplicate objects.

5-16)a:

An inverse document frequency factor is incorporated which diminishes the weight of the terms that occur very frequently in the document set and increase the weight of terms that occur rarely.

5-16)b:

Inverse document frequency is a measure of whether the term is common or rare across all doments to distinguish documents.

5-19)a:

cosine: 1 correlation: 0/0(undefined) euclidean: 2

5-19)b:

cosine: 0 correlation: -1 euclidean: 2 jaccard: 0

5-19)c:

cosine: 0 correlation: 0 euclidean: 2

5-19)d:

cosine: 0.75 correlation: 0.25 jaccard: 0.6

5-19)e:

cosine: 0 correlation: 0

5-24)a:

A proximity matrix P is an m by m matrix containing all the pairwise dissimilarities or similarities between the objects being considered. distance , centroid for points in euclidean space can be used to compute proximity

5-24)b:

The distance between the centroids can be used to define the distance between two sets of points in Euclidean Space.

5-24)c:

If the distance values are proximity measures then the gretaer the proximity values the more similar are the objects.