Mining massive Datasets WS 2017/18

Problem Set 4

Rudolf Chrispens, Marvin, Daniela Schacherer

November 20, 2017

Exercise 01

We use the formula $cos\phi = \frac{a*b}{||a||*||b||}$ where ϕ is the angle between the vectors a and b. The weighting vector w which is multiplied to a and b before calculating the cosine angle is $\begin{pmatrix} 1 \\ \alpha \\ \beta \end{pmatrix}$.

a) Here we have $\alpha = 1$ and $\beta = 1$. We receive the following cosine angles, which indicate that all three vectors point in almost the same direction:
   - $\phi_{AB} = 0.13°$
   - $\phi_{AC} = 0.17°$
   - $\phi_{BC} = 0.28°$

b) Here we have $\alpha = 0.01$ and $\beta = 0.5$. The weighted vectors are thus

$$\begin{array}{c} \\ PS \\ DS \\ MMS \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 3.06 & 2.68 & 2.92 \\ 5 & 3.2 & 6.4 \\ 3 & 2 & 3 \end{pmatrix} \end{array}$$

   We receive the following cosine angles:
   - $\phi_{AB} = 7.74°$
   - $\phi_{AC} = 7.45°$
   - $\phi_{BC} = 14.26°$

c) If we want to select α and β as the invers proportional of the average in the respective component we receive $\alpha = \frac{1}{\frac{500+320+640}{3}} = \frac{1}{487}$ and $\beta = \frac{1}{\frac{6+4+6}{3}} = \frac{1}{5.34}$.

$$\begin{array}{c} \\ PS \\ DS \\ MMS \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 3.06 & 2.68 & 2.92 \\ 1.03 & 0.66 & 1.31 \\ 1.12 & 0.75 & 1.12 \end{pmatrix} \end{array}$$

With possible rounding errors during the calculation we receive for the angles:

- $\phi_{AB} = 6.01°$
- $\phi_{AC} = 5.25°$
- $\phi_{BC} = 10.67°$

Exercise 02

Consider a web shop that sells furniture and uses a recommendation system. When a new user creates an account and likes one product, he will be presented with similar products on his next visit.

How can a competitor - in principle - try to steal the valuable data for recommendation from this website?

-Write a Bot that likes an Item (or a random set of items) and looks into the recommended list.

-Repeat this process numerous times for every item.

-Using this data gathered, the competitor can recreate his own recommendation data out of the suggestions of the web shop.

Does this work better when the web shop implemented a content- based or a collaborative filtering system?

-It works better in a content- based filtering system, because it uses only the user data of one user.

What data would the competitor be able to infer?

-The competitor could infer what kind of items would be recommended for specific users, because of the gathered data.

-He can create sets of recommendations for all combinations of the current set of data from the web shop.

Would this technique have an impact on the recommendation system, i.e., would this attack create a bias on the data?

-If the bot data of the web shop is treated as a user (not filtered out) then it will bias the data with random not useful data. For example: Like on a Deathpunk band and also Like on Justin Bieber, which would not be "normal" user data.

Why is this attack probably not viable in any case?

-Because of the huge real user Dataset and the modern Algorithms for filtering those kind of attacks would not be relevant, they get filtered out.

Exercise 03

a) The Jaccard distance between two sets C_1 and C_2 is defined as $d(C_1, C_2) = 1 - \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$.
For instance regarding user A and B we need to sum up the elements which have been rated by at least one of the two users (which are all given elements and thus 8) and in a second step count the elements which have been rated by both, user A and B (which is the case for element b, d, e, g and thus 4 elements). We receive: $d(C_A, C_B) = 1 - \frac{4}{8} = 0.5$. We perform this calculation for all pairs of users and get:

- $d(C_A, C_B) = 1 - \frac{4}{8} = 0.5$
- $d(C_A, C_C) = 1 - \frac{4}{8} = 0.5$
- $d(C_B, C_C) = 1 - \frac{4}{8} = 0.5$

b) The cosine distance is defined as: $cos(A,B) = 1 - sim(A,B) = 1 - \frac{A*B}{||A||*||B||}$. For every missing rating of a user a zero is inserted, thus the vector for user A would be $A = (4,5,0,5,1,0,3,2)$. We get the following results:

- $cos(A,B) = 1 - 0.601 = 0.399$
- $cos(A,C) = 1 - 0.615 = 0.385$
- $cos(B,C) = 1 - 0.514 = 0.486$

c) If we treat 3,4 and 5 as 1 (interpretation True) and 1,2, and blank as 0 (interpretation False) we receive the following values for the Jaccard distance:

- $d(C_A, C_B) = 1 - \frac{2}{5} = 0.6$
- $d(C_A, C_C) = 1 - \frac{2}{6} = 0.67$
- $d(C_B, C_C) = 1 - \frac{1}{6} = 0.834$

d) If we do the same as in c) we receive for the cosine distances:

- $cos(A,B) = 1 - \frac{2}{2*\sqrt{2}} = 1 - 0.707 = 0.293$
- $cos(A,C) = 1 - \frac{2}{2*2} = 1 - 0.5 = 0.5$
- $cos(B,C) = 1 - \frac{1}{\sqrt{2}*2} = 1 - 0.354 = 0.646$

e) The average user vales are: $avg_A = \frac{4+5+5+1+3+2}{6} = 3.34, avg_B = 2.34, avg_C = 3$. Normalizing the matrix with those averages one receives:

$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{cccccccc} a & b & c & d & e & f & g & h \end{array} \\ \begin{pmatrix} 0.66 & 1.66 & - & 1.66 & -2.34 & - & -0.34 & -1.34 \\ - & 0.66 & 1.66 & 0.66 & -1.34 & -0.34 & -1.34 & - \\ -1 & - & -2 & 3 & - & 1 & 2 & 0 \end{pmatrix} \end{array}$$

f) With the matrix calculated in part e) we finally end up with the following cosine distances:

- $cos(A,B) = 0.42$
- $cos(A,C) = 1.12$
- $cos(B,C) = 1.74$

Exercise 04

Alternatives for numerical ratings are symbols that show how good the connection is or maybe strings. The problem with strings is that someone has to determine its value, no matter if it is a KI or a human.

(a)

In this case the users are the students and the professors constitute as items. The rating is put together out of the different experience. For example: a student gives a rating in every category and the connection (rating) is the mean (maybe some categorys are weighted more than others) out of all ratings. These values are used to build the matrix.

(b)

The users in this context are obviously the users of the online community. Since every user can upload more than one picture, users have to rate the pictures and not other users. That's why the items are the pictures of artworks and not users. I think the best rating for this problem is a star rating. Therefore the matrix will consist of calues from 1 to 5.

(c)

Both the users and items are the users that use this dating platform. The special thing about this scenario is, that we have to pay attention on values for blocked users and that we have to check the "dream partner" of an user. Send messages don't need to be observed because it means nothing if one user send a message. Maybe he is bored or maybe it's a bot, so it's safer to ignore the messages. So if we want to suggest one user to another the value should be negative if this user is blocked and shuold be very high if the user is close to the "dream partner". Liked users just have a constant value.

Exercise 05

see python code attached. (U4_Ex5.py)