

Learning Analytics Tools

Professor Ramkumar Rajendran

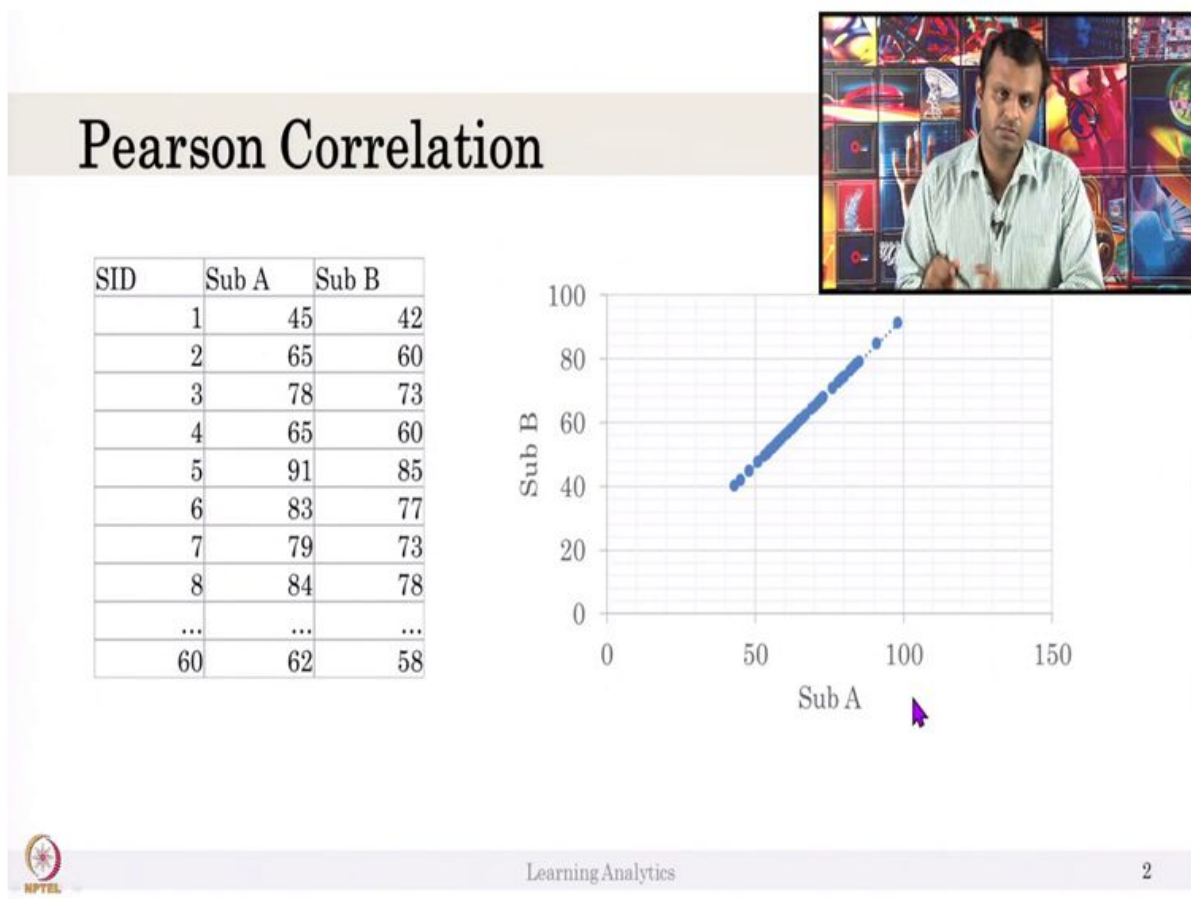
Educational Technology

Indian Institute of Technology, Bombay

Lecture 5.3

Correlation Matrix

(Refer Slide Time: 0:24)



In this video, we will talk about correlation matrix using the Pearson correlation coefficient. So, if you remember, we discussed about students' marks in subject A and subject B. We used a

stacked bar chart or a bar chart to compare or box chart to compare these things in descriptive analytics. Let us look at the same data.

We have 60 students marks in subject A and subject B, same students. So, we can plot the relationship between subject A and subject B. You see that students who are scoring good in subject A definitely scoring good in subject B. Actually, I created these marks with linear relationship say 0.93 multiplied by subject A marks.

So, I created these marks to show there is a linear relationship, so it is a linear relationship and it is perfect and it could be 1, like positive 1. And, you know there is a strong relation between subject A. The student who scores well in subject A, that is math, they might definitely score well in science, this can be established.


(Refer Slide Time: 1:31)

Correlation Matrix

Stud.ID	Attendance in %	Mid Term Marks	Final Marks out of 100
1	56	45	70
2	45	32	50
3	85	56	90
4	80	73	80
5	90	65	75
6	100	80	90
7	95	65	78

x_1
 x_2
 y_1

x_1, \dots, x_{10}
 \rightarrow



	Column 1	Column 2	Col 3
Column 1	1		
Column 2	0.90	1	
Column 3	0.84	0.80	1

\rightarrow Since Att & mid term marks are highly correlated.

\rightarrow Attend to Predict Final marks

Let us look at the other example. We saw we have attendance and midterm marks and final marks in the semester. Now, we have two independent variable and one dependent variable, that is you see this is X_1 , X_2 and one dependent variable and two independent variables. We can compute the correlation between these values like what is the correlation of, between attendance and final marks and what is the correlation of midterm marks and final marks, let us look at that.

So, consider column 1 is attendance, column 2 midterm marks, and column 3 is final marks, this matrix was created using excel, the simple tool available in Excel can help you to do that. So, is there a relationship between column 2 and column 1? Yes, this relationship X and Y highly correlated, 0.9, is there a relationship between column 1 and column 3? So it is 0.84.

This is the correlation coefficient between attendance versus final marks. If the students have more attendance, it is highly likely to scores better score compared to students who have low attendance score. So, this is self-correlation, like attendance versus attendance, midterm marks versus midterm marks, final marks, so obviously it is 1. Hence ignore diagonal part of this matrix.

Let us look at that, column 3 versus column 2, what is the correlation coefficient of midterm marks versus final marks? It indicates 0.8, it says if the students score well in the midterm, like in midterm or mid sem exams, he definitely will score well in the final exams, so that is the final exam score.

This is the correlation matrix, which tells you the relationship between X_1 and Y_1 , X_2 and Y_1 .

Attendance and midterm marks have a high correlation, you know, it is 0.9. Now there is a question that comes in, there are a lot of questions if we want to create a simple classifier, should they use both X_1 and X_2 to predict Y_1 or should I use any one of them. It is a very interesting question, the reason is you see there is a high correlation between attendance and midterm marks because 0.9 is a very high correlation.

If the student has more attendance, definitely he has more midterm marks, if he has less attendance, fewer marks in the midterm exam. So, there is no point in picking both variables for our research, for creating the classic predictive analytics methods. Instead, you have to pick one

of them. The general approach is, we use the one which has the highest correlation with the dependent variable, that is column 3 with column 1, that is attendance.

So it is very simple. Since attendance and midterm marks are highly correlated, we can pick one of these features for the classification, which one? This depends on how strongly they are correlated with the final marks. I would choose, so I would choose the attendance because attendance one is strongly correlated compared to the midterm marks.

Midterm marks are also good, 0.8 compared to 0.85 might be 0.84, so I will choose, so attendance to predict final marks. So that is what I will try to reason. I will try to use the best independent variable, which is correlated with a dependent variable for creating classifier. For these two values, it is easy to predict. For example, since there is a slight correlation between attendance and final marks, it is simple, we just have to create a simple, if the condition or some filtering criteria to figure out that the student will pass the exam or not pass the exam or student will, how many students will score using the equation or something like that.

Consider you have more than say two, that is you have X_1 up to X_{10} , you have assignment score, you have midterm marks, the number of the attendance, the engagement level in the class, the responses in the discussion forum, you have a lot of other information available from the classroom environment or some learning environment. If you have that kind of information, then having all these features to predict it may be a problem, if you want to just remember addition time.

So, you can do the, so you can do the feature selection algorithm. One of the algorithms is doing the correlation between an independent variable with the dependent variable, then picking the strong correlated one and also ignoring the values which are weakly correlated.

There is one technique which was used very widely some time ago. Now, due to the advanced classifiers, we do not really care about this, how to select these features to predict Y_1 , instead we use other techniques, which are available in the tools, you might see them when you are using the tools, which help you to pick the right features for prediction, but still, it is very good practice if you want to go for prediction, try the correlation method with correlation matrix, that might help you.

Also, the correlation matrix, if you have to say 10 columns, it might help you to provide a relationship between each independent variables and how it is related, how these variables can be predicted using the values given in the correlation matrix. So, I would recommend if you have more than one variable, one independent variable, go ahead and complete the correlation matrix to get a sense of the data and see how they are related, is there any linear relationship between these two variables.

(Refer Slide Time: 9:14)

I used a correlation matrix in the previous slide, I hope you understood that. Let us do a small activity. Given the four plots, $X_1 Y_1$, $X_2 Y_2$, $X_3 Y_3$ and $X_4 Y_4$. Given these four plots, can you guess the correlation coefficient of these charts, like chart 1, chart 2, chart 3, chart 4? Can you guess this? Also, if you guess it, what is the inference you are making it out of this correlation

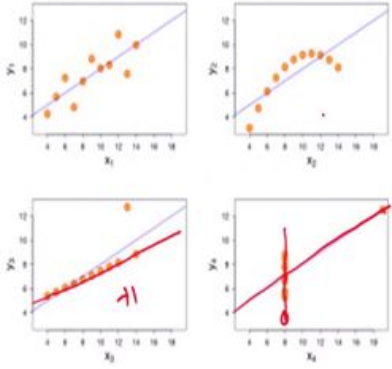
coefficient. No need to do any calculation or something, just guess it, just try to guess what will be the value. This data is from paper, so this is the cited paper.


(Refer Slide Time: 9:54)


Activity

Pearson Correlation

- Same Correlation – 0.816







Learning Analytics

5

So, all of them have the same correlation coefficient, you might have guessed it rightly, if it is, good. All of them have the same correlation coefficient, why? In the last video, we talked about the drawbacks of Pearson correlation, right? This actual example gives you all the details of that. Let us look at it.

The first one, X_1 versus X_2 , it is just scattered around the linear relationship, so there is a 0.816 relationship, it is okay, it is 0.816, good. This is not a linear relationship, right?

So, here the drawback of Pearson correlation, that nonlinear relationship is not identified here, right? If it is good, correlation coefficient, you should be able to identify nonlinear also, but unfortunately, there is no metric which identifies this nonlinear, this kind of nonlinear relationship very accurately.

Here, the correlation if you remove this particular point, for example, if I remove this point, the line, it will be exact +1 because of one outlier, only one outlier, the value is less, 0.816. So, the Pearson correlation coefficient is sensitive to the outlier, okay? How sensitive, that is defined, explained in this particular figure look at this figure and if you remove this outlier, absolutely there is no correlation between X and Y, it is 0.

There is no correlation coefficient, right? Not increasing, it is just simple 0 because of this outlier

So, this tells you that Pearson correlation coefficient is sensitive to an outlier, this tells you how much sensitive it is and this tells you, Pearson, correlation coefficient does not consider a nonlinear relationship, so that is the typical figure which explains the Pearson correlation coefficient, also its drawbacks.

(Refer Slide Time: 12:28)

Summary

- Correlation matrix



So, in this video, we saw what is correlation matrix and we saw a plot which explains the drawbacks of the Pearson correlation coefficient. Thank you.