Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 03
Lecture - 11

Lecture 11: Distance and Similarity

Hello everyone. Welcome to the module 3 where we will be talking about, we are actually starting now the algorithms for collaborative filtering. And in this particular lecture, we will be talking about distance and similarity measure which forms the foundation of neighborhood based collaborative filtering. So, before we talk about the algorithm for neighborhood based collaborative filtering, we will be talking first on distance and similarity measures.

So, these are the concepts distance and similarity matrix we are going to cover. To start with, let us look at collaborative filtering. Collaborative filtering has two major approaches neighborhood based and model based. These neighborhood based methods are also called memory based methods because you have to actually remember the items. So, in case of neighborhood based user, the only input that you get is user item rating matrix. To remind you, we started with discussion on recommender system with three matrices user matrix, item matrix and user item feedback matrix or the rating matrix.

This is also called utility matrix. This is the only matrix which is input to the neighborhood based method. Now, again in neighborhood based method, two major approaches are user user based collaborative filtering and item item based collaborative filtering. In case of user user based collaborative filtering, you find out the distance among the the similarity among the users and you group them together based on the similarity and depending on in a particular group how the items are rated by other users. For a specific user, if that item is not rated, the rating value is derived from the rating pattern of the similar users or the likeminded users.

In case of item item based, again looking at the ranking, we identify the similarity in ranking of the items and we form the groups of the item. Based on that, we can assign we can assign the rating value or predict the rating value to a specific user item pair where the rating is already not given and based on that rating prediction, we can give top n recommendation. So, this distance and similarity are the basis for this user and item based collaborative filtering. Besides this collaborative filtering, though we are talking about distance and similarity measures in the context of collaborative filtering where only rating matrix is the input, in other places also distance and similarity measures you will come across. In fact, while discussing about let us say KNN classifier or let us say K-means algorithm, we have already talked about the distance measure.

So, now let us formally look at this. What is similarity? It is the numerical measure of how alike two data objects are. So, when we talk about data objects, a data object typically will have more than one dimension, two or more. So, those dimensions are called features. So, based on these features, you are supposed to find out the similarity.

So, now when the two objects are very similar, the similarity function which you will be using will be giving a very high value. In fact, while talking about the statistical foundations, we saw correlation at least between two variables. So, correlation was a measure which was giving a value between minus 1 to plus 1. And similarly, for most of the similarity functions, it can be between minus 1 to plus 1. For some, it is also 0 to 1.

You can find out the similarity using and the range can be 0 to 1, but it has a limit either 0 to 1 or minus 1 to 1. Dissimilarity which is also can be called as the distance because higher the distance, more dissimilar the object are, objects are. Lower the distance, more similar the objects are. So, here also you have a set of numerical measures of how different two objects are and we can relate this dissimilarity which is distance to similarity as well. So, instead of similarity, we can also use distance as a criteria.

Let us say with certain function, let us say we can normalize the distance to 0 to 1 and ake the, take it 1 minus that can give you the similarity value. But distance is always a positive quantity and the minimum value is 0 and upper limit can be anything. So, this is the typical property of a distance function. If P and Q are two data points, when I say data points, they can have multiple dimensions. Then distance between P and Q equal to 0 only if P is same or exactly similar to that of Q.

So, this particular property is called positive definiteness. The next property which is symmetry tells that the distance, if we, distance is a function. So, if that the function, the input is P and Q, the first point is P, second point is Q, the distance between Q and P is also going to be same. So, this particular property is called the symmetry. And the third property is called triangle inequality.

So, if you have three points P, Q and R, if you take the distance between P and Q and the distance between Q and R and add them together, this has to have higher or at least equal value to the distance between P and R. So, with these three definitions, we can say that any function which satisfies this above three properties can be called a distance function. So, which means not necessarily Euclidean distance is the only distance function. So, if we have a data matrix where the objects are represented as with characterized with P number of features and there are n number of such elements, then we have to, when we construct the similarity or distance matrix, we have to compare all these n points with each other. So, consequently the distance or similarity matrix will be of dimension n cross n.

So, if the original data matrix is of dimension n cross P, where n is the number of objects, the distance matrix is going to be n cross n, where n is the number of the objects. So, with

each object we have to find out the distance. Now, because this distance function is symmetric, so therefore, the matrix also turns out to be symmetric and because it is symmetric, we can also represent it as a triangular matrix, either upper triangular or lower triangular. So, in fact, when we make it as a full matrix, it becomes a symmetric matrix. So here, let me remind you about the kind of numerical data we deal with, as a measurement scale we deal with.

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

So, if you remember the measurement scale, we had four kinds of measurements. So what are they? Nominal, then we have ordinal, then we have interval and ratio scale. So this last interval and ratio scale were numeric. The first two were actually qualitative, but we were representing them in terms of numeric values. So for example, in this example, there are four students who are getting grades in two subjects, EXA, ABCB, EXA and so on.

So now, this grade, which is a qualitative thing, we have to represent it in quantitative form. So, if you represent this in a quantitative form, how do you do it? You assign certain scale, a nominal scale, let us say EX will be 1, A will be 2, C will be 3 and for the second object, we have only two categories. So, 1 will be 1, sorry, A will be 1 and B will be 2 and so on. So whatever may be the case, the point that I am trying to make is, whenever you have nominal variables like this, you simply cannot find out Euclidean distance by using these numbers that I told you, because the numbers in this case are actually symbolizes these grades. They are truly not number and cannot be added, subtracted or no mathematical operations can be done on them.

So, the first measure, if we have nominal attributes, the first measure which is very simple is based on the matching distance concept. So, in this simple matching distance concept, we simply match the items. Let us say, if we are comparing student 1 and student 2, if we are comparing student 1 and student 2, student 1, student 2, first subject EX, second A, this is A, this is B. So neither in this attribute nor on this attribute there is any match. So P is the

total number of attributes that is 2 and how many matches has happened? 0 matches have happened divided by P that makes it 1.

So in the distance matrix, we can compute between every pairs and we can see that except for 2 and 3, rest are quite, the distance is high. Now let us look at 2 and 3, why the distance is low? Because here at least in one of the subjects, they have the same attribute value. So here it is AC, here it is BB. So that is how they are little bit similar. And if you look at 1 and 4, they have got exactly same grade.

So in that case, 4 and 1, whatever was happening to 1 and 4, 4 and 1 is the same. So they are very similar. So distance is 0. They are exactly same, distance is 0. As more dissimilarities happen, higher is the value.

- If object attributes are all nominal (categorical), then proximity measure are used to compare objects
- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Method 1: Simple matching distance
  - *m*: # of matches, *p*: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

Data matrix

Student	Sub 1	Sub 2
1	Ex	A
2	A	B
3	C	B
4	Ex	A

$$d(1,2) = \frac{2-0}{2} = 1$$

Distance matrix

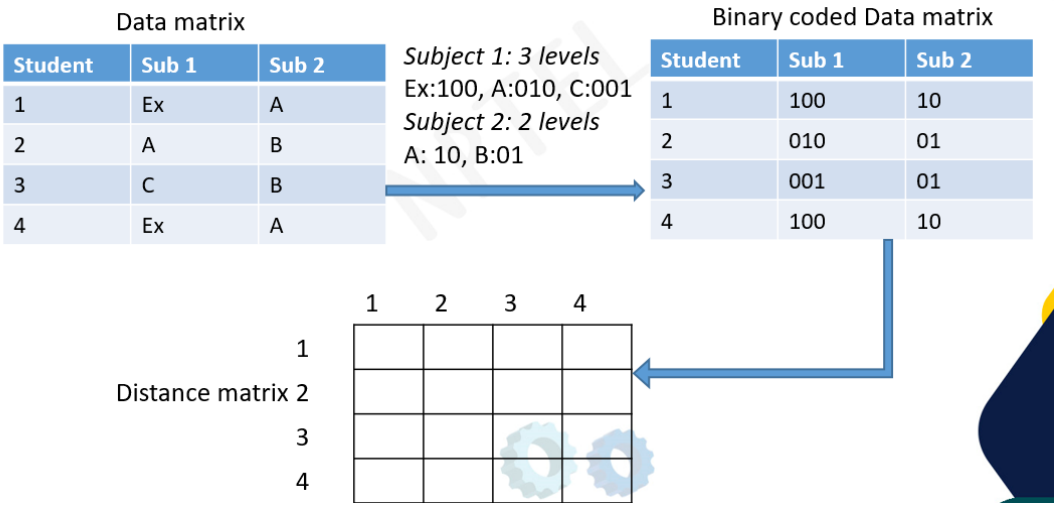
	1	2	3	4
1	0			
2	1	0		
3	1	0.5	0	
4	0	1	1	0

So this is the distance matrix. Now come to the second approach for doing this. So this we can compute using something called binary formatting. Here let me tell you one more thing. The attributes when we consider, in a particular let us say here, we are talking about student 2 subjects.

Student can have other attributes. Let us say he is having age as one of the attribute. Then he is having let us say his attendance number of classes he has attended in subject 1 and subject 2. So those are numeric values. So if we have this combination of numeric as well as some kind of categorical variable, then how exactly we deal with while finding the distance.

So while finding the distance, definitely because there are certain numeric attributes, some numerical measures like that of our Euclidean distance we may like to use. But in that case we need the values to be all numeric. So how do, now the question is how do we convert these nominal variables to numeric form. To know that how to convert them into numeric form, just look at this. Here this process is called binary formatting.

So in the binary formation what do you do? Look at this subject 1. Subject 1 has 3 different symbols. So what are those symbols? X, A and C. So we represent it using 3 binary values. So let us say to EXV give 100 to A, the middle one becomes 1.



For C, the last one becomes 1. So had there been 4, what we would have done? The first one would have been the first one would have been 1. In second category, second point would have been 1. And in the third, the third position would have been 1 and so on. So this is how the subject 1 is represented. And likewise in subject 2 we have only 2 symbols.

So therefore it is represented in this way. Now once you have this, this also becomes numeric values now. Now along with let us say age, number of attendance, etc., we can put together and make a vector. Now based on this vector, we can now find out the distance.

But still the distance matrix is going to be 4 x 4 and symmetric. Now come to the binary attributes which is a special type of nominal attribute only. Now in case of binary attribute, we will be distinguishing binary attributes in 2 types. In fact, I have already told you in one of the earlier lectures if you remember that this binary attribute is of 2 types. Either it is symmetric or it is asymmetric.

In case of symmetric binary variable, both 0 and 1 are important. When I say both 0 and 1 are important, which means when you give 0, it essentially says absence of something. But in case of asymmetric, if you give 0, it is does not necessarily say absence of certain feature. For example, which we have already said, in case of YouTube, you have the option of giving your rating as with thumbs up or thumbs down. So if it is thumbs up, let us say we consider it as 1.

If it is thumbs down, we consider it as 0. So which means the person has actually explicitly stated that he does not like that particular item, particular video. But in case of Facebook, you have the option for only thumbs up. So which means if you give thumbs up, you are liking the item but not giving, you do not have any option for giving thumbs down. So which means if no opinion is given, then it is not necessary that the person is not liking it. Only thing is that he has not seen it probably.

Okay? So under such situations, whether it is symmetric or asymmetric, the first task is to make something called a contingency table for the binary data. When we make this

contingency table, we consider two objects, object 1 and object 2. Object 1 will have many binary variable values. This is object 1 and your second object 2. So it will have values like 0, 1, 1, 0, 0, 0, 1 and this one will have values, let us say 1, 1, 0, 0, 1, 1, 0.

Object *i* \ Object *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

$$d(i, j) = \frac{r + s}{q + r + s}$$

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

Okay? Okay. So now, while making this contingency table, we have to see how many places, how many places both of them have given the value 1. So how many such places? This is one place. This is one place only. Okay? So value in this particular example, q is 1. How many places where object 1 is object, let us say this is object 1 and this is object 2, object i is object 1, this is object 2 or here we can make i and j itself.

Okay? So in how many places when object i is, object i is giving 1, object 2, object j is giving 0. So object i is giving 1, so these two are places, three places and object i has given 0 in, out of that in two places, object i has given 0. Object 1 has given 0, given 1 and object 2 has given 0. Object 1 has given 1, object i has given 1, object j has given 0.

So how many such values? You have two such values. Similarly, you can find out wherever this person has given 0, the other person has also given 1. So this person has given 0, this person 1. 0, 1, 0, 1, so three such points. Okay? Then both will be giving 0.

Both are giving 0 in one, in one instance only. Okay? So, 1, 2, 3, 4, 5, 6, 7. So for all the seven attributes, your p is equal to 7, all the seven attributes. Okay? So, you find out this row and column wise sum, now you find out the distance measure. So for symmetric binary variable, this is the formula. R plus s, what is r? R is the, r and s are the places where they differ and this is total p.

The places where they differ and total p. And in case of asymmetric variable, the places where they differ, but you have to mark here, t is not included. t is not included. Why t is

not included? Because both of them have given 0 and 0 in case of asymmetric setting does not necessarily say the absence of the feature. So we exclude that. Okay? Now, there is another measure called jacquard coefficient, which is the similarity measure for asymmetric binary variable.

Now this jacquard coefficient is 1, is q by q plus r plus s. Now what is q? What is q? q is where all both of them agree, that is this one, both of them agree divided by total number of writing. Now see, once again we have now excluded t. Now come to the numeric data. In numeric data, one very well known distance measure is your Euclidean distance.

$$X = \langle x_1, x_2, \cdots, x_n \rangle \qquad Y = \langle y_1, y_2, \cdots, y_n \rangle$$

$$dist(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^h \right)^{1/h}$$

$$dist(X,Y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$

$$dist(X,Y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

$$dist(X,Y) = 1 - sim(X,Y) \qquad sim(X,Y) = \frac{\sum_i (x_i \times y_i)}{\sqrt{\sum_i x_i^2 \times \sum_i y_i^2}}$$

This Euclidean distance is also called the L 2 norm. In fact, about this L 2 norm, L 1 norm, etcetera, we discussed while talking about, talking about minimizing the error function. And in what context? In the context of supervised learning. Our error function, that mean square error was actually a function which was like Euclidean distance. And we also talked about L 1 norm in some context where we were trying to give some penalty to the mean square error.

But anyway you refer to those lectures for more detail. But anyway now this common distance measures are based on norms. So the first one is actually Minkowski distance which is called LH norm. So if we have let us say two vectors x 1, x and y with elements, n number of elements x 1 to x n and y 1 to i y n, then we can find out the Minkowski distance by taking the mod of the difference between. So positive value of that to the power h, sum it, sum them up, take root over of hth root of that, hth root of that.

In case of L 1 norm, h becomes 1. So essentially you sum them up without raising the power to any other value. In case of Euclidean distance, it is again the same thing. Because it is square, so taking the mod is same as that of making the square. Because it is anyway

going to be positive and take the root over. So here your h is equal to 1, h is equal to 2 and likewise you can continue and you can have h equal to 3 and so on.

So any other for any other h, the formula is that of Minkowski distance. So as the h gets larger, only the dimensions with the largest difference matters. So formally you also can have something called L infinity norm which is defined as the maximum of x i minus y i. What is i? i is from 1 to n, this one, i is from 1 to n. Now as I told you, the distance is actually just the opposite of similarity.

What do you mean by opposite of? It is just dual of similarity. What do you mean by dual of similarity? If something is very similar, the distance is going to be very less and similarity is going to be very high. So you can also compute distance function from the similarity values. Let us say this is one similarity function of which we are going to study next and this is the distance function. So 1 minus this similarity value can give you some quantity which you can term as a distance function. However, you have to make sure that the value that you are getting lies between 0 to 1 because of one property of the distance function because it is that property is called positive definiteness.

Now with this, we wind up this lecture. These are some of the references and let us try finding the concluding steps on this particular lecture. So we understood that distance and similarity approaches are very important concept in recommender system. Specifically, we saw that in case of collaborative filtering, while finding the user-user and item-item based similarity based on the rating matrix, we can use the similarity concepts. Now the objects are more similar when distance is less.

So these concepts are related to each other in a dual manner. Now choice of distance and similarity function depends on the nature of the data and in particular, we discussed about the binary variable, the nominal variable as such and then binary variable. And we saw that nominal variable can also be represented in the form of binary variable using some concept called binary formation. So you can form a binary string out of the categorical or the nominal data. And we also saw if the data and once we make this binary formation and if we have a multivariable setting along with these binary values, if we have certain numeric values taking all of them together, we can consider them as a as the numeric data and using various distance functions, we can find out the distances. And coming to distance functions, we saw that using I mean the most familiar of this is actually Euclidean distance, but in fact Euclidean distance is a more specific form of a generic distance concept called Minkowski distance and this L1 norm, L2 norm which is Euclidean distance and L infinity norm, all of them can be treated as distance functions. Thank you.