10 Recommender System

Exercise 10.1 (Similarity between Users)
Let $v_{i,j}$ the vote of user i for item j. We suppose that there are n users and m items. Let I_i the sets of items for which user i has voted. Let $v_i = (v_{i,1}, \dots, v_{i,m})$ the vector of votes of user i with the convention $v_{i,j} = 0$ for all $j \notin I_i$.

1. Let us consider two users a and b. Let $\mathcal{P}$ be the plane spanned by the two vectors v_a and v_b. The origin of this plane is the origin $O = (0, \dots, 0)$. The two vectors are starting from the origin. The end of vecteur v_a is the point A and the end of vecteur v_b is the point B. Draw the triangle OAB. Using the law of cosines, what is the definition of the angle α between the edges $[OA]$ and $[OB]$ of the triangle OAB ? You can consult https ://en.wikipedia.org/wiki/Law_of_cosines
2. Express $\cos\alpha$ in function of $\|v_a - v_b\|^2$, $\|v_a\|^2$ and $\|v_b\|^2$ where $\|v\|$ is the Euclidean norm of v.
3. Show that $\cos\alpha = \frac{v_a \cdot v_b}{\|v_a\|\|v_b\|}$ where $v_a \cdot v_b$ is the dot product between the vectors v_a and v_b. The angle between the two vectors v_a and v_b is defined as the angle α.
4. The cosine of the angle between two vectors is a distance or a similarity measure ?
5. Recall the definition of the correlation coefficient between two vectors. Is it a distance or a similarity measure ?
6. Let $\varepsilon > 0$. We say that a vector v is a ε-neighbor of v_a if $\|v_a - v\| \leq \varepsilon$. The ε-neighborhood measure $m(v_a, v_b)$ is defined as $m(v_a, v_b) = 1$ is v_a and v_b are some ε-neighbors, and $m(v_a, v_b) = 0$ otherwise. Is it a distance or a similarity measure ?
7. For the following vectors v_a and v_b, calculate the ε-neighborhood ($\varepsilon = 2$), the cosine and the correlation measures.
   (a) $v_a = (1, 1, 1, 1)$ and $v_b = (2, 2, 2, 2)$,
   (b) $v_a = (0, 1, 0, 1)$ and $v_b = (1, 0, 1, 0)$,
   (c) $v_a = (0, -1, 0, 1)$ and $v_b = (1, 0, -1, 0)$.
8. Comment the interest of each measure (cosine, correlation and ε-neighborhood).

Exercise 10.2 (Artificial Data)

1. Create a set of $n = 10$ integers randomly choosen in $\{1, 2, 3\}$. Each of these integers will represent the user type $t_i \in \{1, 2, 3\}$ of user i for $i \in \{1, 2 \dots, n\}$.
2. Create a common rating vector $c_t \in \mathbb{R}^4$ for each user type $t \in \{1, 2, 3\}$. The common rating vector is a vector of $m = 4$ independent realizations of a normal random variable with mean 50 and standard-deviation 10.
3. For user i, create a rating vector $v_i = c_{t_i} + \delta_i$ which is the sum of the common rating vector c_{t_i} (corresponding to the user type t_i) plus a personalized rating deviation δ_i which is a normal random vector of m independent realizations of a normal random variable with mean 0 and standard-deviation 1. You must arrange all these rating vectors in a matrix of size $n \times m$. Rows can be interpreted as "users" and columns as "items".
4. Remove randomly a proportion p of values out of the matrix. You can use a Bernoulli random generator with probability of success p to choose if a value must be kept or remove. The erased values will be replaced by "NA".

5. Code the memory-based algorithm with the Pearson correlation to predict the missing values. The correlation is computed between the users.
6. Compute the Mean Absolute Error (MAE).
7. Study the MAE as a function of p for $p \in \{0.001, 0.01, 0.1, 0.15, 0.2\}$. You should run several times the full simulation testing all the values of p. Discuss the results. You can increase n or m if necessary.

Exercise 10.3 (Real Data)
This exercise is based on the real data sets : “train_v2.csv” for training and “test_v2.csv” for testing. The testing data set is not explicitly used but, if you are motivated to test your algorithm in real conditions, you can exploit the testing data set.

The data fields are :

— id : an anonymous id unique to a (movie, user) tuple
— user : the id of a user
— movie : the id of a movie
— rating : the rating of a movie by a user.

1. Develop a code to load the training data set.
2. Apply the memory-based algorithm to the training data set (as made in Exercise 2).
3. Compute the MAE and discuss the results.