Data Mining: Homework #1

Due on February 4, 2016 at 11:59pm

Professor Predrag Radivojac CSC 565

Anirudh K M (anikamal)

Problem 1

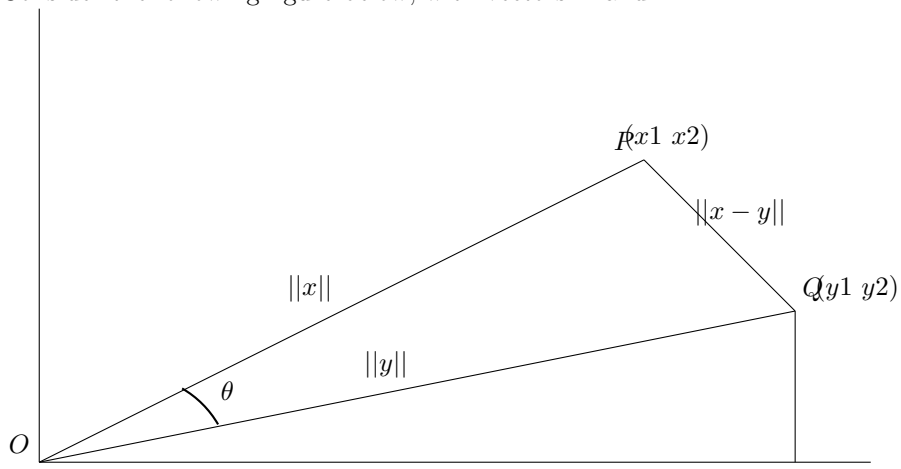
a)The benefits of this coding are as follows
-1 Less weightage is given to words which appear with high frequency in many documents (example : stop words like a,the and and.

-2 More weightage is given to words which has high frequency in few documents.
b) Same weightage would be given to words which appear in large number of doc and less number of docs (as the idf component is omitted)

- Added to it since the document length (Mi) is also not present ,word frequency in bigger documents and smaller documents are treated with out any difference.
More weightage is given to a term which occurs in one document than if it occurs in many documents

Problem 2

Consider the following figure below, with vectors X and Y



Let the angle between X and Y be θ:

$$cos\theta = \frac{x^T y}{||x||\ ||y||}$$

To prove this, we relate to the lengths of the sides of the triangle OPQ.
We know the law of cosine as follows

$$||y - x||^2 = ||x||^2 + ||y||^2 - 2||x||||y||cos\theta$$

When θ is a right angle, we see the above equation follows Pythagorean property:

$$||y - x||^2 = ||x||^2 + ||y||^2$$

for any other angle θ, we see the term $||y - x||^2$ can be represented as $(y - x)^T(y - x)$:

$$y^T y - 2x^T y + x^T x = y^T y + x^T x - 2||x||||y||cos\theta$$

This equation can be simplified as follows

$$x^T y = ||x||||y||cos\theta$$

Problem 2 continued on next page...

So this we have the equation,

$$cos\theta = \frac{x^T y}{||x||\ ||y||}$$

Problem 3

Metric space definition.
A metric space is given by a set X and a distance function d : X ∗ X → ℝ such that

i) **Positivity** For all $x, y \in X$

$$0 \leq d(x, y)$$

ii) **Non-degenerated** For all $x, y \in X$

$$0 = d(x, y) \Leftrightarrow x = y$$

iii) **Symmetry** For all $x, y \in X$

$$d(x, y) = d(y, x)$$

iv) **Triangle inequality** For all $x, y, z \in X$

$$d(x, y) \leq d(x, z) + d(z, y)$$

1. If it is a metric,
   •**Positivity** $d_1(A, B)$ *will be positive since it always returns a positive number*
   • **Non-degenerated** It has becomes zero when A = B, satisfies property 2.
   • **Symmetry** Will satisfy symmetry property, d(A,B) $=|A - B| + |B - A| = |B - A| + |A - B|$ which is equal to d(B,A)

   •**Triangle inequality** : $d_1(A, B) = |A - B| + |B - A| \leq |A - C + C - B| + |B - C + C - A|$

   $$\leq |A - C| + |C - B| + |B - C| + |C - A|$$
   $$\leq d(A, C) + d(B, C)$$

   Hence, $d_1(A, B) is a metric.$

2. • **Positivity** It always returns a positive number.
   • **Non-Degenerate** also it sums to zero only if A=B.
   • **Symmetry** Will satisfy symmetry, $d(A, B) = \frac{(|A-B|+|B-A|)}{|AUB|} = \frac{(|B-A|+|A-B|)}{|BUA|}$ which is equal to d(B,A) since by associative laws, $|AUB| = |BUA|$

   •**Triangle inequality** :

   $$|A - B| = |A| - |A \cap B|$$

   $$|B - A| = |B| - |B \cap A|$$

   So using this in the given function we can get, $d_2(A, B) = 1 - \frac{|A \cap B|}{|AUB|}$
   We know this is the measure of distance based on Jaccard similarity.
   So we can say that $d_2(A, B) \leq d_2(A, C) + d_2(B, C)$
   Hence,$d_2(A, B)$ is a metric

3. If it is a metric,
   •**Positivity and non-degenerated** The distance will always be positive and it is 0 only when A = B.

$$If A = B, then\ d(A,B) = 1 - (\frac{1}{2}.1 + \frac{1}{2}.1) = 0$$

   • **Symmetry** Will satisfy symmetry, $d(A,B) = 1 - (\frac{1}{2}\frac{|B \cap A|}{|B|} + \frac{1}{2}\frac{|B \cap A|}{|A|}) = d(B,A)$

   •To satisfy triangular inequality, we need to prove $d(A,B) =\leq d(A,C) + d(B,C)$

   We know that $|A \cap B| = |A|$ (By set theory property)

   So substituting this in the function gives us

   $d_3(A,B) = 0$

   So we can say that $d_3(A,B) \leq d_3(B,C) + d_3(C,A)$

   So this function follows triangle inequality, So this is also a distance metric.

4. This is similar to the above question, but the inverse is taken. But still it follows the triangle inequality. So this is a distance metrics.

5. $d_5(A,B)$ is also a distance metric because it follows first three properties and also the triangle ineqaulity can be proved using Minkowski inequality.

6. This can also be treated as a distance metric because it follows all properties and triangle inequality can be proved using Minkowski inequality.

Problem 4

1. This program takes the 'n', the data points to be generated as a system argument.
2. Then, with dimensions from 1, 2, 3, 4, ... 100, we generate data points with random data values between 0 and 1 for 5 trails and take the mean of them.
3. Then, we calculate the r(k) value based on the given formula for this.
4. Finally, we plot graph between no of dimensionality and r(k) value.
5. These steps are done for n = 100, 1000, 10000.

From the three graphs we observe that the r(k) is higher for higher 'n' value at the start and decreases as the dimension increases. Also, there is less distortion for higher value of 'n'.
We see that our expectation was when the dimensionality tends to infinity r(k) tends to zero. We also observe this with n = 10000. So we see that our expectation can even match more when we run the script for even higher value of n we can see that r(k) tends infinity for lesser k value.

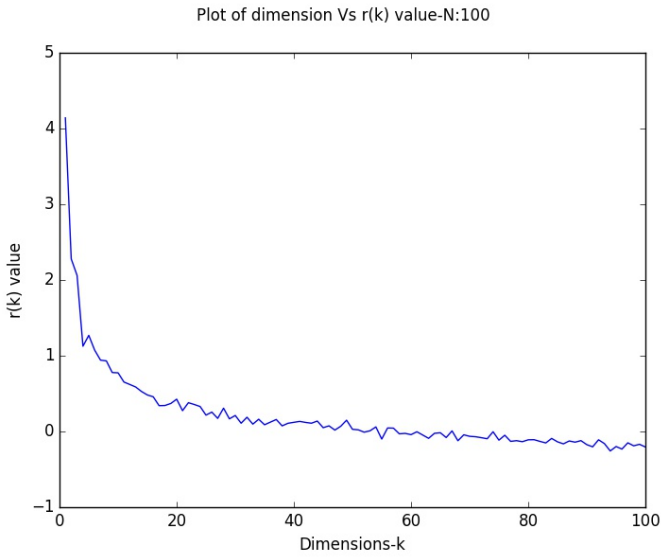


Figure 1: n = 100

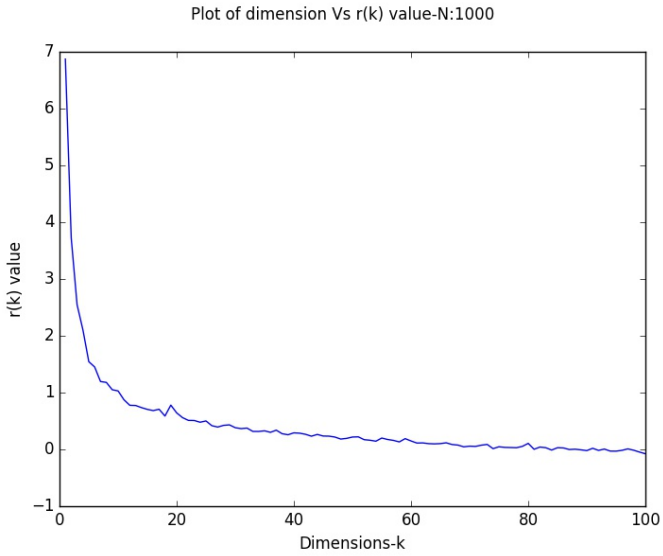


Figure 2: n = 1000

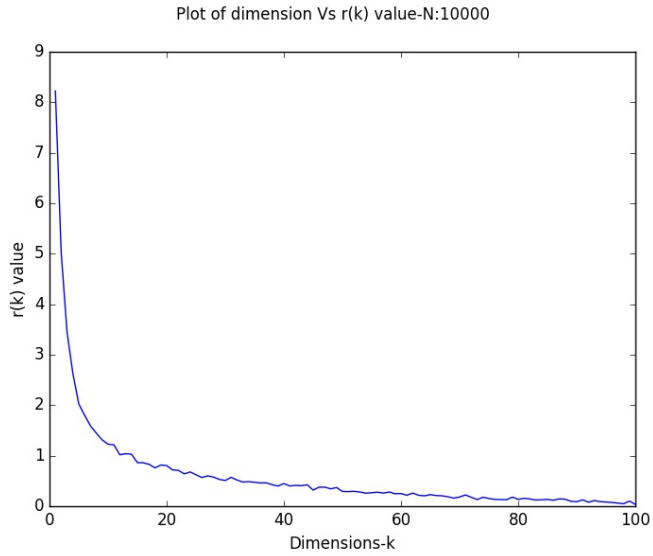


Figure 3: n = 10000

Problem 5

5a)
1. recommender system.py
a) This program is takes the training files from u1.base, u2.base,... u5.base and test files as u1.test, u2.test, ... u5.test and takes the number of similar users to find as a system argument.
b) Then based on the input we get dictionaries based on users and movies.
c) Then we proceed to calculate, the distance between users based on Euclidean, Manhattan and Lmax distances.
d) Then for each distance calculated we find the top k users who saw movie j which was not seen by user i.
e) Based on the distance calculated, we find the similarity and find the users predicted rating based on this.
f) Then, we perform the mean absolute difference with the predicted rating and the user rating from the test data.
2. naive recommender system.py
a) This program is similar to recommender$_s ystem.pyprogram$.
b) But in calculating the user rating we just take the average of the ratings of the person who has seen the movies to assign to the person who has not seen the movie.
c) Then we calculate to the mean absolute difference to find the performance.
The MAD values for various distances is in the below table

K	25	50	75
Euclidean	0.699	0.79	0.79
Manhattan	0.72	0.80	0.80
Lmax	1.22	0.92	0.99

MAD when using the naive algorithm is 0.84

5b) We choose the attributes such as gender, occupation and age. We choose these because these are the data that are available and also that helps us to find the similarity between users.

Problem 5 continued on next page...

a) This program is takes the training files from u1.base, u2.base,... u5.base and test files as u1.test, u2.test, ... u5.test.
b) Then based on the input we get dictionaries based on users and movies.
c) Then we get the attributes of the users such as gender, profession and age.
d) Based on this, we find the similarity between users.
e) Based on this similarity, we predict the rating for the movie 'j' which is not seen by user 'i'.
f) Finally, we predict the mean absolute difference to measure performance.

The MAD when predicted using based on user similarity is 0.78.

5c) When run for the 10M dataset, the above algorithms doesn't work very well because of very high data volume and we need better algorithm like neural networks to predict better user ratings.

5d) The recommendation system can be improved further by getting more attributes like what type of genre a user likes and also who is the favorite actor, director, music director for the user. These type of things helps to find the similarity more efficiently. This can also be improved by using algorithms such as k-nearest neighbours algorithms, neural networks.

References:
1. https://bionicspirit.com/blog/2012/01/16/cosine-similarity-euclidean-distance.html
2. http://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn-note02-2up.pdf
3. https://en.wikipedia.org/wiki/Tf
4. https://en.wikipedia.org/wiki/Metric
5. Probability theory by A. Renyi
6. Linear Algebra and its Applications by Gilbert Strang