

Data Analysis and Decision Making - II
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture - 33
Distance Measures, Normalization

Welcome back my dear friends and dear students; a very good morning, good afternoon, and good evening to all of you wherever you are in this part of this world whether in India or abroad. And, as you know this is the DADM which is Data Analysis and Decision Making - II course on lecture series under the NPTEL MOOC lecture and this total course, this course total number of weeks is 12 which is 30 hours, 30 hours being expanded into 60 lectures because each lecture is a for half an hour and each week we have 5 lectures after each week we have assignments.

So, as you can see we are in the 33rd lecture which is basically the 7th week going on, 2 weeks I have already where 2 classes I have already been completed for the 7th week and my good name is Raghu Nandan Sengupta from IME department IIT Kanpur. So, if you remember we were discussing TOPSIS method and the concept of outranking was coming up in the discussion that how would you outrank decision 1 with respect to decision 2 or alternative 1 with respect to alternative 2.

So, in case that is there you will basically consider the criterias also to make the decision. And when you are doing that you will basically have the concept that there is an ideal solution and you will try to find out the distance positive and a negative distance that will come later on, the concept of distance based on the ideal solution how close or how far it is. So, if it is positive ideal solution which is which you want, so obviously will try it to be as close as possible to the positive ideal solution.

Hence, you will give positive benefit to your decision in that case if i and j , i and j being basically the nomenclature trying to basically mark the alternatives 1 to m and if i and j are being compared alternatives and for any criteria and if it is positive, so obviously, we will give the distances is as close as possible to the positive ideal solution. So, it will be positive benefit if it is closer to the negative ideal solution it will be negative. Similarly if it is farther away from the positive ideal solution it would basically lose its importance that alternative with respect to the other which you are trying to compare. And, if it is

farther away from the negative ideal solution obviously in that case the negativity would start decreasing going towards positivity.

So, we are, as I said we would not consider too close to a positive solution means too far from the negative solution or too close to the negative solution it means that it is too far from the positive solution we would not consider that; in the sense that asymmetry of the decision would be maintained. So, I was discussing the pseudo code based on which we will solve a problem slowly. So, in the last class which is in the 32nd class, I did start, I hardly spend about 1 minute, 2 minute for this algorithm and then I said that I will stop it here because the half an hour was almost over.

So, I said I again take up this algorithm or based on which the procedure based on which will work and try to solve the problem. Now the algorithms can be changed in the sense general flow process would be the same, but the methodology or at each and every step the decision or see for example, the utility function, normalization all these concept may change; so that would depend on the decision maker. As I told that this utility function is important. So, considering those are changing, but we will follow the general structure of the algorithm accordingly.

(Refer Slide Time: 04:34)

Algorithm for TOPSIS (contd...)

1. **DEFINE:** $X_{m \times n}$ (matrix consisting of priority scores assigned to decisions/alternatives A_i based on attributes/decision criteria/goals C_j, w_j) (weight for the attributes/decision criteria/goals) such that $\sum_{j=1}^n w_j = 1, B$ (benefit matrix), C (cost matrix) $r_{ij} = \frac{w_j x_{ij}}{\sum_{i=1}^m w_j x_{ij}}$, $AIS = (r_1^+, \dots, r_n^+)$ = $\left[\begin{matrix} \max_j(r_{1j}) \in B \\ \min_j(r_{1j}) \in C \end{matrix} \right]$ (negative ideal solution), $PIS = (r_1^-, \dots, r_n^-) = \left[\begin{matrix} \min_j(r_{1j}) \in B \\ \max_j(r_{1j}) \in C \end{matrix} \right]$ (positive ideal solution), $d_i^+ = \sqrt{\sum_{j=1}^n (r_{ij} - r_j^+)^2}$, $d_i^- = \sqrt{\sum_{j=1}^n (r_{ij} - r_j^-)^2}$, $R_i = \left(\frac{d_i^-}{d_i^+ + d_i^-} \right)$ (relative closeness), $M = (R_1^+, R_1^-)$ (separation measure), here $i = 1, \dots, m$ and $j = 1, \dots, n$.

2. **INPUT:** $X_{m \times n}$ (matrix consisting of priority scores assigned to decisions/alternatives A_i based on attributes/decision criteria/goals C_j, w_j) (weight for the attributes/decision criteria/goals) such that $\sum_{j=1}^n w_j = 1, B$ (benefit matrix), C (cost matrix), here $i = 1, \dots, m$ and $j = 1, \dots, n$.

3. **START** $i = 1$ to m .

4. **START** $j = 1$ to n .

5. **CALCULATE:** $r_{ij} = \frac{w_j x_{ij}}{\sum_{i=1}^m w_j x_{ij}} = w_j / x_{ij}$ where $i = 1, \dots, m$ and $j = 1, \dots, n$.

6. **END** j .

7. **END** i .

8. **CALCULATE:** $r_1^+, r_1^-, AIS, PIS, d_i^+, d_i^-, R_i$.

9. **REPORT:** AIS, PIS, M, R_i .

10. **END**.

So, as I said you will basically have X matrix which is m cross n; m is the number of alternatives, n is the number of decision the criteria and this X which matrix would basically be each cell would give you the corresponding values which accrues to that

alternative based on the criterias. So, cell see for example, A 1 1 would basically be the corresponding value which is coming out from criteria 1 to alternative 1. Similarly cell A 3 1 would be the corresponding value which is accruing or being given by criteria 1 2 alternative 3 and so on and so forth.

Apart from that and these and A's would be the corresponding alternatives and small c's would basically with the or capital C they are not the concept of criteria. So, c's would basically my mistake c's would definitely be the capital C would definitely be the criteria because that c what I am referring here is not the c what we have referred in the electro process remember that. And we will have basically a matrix W where W 1 1 then W 2 2, W 3 3, W 4 4 so on and so forth which is the principal diagonal with the values which will be assigned to of the weights based on the fact that what weightages or what importance you will give to the criterias in making the distance.

We should also ensure I am still in the line 1, we should also ensure that the sum of the weight should be 1 and we will basically find out the normalization based on the utility function. Now remember one thing I am mentioning it time and again so, please bare with me. The utility function which will be utilizing would be constant for that set of procedure which you are doing, point 1. We will also try to maintain that if the person is going to take decision stage by stage, obviously the utility function should not change. Point number 4 is that when you are trying to normalize, use the concept of normal normalizing along the row or along the column so as that you maintain the parity from step from step 1 to step 2, step 2 to step 3 so on and so forth.

So, once you find out say for example, the normalized matrix you will basically multiply the normalized matrix with the weights to get the so called normalized weighted values which are the percentage wise so called utility, normalized utility which you will be getting for alternative 1 with respect to criteria 1 or alternative 3 with respect to criteria 4 so on and so forth. Because, the size of the matrix would not change what I am saying that when you are trying to find out y, y is equal to basically X into W. Here X is a normalized vector. The size of X is m cross n, size of W is n cross n.

So, m cross n multiplied by n cross n gives you the final matrix of y as m cross n, m as in Mangalore, n as in Nagpur which means that I have basically per for each of the columns, I have the criterias and corresponding to each of the rows I have the basically

the alternatives. So each cells would give you the overall normalized weighted values which accrues from each criteria to each of this alternatives. Once you find that you will basically find out the so called positive distance function and negative distance function. And find out for what sets of values of j 's here you will basically have $p_i s$ that means, it will be an element in that set and for what set of values of j 's it will be in the set of $n_i s$ which is negative ideal solution.

So, $p_i s$ is the positive ideal solution and $n_i s$ is the negative ideal solution. So, once you have this what you will try to do? You will try to basically find out the set of the decisions. So, they would be basically for $p_i s$, you will try to find out the positive one and also so called how farther it is from the positive one. So, one is set of closer values one in the set of farther values. Similarly when you do it with respect to the $n_i s$, you will find out the so how close it is for the negative solution and how far it is from the negative solution.

Now, when you have that you will basically find out the ranking based on these ideal solutions distance both for the case for the $p_i s$ and the $n_i s$ and once you do that you will basically find out the so called the ranking system based on both the 4 columns. 4 columns are nothing to do with the rows or nothing to do the columns what we have discussed about the alternatives and the criteria remember that. These are the basically the so called distance functions normalized distance function which I am getting with compressing to the $p_i s$ positive negative that means, close and far from $p_i s$.

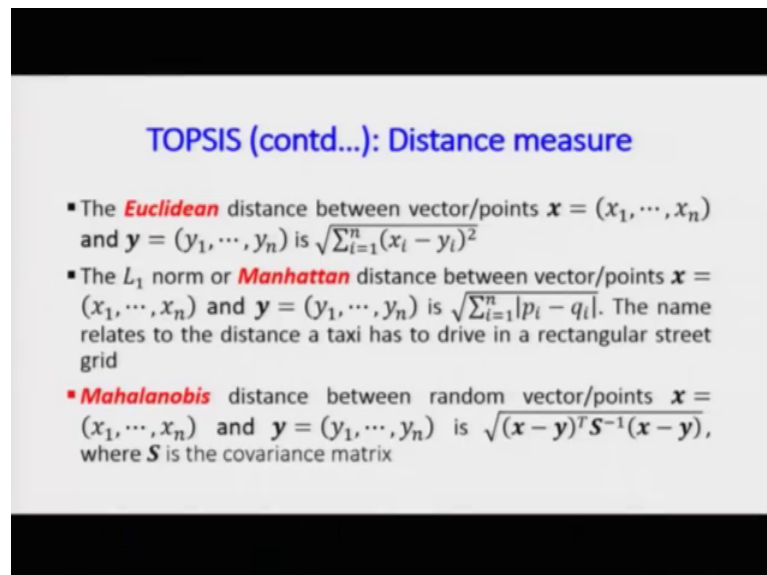
Similarly, the 3rd and 4th column would be the corresponding distance function based on how close and far it is from the $n_i s$. So, once you do that you will basically be able to find out the ranking for each and every alternatives by considering the combination of the criteria. Now remember here I am again repeating when you are going to take the 4 columns, 2 with respect to the $p_i s$ and 2 with respect to $n_i s$, the actual fact is that, the 1st and the 2nd column will be utilized to find out the so called positive benefit and the 3rd and 4th column would be utilized to find out the negative benefit.

It is something to do with the concordance set and then discordant set. So, with this I will basically start and discussing this the once this very simple theoretical concept with a very simple matrix and then go into trying to make a decision that how you choose a

house based on the different criterias and the alternatives which are given. So, um, so let me go into the simple first set of concepts of the distance concept.

Now you remember that whenever we are trying to find out the distance, so distance can be of different varieties. So, distance can be Euclidean distance, so in the Cartesian coordinate you have say for example, 2 dimensional one, you have 2 values x_1, y_1 and x_2, y_2 . So, if you want to find out the distance between those 2 points, consider them as a and b. So, where a has a coordinate system as x_1, y_1 and b has a coordinate system as x_2, y_2 .

(Refer Slide Time: 11:48)



TOPSIS (contd...): Distance measure

- The **Euclidean** distance between vector/points $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ is $\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
- The L_1 norm or **Manhattan** distance between vector/points $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ is $\sum_{i=1}^n |x_i - y_i|$. The name relates to the distance a taxi has to drive in a rectangular street grid
- **Mahalanobis** distance between random vector/points $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ is $\sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{y})}$, where \mathbf{S} is the covariance matrix

What you do is that you basically find out the difference between x's square them up, find out the distance between the y's square them up, add them up and find out the square root. So, you will basically have the Euclidean distance and as per the concept you have if you want to find out the difference between the points, the points are given as x_1 to x_n and y_1 to y_n , you basically find out the difference, square them up, add them up, find out the square root.

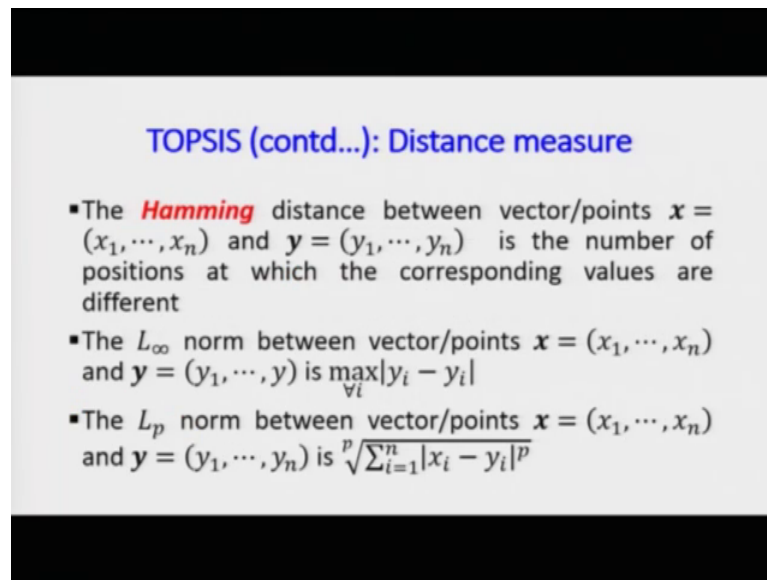
Now, there are different norms of the distance function. Some of them up in statistics concept there are different type of norms, this is the Mahalanobis distance, Pattazhy distance then there in concept of information theory they have the Hamming distance, in general concept of mathematics you are basically the L_1 norm, L_∞ norm, L_2 norm, L_p norm. So, all these things are there they can be utilized, but I am only going to

consider the Cartesian coordinate in a very simple way to solve our problem and obviously, these norms would have some concept of the distance which would be basically coming out from the utility function. We will consider the utility function to be because they are only 4 utility functions which we are considering. So, we will consider the utility function to be quadratic hence they returns to be normally distributed based on this will be proceed.

So, trying to basically minimize the distance if you remember gives us the information you will try to basically find out the difference in other squares and then try to basically find out the average value of that and try to minimize that which leads us to the fact that you are trying to basically minimize the variance which is a characteristics which I have been talking time and again. So, the L 1 norm or Manhattan distance, so in Manhattan, New York all the roads are perpendicular to each other. So, based on that they are just rectangle set up so that is why it is known as the Manhattan norm.

So, L 1 norm is basically this distance where you find out the mod of the difference of these values which is basically here I do not just let me change it, I have not done it. So, the Manhattan norm would be given by the sum of the mod of these values of the errors, not the differences and then you will basically find out the values one minute. So, in the Mahalanobis distance you basically find out the there are two random vectors you basically find out the multiplication of those differences and find out them multiplied by the covariances. So, what you are trying to do is in some way trying to find out the variance and trying to the minimize that and try to try and find out the minimum one for the Mahalanobis one.

(Refer Slide Time: 14:46)



TOPSIS (contd...): Distance measure

- The **Hamming** distance between vector/points $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ is the number of positions at which the corresponding values are different
- The L_∞ norm between vector/points $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ is $\max_i |x_i - y_i|$
- The L_p norm between vector/points $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ is $\sqrt[p]{\sum_{i=1}^n |x_i - y_i|^p}$

Similarly, for the Hamming distance, the values would be basically the position that is the number of position at which the corresponding values are different you will find out the differences, find out whether it is 10 or 15 in number, you report that number that will give you what is the overall n, num n, n means the number of such differences which are there which will give you how far or how close those 2 vector setup points are.

Then for the L infinity norm, you basically find out the max of these differences between x and y. So, you have x and y's the, the difference you find out and then you find out the maximum this distance for the in infinity norm and for the L p norm you basically find out the differences in the mod value try to find out to the pth power and then add them up and then find out the square or say for example, the power 1 to the power p based on which you will try to basically find out that the L p norm distance.

Now, here as I mentioned we will only consider the L 2 norm which is the Euclidean one based on which we will do the calculation.

(Refer Slide Time: 16:11)

TOPSIS: Step # 01 (Construct the normalized decision matrix)

- Assume the decision matrix, $X = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{bmatrix}$
- Convert the entries in X into scaled **normalized** values, $r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{k=1}^m x_{kj}^2}}$, which has no dimension
- Thus we get $R = \begin{bmatrix} r_{11} & \dots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \dots & r_{mn} \end{bmatrix} = \begin{bmatrix} \frac{x_{11}}{\sqrt{\sum_{k=1}^m x_{k1}^2}} & \dots & \frac{x_{1n}}{\sqrt{\sum_{k=1}^m x_{kn}^2}} \\ \vdots & \ddots & \vdots \\ \frac{x_{m1}}{\sqrt{\sum_{k=1}^m x_{k1}^2}} & \dots & \frac{x_{mn}}{\sqrt{\sum_{k=1}^m x_{kn}^2}} \end{bmatrix}$

So, now first the part what we will consider is a very simple theoretical example and then I will go into the, a somewhat practical example where their mean in set of informations being given where you want to find out how you want to purchase a house depending on different criterias. So, assume the value of X which is the matrix of the decision values so these are given by and this is the, remember this is the m cross n matrix. These are cells values are given by x_{11} , I am only reading the topmost row; x_{11} to x_{1n} so that means, corresponding to the first alternative what are the so called values you accrue for that alternative based on the fact that x_{11} would be for the first criteria, x_{12} would be the for the second criteria.

Similarly the last one x_{1n} would be for the last criteria. Similarly, if I go to the last row, so, x_{m1} would basically the value which is accruing from the criteria 1 to the m th alternative. Similarly x_{mn} would basically be the value which is which is accruing from the n th criteria to the m th alternative. So, once you have that so called decision values they are given remember that that will be given. So, you need to find out the normalized values. So, we are using the very simple concept of the normalization concept which is x_{ij} divided by the square root of the sum of the squares of these values for each and every cell.

Now the word what I said last is basically each and every cell would be either considering the row wise or the column wise. So, say for example, I want to find out the

normalized value for x_{11} , so, I can either do the normalization along the 1st column or along the 1st row. So, if it is along the 1st column the actual values in the denominator I am not talking about the square root, I am not talking about the summation, I am only talking about the squares inside so it will make you easy so basically you find out the squares, add them and find out the square root.

So, the values which you will basically add them are corresponding due to the fact if you are going through the column wise, it will be x_{11}^2 whole square, then you will basically have x_{21}^2 whole square the so x_{31}^2 which I am mentioning is in the subscript and the square values is in the in the superscript in on the top, then the 3rd value will be x_{31}^2 square till the last value it will be x_{m1}^2 square, you add them up, square root and if you find out then each values along the 1st column would be divided by the fact which I mentioned.

Now, if you are trying to find out the normalization another thing which I should mention, so once you have normalized, so if you go to the 2nd column you will basically use each and every cell values in which are there in the 2nd column you will basically square each of them, add them up square root. And, that would be the denominator in each and every values which these would be utilized to divide each and every values in the second column.

Similarly, for the 3rd column the value which you will utilize to divide each and every cell for the 3rd column would be the corresponding square of each and every value for the 3rd column which would basically be x_{13}^2 whole square, x_{23}^2 whole square, x_{33}^2 whole square till the last one which will be x_{m3}^2 whole square. Now, if you are doing the denomination along the rows. So, let us consider the 1st row the values which will be coming in a denominator would be likewise like this it will be x_{11}^2 whole square, then plus it will be x_{12}^2 whole square, then plus x_{13}^2 whole square till the last value which is x_{1n}^2 whole square.

Now, if I go corresponding you use this concept go to the last row, the corresponding value which will be coming in at the denominator again I am not going to mention the square root value I am not going to mention the sum. I am only mentioning the values which will be summed up these are like this x_{m1}^2 whole square plus x_{m2}^2 whole square dot dot till x_{mn}^2 whole square.

So, once you have that you will basically have so called the normalized set which I will mention as R. So, these are the values which I have done, so their normalization can be done either row wise or column wise. I am not going to do details what is done here, but you will surely understand.

(Refer Slide Time: 21:10)

TOPSIS: Step # 01 (Construct the normalized decision matrix) (contd..)

Assume, $X = \begin{bmatrix} 10 & 25 & 25 & 30 & 15 \\ 05 & 15 & 35 & 40 & 10 \\ 15 & 25 & 40 & 45 & 10 \\ 20 & 30 & 30 & 35 & 05 \end{bmatrix}$

Scale the values using *normalization* concept, i.e., $r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{k=1}^m x_{kj}^2}}$ (you can use any other concept of utility also)

$R =$

10	625	625	900	225
$\sqrt{100+25+225+400}$				
5	225	1225	1600	100
$\sqrt{100+25+225+400}$				
15	625	1600	2025	100
$\sqrt{100+25+225+400}$				
20	900	900	1225	25
$\sqrt{100+25+225+400}$				

So, assume the values which are given for the X matrix the priority so called values which are there are the first I am only read the first row which is basically 10, 25, 25, 30, 15 similarly the last row is 20, 30, 30, 35, 05. Now, remember one thing here the number of this alternatives is 4 in number which is the 1st row, 2nd row, 3rd row, 4th row and the number of criterias which are there which is the 1st column, 2nd column, 3rd column, 4th column, 5th column that means 4 number of alternatives and 5 number of criteria. Now consider the value of 40. So, what would what 40 mean? That means, it would mean for the 3rd alternative the value being accrued by the 3rd criteria would be 40.

Similarly, if I consider the 5 here, so it means that for the 4th of alternative the value being accrued by the last criteria is 5; so, once you have that you normalize again used either the row wise or the column wise. Now here what I have used is, I have given these values, but I have used two different calculations in order to make you understand. So, in this case what you can do if you find out the square root so obviously, the values would be 10 divided by the sum which square which is 10 square 100, 5 square 25, 15 square 225 for 20 square 400.

So, this value would be utilized for each and every value calculation in the first column similarly. So, if I go to the second column, the corresponding square values are 625 square 625, 15 square 225, 25 square 625, 30 square 900. So, these values are given. So, here I have considered the square root, the corresponding 3rd, 4th, 5th I am not considering this square root, but I am just giving an example so, they are again square values of each and every corresponding cell values in the 3rd column similarly 4th.

So, we basically find out the square root of, so, I should also mark it with blue colour so you will understand. So, blue goes blue, green goes green, red goes red, yellow goes yellow. So, the blue one means I am squaring 30 which is 900, 40 which is 1600, 45 which is 2025, 35 which is 1225. Similarly, come to the last column 15 square, 10 square, 5 square. So, these are the values.

(Refer Slide Time: 24:50)

TOPSIS: Step # 01 (Construct the normalized decision matrix) (contd..)

▪ $R =$

0.133333	0.263158	0.143678	0.156522	0.500000
0.033333	0.094737	0.281609	0.278261	0.222222
0.300000	0.263158	0.367816	0.352174	0.222222
0.533333	0.378947	0.206897	0.213043	0.055556

▪ Check each column adds up to 1 as it should be

So, once you have this the normalized values considering the square root one, I have these so, if you change it, it will be giving you different values. So, if you want it let me show you the differences. So, it will be easy. So, I will use the excel sheet. So, I will use this table these values

(Refer Slide Time: 25:09)

	C1	C2	C3	C4	C5	
19		0.11594	0.26087	0.26087	0.35507	0.00725
20						1
21	C1	C2	C3	C4	C5	
22	A1	10	25	25	30	15
23	A2	5	15	35	40	10
24	A3	15	25	40	45	10
25	A4	20	30	30	35	5
26						
27	C1	C2	C3	C4	C5	
28	A1	0.2	0.26316	0.19231	0.2	0.375
29	A2	0.1	0.15789	0.26923	0.26667	0.25
30	A3	0.3	0.26316	0.30769	0.3	0.25
31	A4	0.4	0.31579	0.23077	0.23333	0.125
32						
33		1	1	1	1	1
34						
35						

So, let me write them 10, I will correct myself and I wrote it down so there is no mistake. I am calling for the first column 10, 05, 15, 20, 10, 05, 15, 20 which is 10, 05, 15, 25, 15, 25, 30 25, 15, 25, 30 I go to the so, I should also mention that what I mention. So, it will A 1, A 2. So, it will be easy for you to understand. So, these are the alternatives and so obviously, criterias and you see C 1, C 2 not that c concept which is coming out when you do the concordance discordance set, I will continue using C here. So, so they were 5 criteria. So, it is done. So, let me come to the 3rd. 25, 35, 40, 30, 25, 35, 40, 30, 25, 35, 40, 30, 30, 30, 45, 35, 30, 30, 30, 40, 45, 35, 30, 40, 45, 35, 15, 10, 10, 05, 15, 10, 10, 05.

Now, I will calculate different values. I will write the weights later on. So, let me write them I would let me if you are if you are insisting. So, I will do first the simple normalization which is sum. So, this should be dollar in order that it does not move values. So, I have. So, now, see, so this value is basically 1st cell divided by the sum of all the cells this is the 2nd cell divided by the sum of the cells, this is the 3rd cell divided by the sum of the cells, this is the 4th cell divided by the sum of the cells.

So, I will take another few extra minutes for this class it will be exceeding 30 because I want to finish 30 minutes because I want to finish this table. So, this would be divided by sum. So, it will be dollar here dollar here, dollar here, dollar here so done. So, this will be 25 divided by sum. So, put a dollar. So, we fixed it copy, done come to the 4th column

put a dollar here dollar here dollar here. So, this is done and last column sum of this so, put a dollar here so copy.

So, last now to wrap it up this class so obviously, you have normalized along the columns. So, let us find out obviously, this will be true, but I want to show it to you. So, this is 1, 1, 1. So, you could have done it along the rows that would have been the same. So, with this I will end this 33rd third lecture and in the 34th and then 35th which will be the last two lectures, for the 7th week I will try to wrap up the example of buying the house in the using the TOPSIS method. Have a good day and.

Thank you very much for your attention.