Data Mining Assignment 1

ADITYA TANIKANTI - 0003567018

February 2, 2016

Solution 1

a)

Given $x_{ij} = \frac{m_{ij}}{m_i}.log\frac{n}{n_j}$ where m_{ij} is the number of times the jth term appears in the ith document,m_i is the number of terms in the i-th document and n_j is the number of documents containing the j-th term in the data set and n is the number of documents.

While using this encoding in text mining, if we consider the first fraction $\frac{m_{ij}}{m_i}$ it computes the Term Frequency, i.e the number of times a word appears in a document, divided by the total number of words in that document.

The second fraction $log\frac{n}{n_j}$ which is the Inverse Document Frequency (IDF), computes the logarithm of the number of the documents which are taken into consideration(a.k.a corpus) divided by the number of documents where the specific term appears. This helps in determing how important a term is.

a)The main advantage is that it helps in reflecting the how important a word is to a document in a collection of documents.

b)It also helps in finding out the similarity of two documents with respect to word-count.

c)It helps in removal of stop words / words with not much relevance. Words like “is”, “of”, “the”and “that”, may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing this fraction

d)The encoding is easy to compute

e)Since every document is different in length it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization.

b)

When compared to $x_{ij} = \frac{m_{ij}}{m_i}$, the $log\frac{n}{n_j}$ part of the encoding $x_{ij} = \frac{m_{ij}}{m_i}.log\frac{n}{n_j}$, intuitively considers only those words which rarely occur over a collection of documents. It considers this to be valuable.The importance of each term is assumed to be inversely proportional to the number of documents that the term occurs.

Problems with text classification would end up ignoring important and relevant words using this encoding. i.e. there might be cases that the term which occurs widely in the document collection is relevant and that these words represent document or the topics of the document better but such text classification scenarios may be completely disregarded.

c)

In the encoding $x_{ij} = \frac{m_{ij}}{m_i}.log\frac{n}{n_j}$ If a term occurs in one document then the $log\frac{n}{n_j}$ part of the encoding gives a significant value greater than 0 when the total number of documents is greater than 1, however the $\frac{m_{ij}}{m_i}$ tends to 0 as there are documents where the word never occurs, thus x_{ij} equates to zero.

When the word occurs in every document the $log\frac{n}{n_j}$ of the encoding always tends to 0 as the value of numerator is always equal to the denominator thus always eliminating the word which appears across all the documents and thus the weight of x_{ij} for that word always equates to zero.

Solution 2

To prove $cos(x, y) = \frac{x^T.y}{||x||.||y||}$

Let's begin with a right triangle with sides x, y and hypotenuse z, trigonometry determines the cosine of the angle θ between side x and the hypotenuse as: $\cos\theta = \frac{x}{z}$.

If we consider the dot product for two vectors: $\vec{x} = (x_1, x_2, x_3, \ldots)$ and $\vec{y} = (y_1, y_2, y_3, \ldots)$, where x_n and y_n are the components of the vector.

$\vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$

To understand this, we need to understand what is the geometric definition of the dot product:

$\vec{x} \cdot \vec{y} = \|\vec{x}\| \|\vec{y}\| \cos\theta$

Rearranging the equation to understand it better using the commutative property, we have:

$\vec{x} \cdot \vec{y} = \|\vec{y}\| \|\vec{x}\| \cos\theta$

The term $\|\vec{x}\| \cos\theta$ is the projection of the vector $\vec{x}$ into the vector $\vec{y}$

The length or magnitude or norm of the vector x is denoted by ||x||.

The length of the vector x can be computed with the Euclidean norm

$\|\mathbf{x}\| = \sqrt{{x_1}^2 + {x_2}^2 + {x_3}^2}$ which is a consequence of the Pythagorean theorem since the basis vectors x1, x2, x3 are orthogonal unit vectors.

Thus we get $\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \, \|\mathbf{y}\| \cos\theta$ where θ is the measure of the angle between x and y . Geometrically, this means that x and y are drawn with a common start point and then the length of x is multiplied with the length of that component of y that points in the same direction as x.

As we know the dot product can also be defined as the sum of the products of the components of each vector i.e.

$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + x_3 y_3$.

if we extend this for for k dimension column vectors of R^K we get $\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + x_3 y_3 ... x_K y_k = \sum_{i=1}^{k} x_i y_i$

The dot product of two column vectors x and y can be computed as the single entry of the matrix product $[\mathbf{x} \cdot \mathbf{y}] = \mathbf{x}^{\mathrm{T}}\mathbf{y}$, which is written as xi.yi in Einstein summation convention.

To illustrate the equivalent usage, consider three-dimensional Euclidean space, letting:

$\mathbf{x} = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$ $\mathbf{y} = y_1\mathbf{i} + y_2\mathbf{j} + y_3\mathbf{k}$ be two vectors where i, j, k (also denoted e1, e2, e3) are the standard basis vectors in this vector space (see also Cartesian coordinates). Then the dyadic product of a and b can be represented as a sum:

$$\begin{aligned} \mathbf{xy} = \; & x_1y_1\mathbf{ii} & +x_1y_2\mathbf{ij} & +x_1y_3\mathbf{ik} \\ & +x_2y_1\mathbf{ji} & +x_2y_2\mathbf{jj} & +x_2y_3\mathbf{jk} \\ & +x_3y_1\mathbf{ki} & +x_3y_2\mathbf{kj} & +x_3y_3\mathbf{kk} \end{aligned}$$

or by extension from row and column vectors, a 3×3 matrix (also the result of the outer product or tensor product of a and b):

$$\mathbf{xy} \equiv \mathbf{x} \otimes \mathbf{y} \equiv \mathbf{x}\mathbf{y}^{\mathrm{T}} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix} = \begin{pmatrix} x_1y_1 & x_1y_2 & x_1y_3 \\ x_2y_1 & x_2y_2 & x_2y_3 \\ x_3y_1 & x_3y_2 & x_3y_3 \end{pmatrix}.$$

For k dimensional matrix

$$
\begin{aligned}
\mathbf{x} \cdot \mathbf{y} &= \mathbf{x}^{\mathrm{T}}\mathbf{y} \\
&= \begin{pmatrix} x_1 & x_2 & \cdots & x_k \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix} \\
&= x_1 y_1 + x_2 b_2 + \cdots + x_k y_k \\
&= \sum_{i=1}^{k} x_i y_i,
\end{aligned}
$$

Hence as we can see $x.y = x^T y$

and thus we can rewrite the cosine rule from $x.y = ||x||||y||cos(x, y)$ to $x^T y = ||x||||y||cos(x, y)$

and thus to $cos(x, y) = \frac{x^T y}{||x||||y||}$

Solution 3

To prove metrics on sets we need to satisfy the following four laws A metric space is given by a set X and a distance function d : S × S -> R such that

i) (Positivity) For all A, B ϵ S d(A, B) >= 0.
ii) (Non-degenerated) For all A, B ϵ S d(A, B) = 0 iff A = B .
iii) (Symmetry) For all A, B ϵ S d(A, B) = d(B, A)
iv) (Triangle inequality) For all A, B, C ϵ S d(A, B) <= d(A, C) + d(C, B) .

a)

d1(A,B) = |A - B| + |B - A|

To prove it is a metric we verify (1)-(4).

For (1): d1(A, B) = |A - B| >= 0 & |B - A| >= 0

by the definition of the absolute value functions, hence |A - B| + |B - A| >= 0 so positivity is proved.

For (2): d1(A, B) = 0 if and only if |A - B| = 0 & |B - A| = 0 which can be the case if and only if A = B, Hence Non Degenerated is proved.

For (3): Since d1(A, B) = |A - B| + |B - A|

= |B - A| + |A - B| {as |A - B| = |B - A|}

= d(B, A). Symmetry is proved.

For (4): d1(A, B) = |A - B| + |B - A|

= |A - C + C - B| + |B - C + C - A|

= |A - C| + |C - A| + |C - B| + |B - C| {as |A + B| = |A| + |B|}

= d1(A, C) + d1(C, B).

Hence triangle inequality is proved

As the four rules of metrics for a set is proved d1(A,B) = |A - B| + |B - A| is a metric

b)

d2(A,B) = $\frac{|A-B|+|B-A|}{|AUB|}$

To prove it is a metric we verify (1)-(4).

For (1): d2(A, B) = $\frac{|A-B|}{|A\cup B|}$ >= 0 & $\frac{|B-A|}{|A\cup B|}$ >= 0 by the definition of the absolute value functions, hence $\frac{|A-B|+|B-A|}{|A\cup B|}$ >= 0.

We can also prove this by considering the denominator of the fraction to be

$|A \cup B|$ = |A| + |B| - |A n B|.

If $|A \cap B|$ = null, i.e. A and B are independent then the numerator of the fraction i.e. |A - B| = |A| & |B - A| = |B| because length of set difference is equal to the length of the first term of the difference.

= $\frac{|A-B|+|B-A|}{|A\cup B|}$

= $\frac{|A|+|B|}{|A|+|B|}$

hence $\frac{|A-B|+|B-A|}{|A\cup B|} = 1 > 0$

Thus positivity is proved.

For (2): d2(A, B) = 0 if and only if $\frac{|A-B|}{|A\cup B|} = 0$ & $\frac{|B-A|}{|A\cup B|} = 0$ which can be the case if and only if A = B, Hence Non Degenerated is proved.

For (3): Since d2(A, B) = $\frac{|A-B|}{|A\cup B|}$ + $\frac{|B-A|}{|A\cup B|}$

= $\frac{|B-A|}{|A\cup B|}$ + $\frac{|A-B|}{|A\cup B|}$ {as length of A - B = length of B - A}

= d2(B, A). Symmetry is proved.

For (4):

To prove (Triangle inequality) For all A, B, C ϵ S

d(A, B) <= d(A, C) + d(C, B) .

Consider L.H,S

d2(A, B)

= $\frac{|A-B|+|B-A|}{|A\cup B|}$

= $\frac{|A-B|}{|A\cup B|}$ + $\frac{|B-A|}{|A\cup B|}$

Consider A & B to be independent with no matching pairs

= $\frac{|A|+|B|}{|A|+|B|}$ {as $|A - B| = |A|$ & $|A \cup B| = |A| + |B|$}

= 1

Now consider R.H.S

d2(A, C) + d2(C, B)

= $\frac{|A-C|}{|AUC|} + \frac{|C-A|}{|A\cup C|} + \frac{|C-B|}{|C\cup B|} + \frac{|B-C|}{|C\cup B|}$

if A & C are independent , B & C are independent

= $\frac{|A|}{|A+C|} + \frac{|C|}{|A+C|} + \frac{|C|}{|C+B|} + \frac{|B|}{|C+B|}$

= $\frac{|A|+|C|}{|A|+|C|} + \frac{|C|+|B|}{|C|+|B|}$ { as |A + B| = |A| + |B| }

= 2

Thus R.H.S > L.H.S

Hence triangle inequality is proved

As the four rules of metrics for a set is proved d2(A,B) = $\frac{|A-B|+|B-A|}{|A \cup B|}$ is a metric

c)

d3(A,B) = 1 - $\left(1/2 * \frac{|A \cap B|}{|A|} + 1/2 * \frac{|A \cap B|}{|B|}\right)$

To prove it is a metric we verify (1)-(4).

For (1): d3(A, B) =

If $|A \cap B|$ = null, i.e. A and B are independent then

= 1 - $\left(1/2 * \frac{|A \cap B|}{|A|} + 1/2 * \frac{|A \cap B|}{|B|}\right)$

= 1 - $(1/2 * 0 + 1/2 * 0)$

= 1

Hence d3(A,B) >= 0

Thus positivity is proved.

For (2): To prove d3(A, B) = 0

Consider B = A. = 1 - $\left(1/2 * \frac{|A \cap B|}{|A|} + 1/2 * \frac{|A \cap B|}{|B|}\right)$

= 1 - $\left(1/2 * \frac{|A \cap A|}{|A|} + 1/2 * \frac{|A \cap A|}{|A|}\right)$

= 1 - $\left(1/2 * \frac{|A|}{|A|} + 1/2 * \frac{|A|}{|A|}\right)$

= 1 - $(1/2 * 1 + 1/2 * 1)$

= 1 - (1)

= 0

Hence if and only if A = B d3(A, B) = 0 , thus Non Degenerated is proved.

For (3): Since d3(A, B) = 1 - $\left(1/2 * \frac{|A \cap B|}{|A|} + 1/2 * \frac{|A \cap B|}{|B|}\right)$

can be rearranged as = 1 - $\left(1/2 * \frac{|B \cap A|}{|B|} + 1/2 * \frac{|B \cap A|}{|A|}\right)$ {as $|A \cap B| = |B \cap A|$}

= d3(B, A). Symmetry is proved.

For (4):

To prove (Triangle inequality) For all A, B, C ϵ S

d3(A, B) <= d3(A, C) + d3(C, B) .

Consider L.H,S

d3(A, B)

= 1 - $\left(1/2 * \frac{|A \cap B|}{|A|} + 1/2 * \frac{|A \cap B|}{|B|}\right)$

Consider A & B to be independent with no matching pairs then $|A \cap B| = 0$

= 1 - $\left(1/2 * \frac{|A \cap B|}{|A|} + 1/2 * \frac{|A \cap B|}{|B|}\right)$

= 1 - 0

= 1

Now consider R.H.S

d3(A, C) + d3(C, B)

= 1 - $\left(1/2 * \frac{|A \cap C|}{|A|} + 1/2 * \frac{|A \cap C|}{|C|}\right)$ + 1 - $\left(1/2 * \frac{|c \cap B|}{|C|} + 1/2 * \frac{|C \cap B|}{|B|}\right)$

if A & C are independent , B & C are independent

= 1 + 1

= 2

Thus R.H.S > L.H.S

Hence triangle inequality is proved

As the four rules of metrics for a set is proved d3(A,B) = 1 - $\left(1/2 * \frac{|AnB|}{|A|} + 1/2 * \frac{|A \cap B|}{|B|}\right)$ is a metric.

d)

d4(A,B) = 1 - $\left(1/2 * \frac{|A|}{|A \cap B|} + 1/2 * \frac{|B|}{|A \cap B|}\right)^{-1}$

To prove it is a metric we verify (1)-(4).

For (1): d4(A, B) =

If $|A \cap B|$ = null, i.e. A and B are independent then

= 1 - $\left(1/2 * \frac{|A|}{|A \cap B|} + 1/2 * \frac{|B|}{|A \cap B|}\right)^{-1}$

= 1 - $\left(1/2 * \frac{|A|+|B|}{|A \cap B|}\right)^{-1}$

= 1 - $\left(2 * \frac{|A \cap B|}{|A|+|B|}\right)$

As $|A \cap B|$ = null

= 1 - (0)

= 1

Hence d4(A,B) >= 0

Thus positivity is proved.

For (2): To prove d3(A, B) = 0

Consider B = A. d4(A,B) = 1 - $\left(1/2 * \frac{|A|}{|A \cap B|} + 1/2 * \frac{|B|}{|A \cap B|}\right)^{-1}$

= 1 - $\left(1/2 * \frac{|A|}{|A \cap B|} + 1/2 * \frac{|B|}{|A \cap B|}\right)^{-1}$

= 1 - $\left(1/2 * \frac{|A|+|B|}{|A \cap B|}\right)^{-1}$

= 1 - $\left(2 * \frac{|A \cap B|}{|A|+|B|}\right)$

if A = B

= 1 - $\left(2 * \frac{|A \cap A|}{|A|+|A|}\right)$

= 1 - $\left(2 * \frac{|A|}{2|A|}\right)$

= 1 - $\left(2 * \frac{1}{2}\right)$

= 0

Hence if and only if A = B d4(A, B) = 0 , thus Non Degenerated is proved.

For (3): Since d4(A,B) = 1 - $\left(1/2 * \frac{|A|}{|A \cap B|} + 1/2 * \frac{|B|}{|A \cap B|}\right)^{-1}$

can be rearranged as = 1 - $\left(1/2 * \frac{|B|}{|B \cap A|} + 1/2 * \frac{|A|}{|B \cap A|}\right)^{-1}$ {As $|A \cap B| = |B \cap A|$}

= d4(B, A).

Thus Symmetry is proved.

For (4):

To prove (Triangle inequality) For all A, B, C ϵ S

d4(A, B) <= d4(A, C) + d4(C, B) .

Consider L.H.S

d4(A, B)

= 1 - $\left(1/2 * \frac{|A|}{|A \cap B|} + 1/2 * \frac{|B|}{|A \cap B|}\right)^{-1}$

= 1 - $\left(1/2 * \frac{|A|+|B|}{|A \cap B|}\right)^{-1}$

= 1 - $\left(2 * \frac{|A \cap B|}{|A|+|B|}\right)$

If we consider $|A \cap B| = 0$

= 1 - (0)

= 1

Now consider R.H.S

d4(A, C) + d4(C, B)

= 1 - $\left(1/2 * \frac{|A|}{|A \cap C|} + 1/2 * \frac{|C|}{|A \cap C|}\right)^{-1}$ + 1 - $\left(1/2 * \frac{|C|}{|C \cap B|} + 1/2 * \frac{|B|}{|C \cap B|}\right)^{-1}$

= 1 - $\left(1/2 * \frac{|A|+|C|}{|A \cap C|}\right)^{-1}$ + 1 - $\left(1/2 * \frac{|C|+|B|}{|C \cap B|}\right)^{-1}$

= 1 - $\left(2 * \frac{|A \cap C|}{|A|+|C|}\right)$ + 1 - $\left(2 * \frac{|C \cap B|}{|C|+|B|}\right)$

If we consider $|A \cap C| = 0, |C \cap B| = 0$

= 1 - (0) + 1 - (0)

= 1 + 1

= 2

Thus R.H.S > L.H.S

Hence triangle inequality is proved

As the four rules of metrics for a set is proved d4(A,B) = 1 - $\left(1/2 * \frac{|A|}{|A \cap B|} + 1/2 * \frac{|B|}{|A \cap B|}\right)^{-1}$ is a metric.

e)

d5(A,B) = $(|A-B|^p + |B-A|^p)^{\frac{1}{p}}$, p>=1

To prove it is a metric we verify (1)-(4).

For (1): d5(A, B) =

If $|A \cap B|$ = null, i.e. A and B are independent then |A - B| = |A| & |B - A| = |B|, on solving d5 we get

d5(A,B) = $(|A-B|^p + |B-A|^p)^{\frac{1}{p}}$

= $(|A|^p + |B|^p)^{\frac{1}{p}}$

If we consider A & B to be empty sets then $(|A|^p + |B|^p)^{\frac{1}{p}}$ for all p>=1 is 0

If we consider A or B to be non empty sets then $(|A|^p + |B|^p)^{\frac{1}{p}}$ will always be a positive value as number of elements will always be greater than or equal to 1 hence for all p>=1 it's > 0

Hence positivity is proved.

For (2): To prove d5(A, B) = 0

d5(A,B) = $(|A-B|^p + |B-A|^p)^{\frac{1}{p}}$

If A = B then

= $(|A-A|^p + |A-A|^p)^{\frac{1}{p}}$ for all p>=1

= 0

Hence if and only if A = B d5(A, B) = 0 , thus Non Degenerated is proved.

For (3): Since d5(A,B) = $(|A-B|^p + |B-A|^p)^{\frac{1}{p}}$

This can be rearranged as $(|B-A|^p + |A-B|^p)^{\frac{1}{p}}$ for all p>=1

Thus this can be considered as d5(B,A)

Symmetry is proved.

For (4):

To prove (Triangle inequality) For all A, B, C ϵ S

d5(A, B) <= d5(A, C) + d5(C, B) .

Consider L.H.S

d5(A, B) = $(|A-B|^p + |B-A|^p)^{\frac{1}{p}}$

If |A n B| = null, i.e. A and B are independent then |A - B| = |A| & |B - A| = |B|, on solving d5 we get

d5(A,B) = $(|A-B|^p + |B-A|^p)^{\frac{1}{p}}$

= $(|A|^p + |B|^p)^{\frac{1}{p}}$

here if we consider p = 1

we get $(|A|^1 + |B|^1)^{\frac{1}{1}}$

= |A| + |B|

Consider R.H.S

d5(A, C) + d5(C, B)

$(|A-C|^p + |C-A|^p)^{\frac{1}{p}}$ + $(|C-B|^p + |B-C|^p)^{\frac{1}{p}}$

We can extend the previous defined to rule A and C , B and C and we get

$= (|A|^p + |C|^p)^{\frac{1}{p}} + (|C|^p + |B|^p)^{\frac{1}{p}}$

if we consider p = 1, then

$= (|A|^1 + |C|^1)^{\frac{1}{1}} + (|C|^1 + |B|^1)^{\frac{1}{1}}$

=$(|A| + |B| +2|C|)

we clearly see that even if we consider set C to be null and no of elements i.e. |C| = 0 we get L.H.S = R.H.S —- (a)

as value of p increases and length of C is not 0 but always greater than 0 we get L.H.S as

$\sqrt[p]{|A|^p + |B|^p}$

and the R.H.S could be modified as

$(|A|^p + |C|^p)^{\frac{1}{p}} + (|C|^p + |B|^p)^{\frac{1}{p}}$

$= \sqrt[p]{|A|^p + |C|^p} + \sqrt[p]{|C|^p + |B|^p}$

thus R.H.S > L.H.S when p>1 — (b)

Using (a) and (b) we can conclude that R.H.S >= L.H.S i.e. d5(A, B) <= d5(A, C) + d5(C, B)

As the four rules of metrics for a set is proved d5(A, B) = $(|A - B|^p + |B - A|^p)^{\frac{1}{p}}$ is a metric.

f)

d6(A,B) = $\frac{(|A-B|^p+|B-A|^p)^{\frac{1}{p}}}{|A \cup B|}$, p>=1

To prove it is a metric we verify (1)-(4).

For (1): d6(A,B) = $\frac{(|A-B|^p+|B-A|^p)^{\frac{1}{p}}}{|A \cup B|}$, p>=1

If |A n B| = null, i.e. A and B are independent then |A - B| = |A| & |B - A| = |B|, on solving d5 we get

d6(A,B) = $\frac{(|A-B|^p+|B-A|^p)^{\frac{1}{p}}}{|AuB|}$

= $\frac{(|A|^p+|B|^p)^{\frac{1}{p}}}{|A+B|}$

If we consider A & B to be empty sets then $(|A|^p + |B|^p)^{\frac{1}{p}}$ for all p>=1 is 0

If we consider A or B to be non empty sets then $\frac{(|A|^p+|B|^p)^{\frac{1}{p}}}{|A+B|}$ will always be a positive value as number of elements will always be greater than or equal to 1 hence for all p>=1

i.e if p = 1 we get $\frac{(|A|^1+|B|^1)^{\frac{1}{1}}}{|A+B|}$

= 1

it's > 0 Hence positivity is proved.

For (2): d6(A,B) = $\frac{(|A-B|^p+|B-A|^p)^{\frac{1}{p}}}{|A \cup B|}$, p>=1

If A = B

then

= $\frac{(|A-A|^p+|A-A|^p)^{\frac{1}{p}}}{|A \cup B|}$

= 0

Hence if and only if A = B d6(A, B) = 0 , thus Non Degenerated is proved.

For (3): Since

d6(A,B) = $\frac{(|A-B|^p+|B-A|^p)^{\frac{1}{p}}}{|A \cup B|}$

This can be rearranged as $\frac{(|B-A|^p+|A-B|^p)^{\frac{1}{p}}}{|B \cup A|}$ for all p>=1

As $|A \cup B| = |B \cup A|$ Thus this can be considered as d6(B,A)

Thus Symmetry is proved.

For (4):

To prove (Triangle inequality) For all A, B, C ϵ S

d6(A, B) <= d6(A, C) + d6(C, B) .

Consider L.H.S

d6(A,B) = $\frac{(|A-B|^p+|B-A|^p)^{\frac{1}{p}}}{|A \cup B|}$

If |A n B| = null, i.e. A and B are independent then |A - B| = |A| & |B - A| = |B| and |A u B| = |A| + |B|, on solving d6 we get

d6(A,B) = $\frac{(|A|^p+|B|^p)^{\frac{1}{p}}}{|A|+|B|}$

here if we consider p = 1

we get $\frac{(|A|+|B|)^{\frac{1}{1}}}{|A|+|B|}$

= 1

Consider R.H.S

d6(A, C) + d6(C, B)

$\frac{(|A-C|^p+|C-A|^p)^{\frac{1}{p}}}{|AuC|} + \frac{(|C-B|^p+|B-C|^p)^{\frac{1}{p}}}{|CuB|}$

$= \frac{(|A-C|^p+|C-A|^p)^{\frac{1}{p}}}{|A|+|C|} + \frac{(|C-B|^p+|B-C|^p)^{\frac{1}{p}}}{|C|+|B|}$

We can extend the previous defined to rule A and C , B and C and we get

$= \frac{(|A|^p+|C|^p)^{\frac{1}{p}}}{|A|+|C|} + \frac{(|C|^p+|B|^p)^{\frac{1}{p}}}{|C|+|B|}$

if we consider p = 1, then

= 1 + 1

= 2

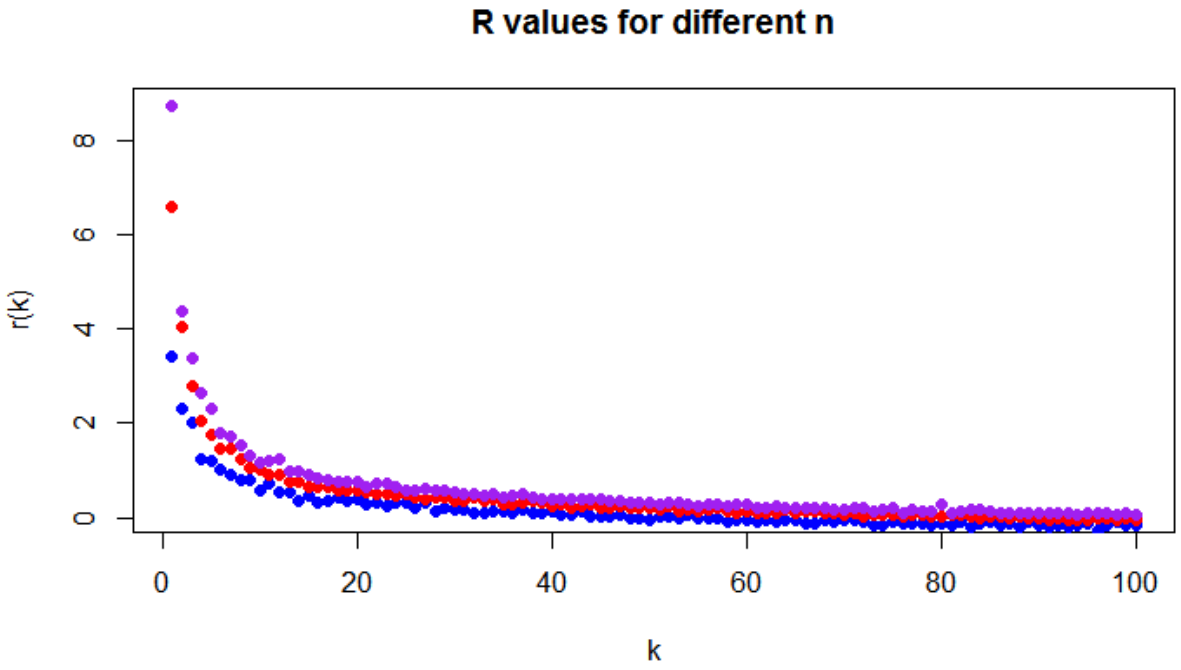
thus R.H.S > L.H.S when p>=1 — (b)

Using (a) and (b) we can conclude that R.H.S >= L.H.S i.e. d6(A, B) <= d6(A, C) + d6(C, B)

As the four rules of metrics for a set is proved d6(A,B) = $\frac{(|A-B|^p+|B-A|^p)^{\frac{1}{p}}}{|A \cup B|}$ is a metric.

Solution 4

a)



b)

Before carrying out the experiment the expectation was that with increase in dimensionality and an increase in the number of data points the data will become sparse and the three curves would appear to be dissimilar, however we see the opposite upon applying the current algorithm. The curve shows that the data points seem to become denser with increase in dimensions and the curves appear more and more simmilar.

Solution 5

a)

The best distance as can be seen from running the code is Lmax distance whose value came out to be 0.54, while that euclidean is 0.58 and Manhattan too

b)

We can use age as the best attribute to calculate metric because the absolute of t-value for age is 6.156 which the highest function of R and thus shows that age is the most significant coefficent so I believe age is the best parameter to incorporate in the distance matrix

c)

Here too LMax outperforms the other distance metrics. The error value increased to 0.84 when compared to euclideans 0.96, and Manhattans 0.95

d)

I believe that we can improve the recommendation system by normalizing the features i.e. by calculating the z score of attributes. $z = (X_(i, j) - /mu)//sigma$.