CSCI B – 565 Data Mining

Abhinandan SampathKumar

Question 1)

Term Frequency, is a measure of frequency of the words in a document. Depending on the size of the documents, the frequency might vary. So this is a function of the number of times word is document (M_{ij}), divided by the total number of words in the document (M_i). It is given by M_{ij}/M_i

Inverse Document Frequency, is measure for significance of words in the document. Although all the word are to be given equal importance, in documents sometimes we have words like - "the", "an" ,"in"; which have high frequency but are of little importance. The words like "Indiana", "Computer Science" can be words which are rare but important, while text mining. So to give more weight to such words we calculate inverse document frequency: $\log(N/N_j)$

A)Advantages of the encoding

- Importance to rare words using the above encoding
- Similarity between two documents I_1 and I_2 is easy to compute. Clustering of similar documents becomes easier
- Search in faster, suppose we are looking for data related to "Indiana" it returns all the documents with the term, mining for a particular Query is easier.

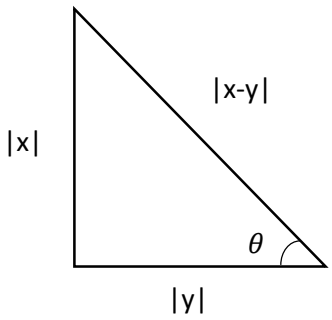
B) Disadvantages of the encoding

- It would cluster the documents based on 'terms', sometimes same terms might occur in different documents with different context. But even though they are in different contexts, the documents are prone to be classified as similar.
- In TF-IDF, if most of the document has a 'term' which is a key term and important for mining. It will not consider the term significant since it occurs in all documents. While achieving rarity, it would leave out a key word

C) Effect of this transformation if a term occurs in one/all documents

- If the term occurs in every document, it would not be considered significant. In this case ($N_j = N$), i.e log1=0, so the term would lose its importance/weight.
- If a term occurs in only one document, N_j would be 1, i.e (LogN/1), the inverse frequency would be high and the term would be very significant.

Question 2)



Cosine similarity can be measured based on the magnitude of the vectors.

$$||\vec{x} - \vec{y}||^2 = ||\vec{x}||^2 + ||\vec{y}||^2 - 2\left|\left|\vec{x}\right|\right|.||\vec{y}|| \cos\theta \quad \text{--------- 1}$$

Since,

$$||\vec{x} - \vec{y}||^2 = \left|\left|\vec{x} - \vec{y}\right|\right|.||\vec{x} - \vec{y}||$$

$$= ||\vec{x}||^2 + \left|\left|\vec{y}\right|\right|^2 - 2\left|\left|\vec{x}\right|\right|.||\vec{y}|| \quad \text{--------- 2}$$

From Equations 1 and 2

$$||\vec{x}||^2 + \left|\left|\vec{y}\right|\right|^2 - 2\left|\left|\vec{x}\right|\right|.||\vec{y}|| \cos\theta = ||\vec{x}||^2 + \left|\left|\vec{y}\right|\right|^2 - 2\left|\left|\vec{x}\right|\right|.||\vec{y}||$$

$$\cos\theta = \frac{\vec{x}\vec{y}}{\left|\left|\vec{x}\right|\right|.||\vec{y}||}$$

Vector Multiplication of x and y is same as multiplication of x^T and y. So x. y=x^T.y

$$=> \cos(x, y) = \frac{x^T y}{\left|\left|x\right|\right| . ||y||}$$

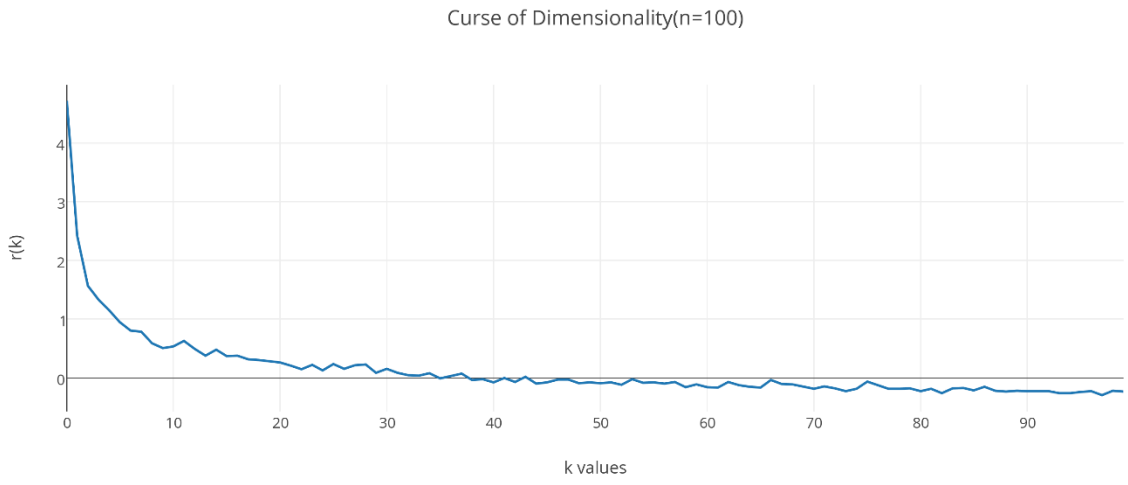
Question 4)

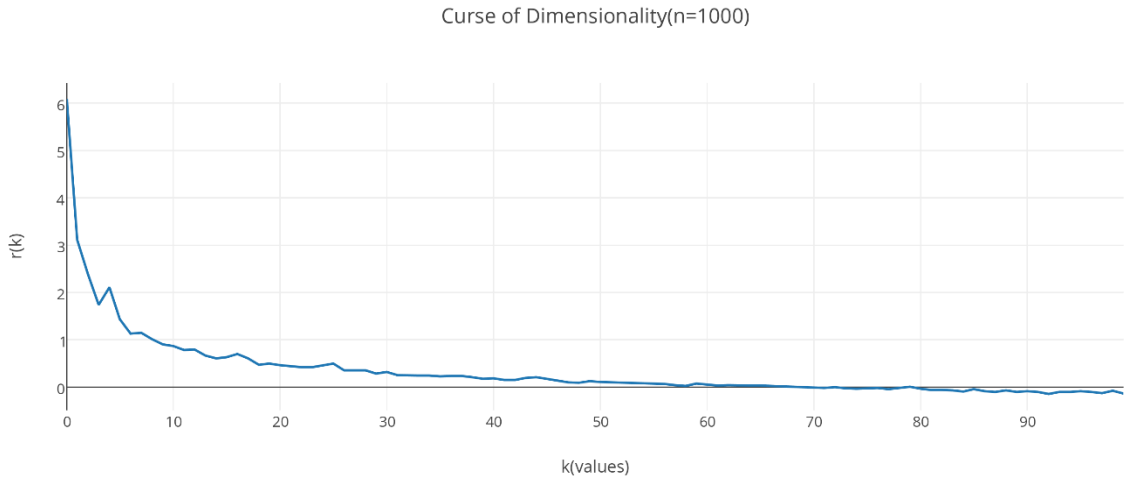
The value of r(k) is inversely proportional to the value of k. With increasing k, the r(k) moves closer towards to O and I also got few negative values for last few values of k. It's also quite clear that distance function is not effective with high dimensions, as the maximum and minimum points are getting closer with increasing dimension.
In higher dimension, computation time increases. Also, Searching/Mining becomes difficult with increasing dimensions. Learning algorithms will suffer greatly in high dimensions.

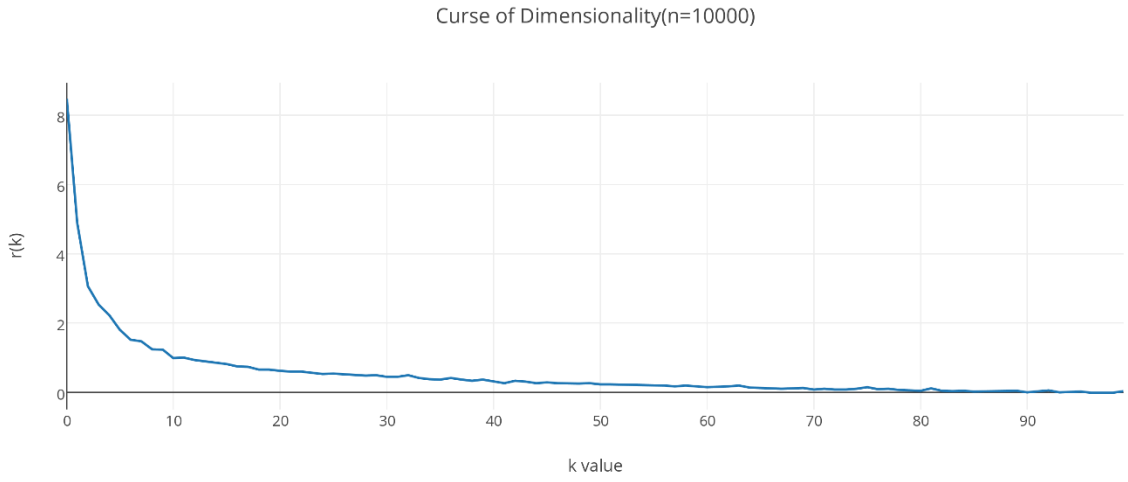
Some of the things observed from the graph and r(k)points were,

1) Highest value of r(k) increases slightly with the increasing dimensions.
2) I did not expect the r(k) values to negative
3) The number of values going to negative decreases with higher value of n.

Below are the attached images of the graphs for Curse of Dimensionality.




Curse of Dimensionality(n=1000)
r(k)
0
1
2
3
4
5
6
0
10
20
30
40
50
60
70
80
90
k(values)


Curse of Dimensionality(n=10000)
r(k)
0
2
4
6
8
0
10
20
30
40
50
60
70
80
90
k value

Question 3)

3. a.

i) $d_1(A,B) = |A-B|+|B-A|$

Considering A=B and Substituting
|A-A| + |A-A| = 0+0 = 0 proves positive Definiteness.

ii) $d_1(A,B) = |A-B|+|B-A|$
$d_1(B,A) = |B-A|+|A-B|$
$d_1(A,B)$= distance (B,A) proves symmetry

iii) Triangle inequality distance

To prove, $d_1(A, B) + d_1(B,C) >= d_1(A,C)$
$D_1(A,B)=|A-B|+|B-A|$

$d_1(B,C)=|B-C|+|C-B|$

$d_1(A,C)=|A-C|+|C-A|$

$d_1(A,B)+d_1(B,C) >= d_1(A,C)$
|A-B|+|B-A| + |B-C|+|C-B| >= |A-C|+|C-A|

Considering LHS,

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

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

Using the above conditions,

|A-B|+|B-C| >= |A-C| and |B-A|+|C-B| >= |C-A|

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

i.e → $d_1(A,B)+d_1(B,C) >= d_1(A,C)$

3. b). $d_2(A,B) = \frac{|A-B|+|B-A|}{|A \cup B|}$

When A=B

$d_2(A,B) = \frac{|A-A|+|A-A|}{|A \cup B|}$ =0

$$d_2(A,B) = \frac{|A-B| + |B-A|}{|A \cup B|} \; and \; d_2(B,A) = \frac{|B-A| + |A-B|}{|A \cup B|} => \; d_2(A,B) = d_2(B,A)$$

We Know,

$$d_2(A,B) = \frac{|A-B| + |B-A|}{|A \cup B|}$$

$$d_2(B,C) = \frac{|B-C| + |C-B|}{|B \cup C|}$$

$$d_2(A,C) = \frac{|A-C| + |C-A|}{|A \cup C|}$$

$d_2(A,C) \leq \; d_2(A,B) + d_2(B,C)$ is satisfied

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

Proving $d_2 \; is \; a \; metric$

3. c) $d_3(A,B) = 1 - (\frac{1}{2}\frac{|A \cap B|}{|A|} + \frac{1}{2}\frac{|A \cap B|}{|B|})$

Let A ={a,b,c} B={a,b,c,d,e,f} C={d,e,f}

|A|=3, |B|=6, |C|=3, $|A \cap B| = 3$

$d_3(A,B) = 1 - (\frac{1}{2}\frac{|A \cap B|}{|A|} + \frac{1}{2}\frac{|A \cap B|}{|B|})$=$1 - (\frac{1}{2}\frac{|A \cap B|}{|A|} + \frac{1}{2}\frac{|A \cap B|}{|B|})$ =$1 - (\frac{1}{2} * \frac{3}{3} + \frac{1}{2} * \frac{3}{6})$ =$1 - \frac{1}{2}\left(1 + \frac{1}{2}\right) = \frac{1}{4}$

$$d_3(B,C) = \; 1 - (\frac{1}{2}\frac{|B \cap C|}{|B|} + \frac{1}{2}\frac{|C \cap B|}{|C|}) = 1 - (\frac{1}{2} * \frac{3}{6} + \frac{1}{2} * \frac{3}{3}) \; = \frac{1}{4}$$

$$d_3(A,C) = \; 1 - \left(\frac{1}{2}\frac{|A \cap C|}{|A|} + \frac{1}{2}\frac{|C \cap A|}{|C|}\right) = 1 - \left(\frac{1}{2} * 0\right) = 1$$

Here,

$d_3(A,C) \geq \; d_3(A,B) + d_3(B,C)$ which is why, d_3 is not a Metric.

3. d)

$$d_4(A,B) = \; 1 - \left(\frac{1}{2}\frac{|A|}{|A \cap B|} + \frac{1}{2}\frac{|B|}{|A \cap B|}\right)^{-1}$$

Let A ={a,b,c} B={a,b,c,d,e,f} C={d,e,f}

|A|=3, |B|=6, |C|=3, $|A \cap B| = 3, |C \cap B| = 3, |A \cap C| = 0$

$$d_4(A,B) = 1 - \left(\frac{1}{2}\frac{|A|}{|A \cap B|} + \frac{1}{2}\frac{|B|}{|A \cap B|}\right)^{-1} = 1 - \frac{1}{2}\left(1 + \frac{1}{2}\right)^{-1} = \frac{2}{3}$$

$$d_4(B,C) = 1 - \left(\frac{1}{2}\frac{|B \cap C|}{|B|} + \frac{1}{2}\frac{|C \cap B|}{|C|}\right)^{-1} = 1 - \frac{1}{2}\left(1 + \frac{1}{2}\right)^{-1} = \frac{2}{3}$$

$$d_4(A,C) = 1 - \left(\frac{1}{2}\frac{|A \cap C|}{|A|} + \frac{1}{2}\frac{|C \cap A|}{|C|}\right)^{-1}$$

$$= 1 - \frac{1}{2}(0)^{-1}\ ; which\ is\ clearly\ not\ defined\ hence, Triangular\ Inequality\ is\ not\ being\ satisfied$$

d_4 isn't a metric.

5.

a)

I ran my algorithm for predicting movie ratings by loading all the data from u1base through u5base and stored it in a userMovieMatrix with ratings as the value. I create a user-user matrix where I calculate distance between users (all 3 distances). I then take k similar users from the matrix for every user based on the similarity. I then take in test Data and predict scores learning from train data for (user, movie) values in test data. Compare these values and then calculate my 'mad' value.

I initially ran for n fold validations and got almost similar values for all folds. I ran all n folds together and it took less computational time with similar results.

Below are 'mad' values for different values of K which changes with distance.
Euclidean was found as the best and Lmax results was the worst.

K= 30
Euclidean -> 0.98924, Lmax ->1.34525, Manhattan -> 1.00271

K = 50
Euclidean -> 0.96913, Manhattan -> 0.9765, Lmax -> 0.99345

K =100
Euclidean -> 0.9572, Manhattan -> 0.9683, Lmax -> 0.99428

K=10
Euclidean = 1.03755, Manhattan -> 1.03426, Lmax -> 1.09234

b)

I took gender of the users and tried to get better results. I tried to give more weightage to users of same gender and find k similar users with this and predict the value better. My results did not improve, may be because of my algorithm.

C)

d) Do sampling, i,e., 30 folds (minimum) of repetitive data, Thats one way to ease the algorithm performance over a large data set.

On each sample, We could run Logistic regression, as the data size is enormously big.

1. here also, we could Divide data into k folds as professor suggested, essentially with out repeating data,

2. Hold out 1 fold as test set and others as trainset.

3. Train and record the test set result.

4. Perform 2 and 3 on each fold in turn as test set

5. Calculate the average and standard deviation of all the fold results.

6. Get the best values of hyper paramter (as a result of Cross validation methodology) and run it over the original test data. Get the accuracy accordingly.

Also,Feature scaling come in hamdy, i.e., Running this algorithm over Selected features like "gender", "age", "Location" can render better results.

References : Used Plot.ly and numpy,scipy websites for plotting graph and numpy array and scipy to calucalte distances.

Discussed with Vivek Patani and Thanmai Bindi for some of the problems regarding how to approach problems.