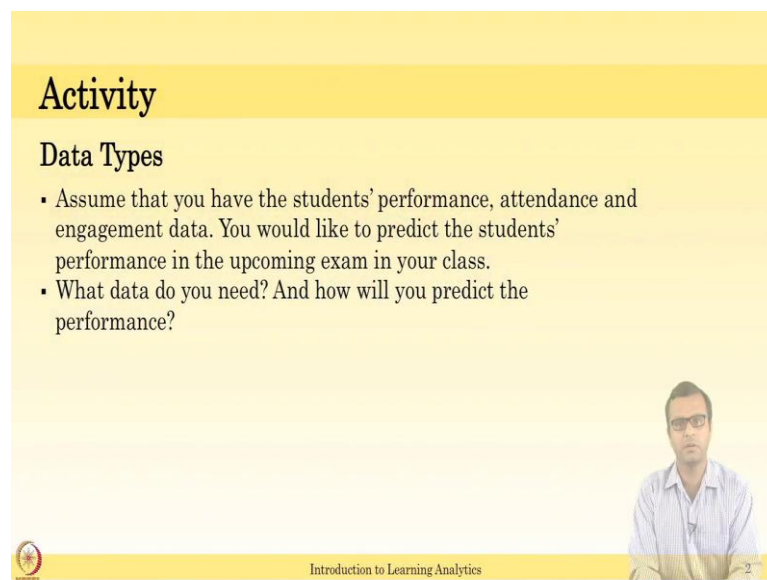


Introduction to Learning Analytics
Prof. Ramkumar Rajendran
Department of Interdisciplinary Programme in Educational Technology
Indian Institute of Technology, Bombay

Lecture - 15
Linear Regression

In this week we will start with one of the prediction algorithm called Linear Regression, then we will have a demo of tool called Weka, this is a machine learning tool. And we will have a small course project and description about the data. In last week's we saw what is predictive analytics, let us look at what is one of the predictive analytics algorithm called linear regression.


(Refer Slide Time: 00:41)



Activity

Data Types

- Assume that you have the students' performance, attendance and engagement data. You would like to predict the students' performance in the upcoming exam in your class.
- What data do you need? And how will you predict the performance?

 Introduction to Learning Analytics 2

Let us start with a activity, assume that you have access to data such as attendance, engagement and performance of the students in a class. So, in a class of 60 you will have a students attendance over the period of time, also their engagement in the class or engagement in the Moodle and their performance in the test. Now, you would like to predict the students performance based on their attendance and engagement in a upcoming exam.

So, what data you need or how do you predict the student's performance in the upcoming exam? Please think about the answer, write down the answer, after completing the writing your answer you can resume the video to continue.

(Refer Slide Time: 01:28)

Activity

Predicting Performance

- Identify patterns from the historical data
- Attendance vs performance, engagement vs performance etc.
- Questions asked, difficulty level, topic covered, questions reused!
- Develop model from the data
- Extend the model
 - Apply the learned model on new data – Current attendance



Introduction to Learning Analytics



As you saw in a last week predictive analytics or to predict performance in this particular scenario, we need to identify a historical data. We need to collect historical data, that is in our case it is attendance and engagement and performance. From the historical data we need to compute the correlation between attendance versus performance or engagement versus performance. When we compute the correlation, if the correlation is average say 0.5 on positive correlation or negative correlation; if there is a some medium to high correlation then only we will consider this particular variable otherwise we may not consider this variable.

So, first thing is we need to compute the correlation between the variables like attendance versus performance or engagement versus performance etcetera. This we will do in a diagnostic analytics and descriptive analytics to see the data what is actually how the data looks like, is there any relation between these two data. Once we have a statement saying there is a relationship between attendance and performance, also engagement and performance then we will go for a next level that is predictive analytics.

Or, you might have a very fine grained data such as not just a performance instead the students performance in a each questions, for that you might have a data of what is the difficulty level of the questions or what is the topic covered. So, you might be asking a questions related to a topic you covered in the class or are you reusing the questions. Because, the students might have a questions borrowed from the seniors or from the last

year exam papers or you reusing the questions; students might have prepared those questions and they might do better, even though they do not understand the concept. So, you can consider these factors to predict the students performance or simply we will start with the engagement or attendance in the class.

From this data we develop the model, then we extend the model or extrapolate the model to predict the future events that is predictive analytics. Let us talk about one such predictive analytics algorithm or a algorithm or ML algorithm to predict a future events. The basic and very simple one is linear regression, linear regression gives you intuition that how it is related, how the performance is related to attendance or engagement and it is easy to understand. Also, it gives you the intuitive weight between two different variables in a linear regression. We will talk about this the variables and performance in detail in this LED.

(Refer Slide Time: 04:23)

Linear Regression

- Given a dataset X and Y or Attendance and Performance, Linear Regression models assumes that there is a relationship between Y and X and that the relationship is linear
- Regression analysis analyses the relation between dependent variable and independent variable



Introduction to Learning Analytics



Consider given a data set X and Y, X can be attendance, Y can be performance. So, we want to predict student's performance from the attendance, if we have a single independent variable, independent variable here is the attendance and dependent variable is performance, because performance depends on some other variable. The performance is depending performance might change based on the other variable so, it dependent variable or the target variable. So, you might consider that the attendance and

performance, then you have to come up with the linear regression model and identify what is the relationship between these two variables..

The basic assumption is there is a linear relationship between independent variable and dependent variable, that is attendance and performance. How do you establish this basic assumption? You simply plot them, when you plot them there is a correlation then you can say there is a linear relationship. If the plot looks like a fully scattered dots, then you can understand there is no correlation, then the data may not be used for a linear regression; you might need to use the some other algorithms. Regression analysis analyze the relationship between dependent and independent variable to create the linear regression model.

(Refer Slide Time: 05:58)

Simple Linear Regression

Simple Linear Regression Model: $y = wx + c$

W – Slope

C · Intercept

If we have 6 students data, we will have 6 pairs of attendance and performance data $\{(x_i, y_i), i = 1, 2, 3, 4, 5, 6\}$

In Linear regression to goal is to find a linear relationship that best fits the data.



Introduction to Learning Analytics



Let us start with the simple linear regression model. In a simple linear regression model we consider one independent variable and one dependent variable. The generic formula for simple linear regression model is equal is given here

$$y = wx + c$$

it is

$y = mx + c$, we draw a slope in a graph in a class 6 or class 7.

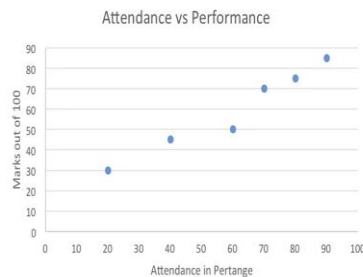
In y its the performance you are predicting and w is the weight associated with x , x is the attendance and c is the intercept and w indicates the slope of this line.

Let us consider we have a 6 students data, 6 students attendance and their performance. Like now we can say the data is $x_i y_i$ for a student 1; what is the student 1's attendance percentage? Say 80 percent, 60 percent. What is the student's performance? Similarly, for 6 students you have pair of data 6 pair of data. In a linear regression, the goal is to find the linear relationship that best fits the line for this data.

(Refer Slide Time: 07:06)

Linear Regression

Attendance in Marks out of	
%	100
20	30
40	45
60	50
70	70
80	75
90	85



Descriptive: Students who attend the class regularly scored good in exams



Let us look at the data, in this linear regression there are 6 students data like a 20 is the attendance percentage, the mark out of 100 is 30; for attendance percentage 40 the mark out is 45. Similarly, for the other data we have a marks out of the attendance percentage. Students, now, from in this data we can see that student do attend the class regularly scored good in exams, this is just by plotting the data in a scatter plot.

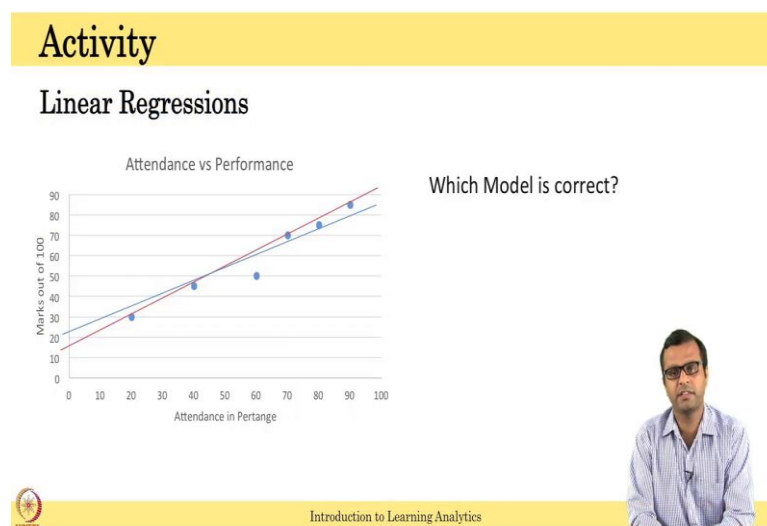
Now, you can see there is a relationship when the attendance percentage increases the marks also increases. If there is a linear relationship you can see then you can apply a linear regression model for this particular data. Let us try to fit a line for this particular data.

(Refer Slide Time: 07:57)



So, we are trying to fit a line, say there is a line which tries to fit almost all the data here and this data this line has a intercept value say 15 year and there is a slope is there and this line looks good. So, is this linear model correct? How do you verify it? How do you validate it? Please think about it and write your answer in a paper, after completing your task please resume the video to continue.

(Refer Slide Time: 08:32)



So, which model is correct? Here we have two lines that is fitting this particular data, the red and the blue one. So, which one is correct?

(Refer Slide Time: 08:51)

Linear Regression



Predicted Marks using Regression
Model: For a 20% attendance, marks
is 35%

Error at $x_1 = (\text{Predicted marks} - \text{Actual marks})^2$
 $= (35 - 30)^2 = 25$

Least Square Method to pick the best
fit line



Let us do the linear regression calculation. In order to identify the best fit line we need to compute a least square method. Least square method is very simple it is nothing, but comparing the observed value that is the line fit value into the original value. For example, if I have drawn this particular line, here are the 6 data; if you look at the data or attendance percentage is equal to 20, the mark is almost equal to 35. For given this line the linear regression model, for attendance percentage of 20 the student is expected to get a pass percentage pass mark of or the performance is 35.

However, we know that from the picture for from the data we know that for the attendance percentage is equal to 20, the marks out of 100 is only 30. But, this particular model predicts for attendance percentage equal to 20 will have 35. So, that predicted mark is 35 minus the actual mark that is 30, this difference is 5 and we are squaring it. So, squaring it just because sometimes the predicted mark will be lower than the actual mark, in order to address the lower mark or higher mark we just want to find the difference and square it. So, that we will have a difference square that is called least square method this 25.

If you sum up all these least errors at each point x_1, x_2, x_3, x_4, x_6 then you will have some of these values, it gives the least square value. If there are two or three lines fitting this particular linear regression model, you need to select the line which has the least error. For example, this particular line the least, the error value at x_1 is equal to 25,

similarly you have to compute for x_2 and x_3 , x_4 , x_5 and x_6 . Then you have to add everything, sum of all these values computed for this particular model, linear regression models error value. Similarly, you can draw other lines, then compute the error value for the each linear model.

However, the mathematical formula to do this is not like this, instead of computing standard deviation and other methods like how to calculate the intercept; there is a detailed mathematical description how to compute this, but in this course we are not going to discuss that. The basic abstract idea is that, in order to find the best fit you need to find a least square method, this is the basic concept for linear regression.

(Refer Slide Time: 11:28)



And you can compute a linear regression in a existing software like Excel Microsoft Excel and we just use a same data and we have calculated the linear regression model; y is equal to 0.7794 that is the weight of this particular the slope of this line into the x value plus 12.40 12.4, will be the intercept. If they extend the line here, it might here it may come here to 12.402. The intercept also decides where the slope is based on and the linear regression coefficient is equal to point the and the linear regression coefficient is 0.9515.

So, it indicates there is a very strong correlation between attendance and the performance. As you might know that this data is not actual data; we just populated this data to show the linear regression. In one of the LbD's in this week we might give you

other data of students and we might ask you to use the excel to create a linear regression model, to get your hands dirty how to use the excel for the linear regression.

(Refer Slide Time: 12:44)

Multiple Linear Regression

- Performance may not be only related to attendance
- Multiple independent variables
- Engagement in Discussion forum, attendance

Multiple Linear Regression Model:

$$\text{Performance} = W1.X1(\text{Attendance}) + W2.X2(\text{Engagement}) + C$$

X_i – i can be $\{1, 2, \dots, n\}$

W_i – indicates the strength of independent variable on dependent variable



The next topic; the next topic is multiple linear regression. Sometimes the dependent variable performance might have depend on multiple independent variables. For example, performance depends on engagement also in the class, also the attendance. The engagement can be collected from students engagement in the class or the engagement in the LMS like Moodle or blackboard.

So, we will have multiple independent variables, performance is dependent variable and engagement and attendance is the independent variable. So, the generic equation for multiple linear regression model is given here,

$$\text{Performance} = W1.X1(\text{Attendance}) + W2.X2(\text{Engagement}) + C$$

performance is equal to weight $W1$ into the weight is actually slope in a previous generic model, for simple linear regression here the weight 1 into the $X1$; the value that is attendance value plus weight 2 and engagement then intercept. We will train the model to learn these weights from the existing data.

The weights values will be learned and also the intercept values will be learned from the existing data, the historical data. Then we will apply this data to predict the students future performance in the next class or next exam. So, the X_i that is the X_i can be more

not just two performance, it is depending only on attendance and engagement. It can also depend on multiple variables X_i can be 1 to n , maybe the students submission date of assignments or the students number of upvotes in the discussion forum. Or, the students performance in the midterm exams, a lot of factors, a lot of variables can be involved. The W_i actually indicates the strength of independent variable on dependent variable. So, this is the very beautiful feature of linear regression.

It is a very very intuitive method, when we have a multiple independent variables like performance and sorry like attendance and engagement; we can see which weight which is more strength on the dependent variable like performance. So, W_1 indicates the strength of independent variable on dependent variable. However, the linear regression have a drawback that is it assumes the relationship between the two variables is linear. And we cannot use the linear regression if the number of independent variables is too much or more than number of samples data we have. So, if the number of independent variables is really high, you might need to consider some other algorithms.

Thank you.