**Introduction to Data Analytics**

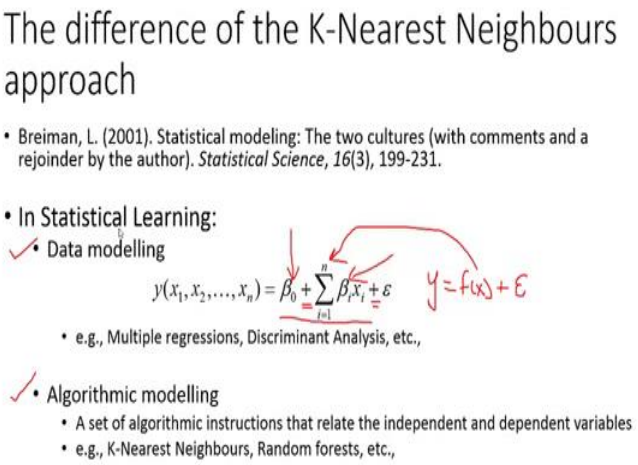
**Prof. Nandan Sudarsanam**
**and Prof. B. Ravindran**
**Department of Management Studies**
**and Computer Science and Engineering**
**Indian Institute of Technology, Madras**

**Module – 04**
**Lecture – 22**
**Data Modelling and Algorithmic Modelling Approaches**
**Regression and K-NN Techniques**

Hello and welcome to this lecture on K-Nearest Neighbors Techniques. So, in this lecture we will introduce a new supervised learning approach called K Nearest Neighbors and a broader thing that we are trying do to with this lecture is by introducing this technique and comparing it to something that you are already familiar with, which is regression analysis. By doing that comparison, we are hoping to kind of illustrate and help you appreciate two very different styles of performing the supervised learning analysis and you know and therefore, the process of prediction itself.

Two very different approaches, two predictions and it is a kind of important to know that, because up until now you heard of what supervised learning is in theory and the first technique that we gone into and explained is regression, which is the certain way of for instance performing a prediction. The hope is by introducing k nearest neighbors, you see a very different way of achieving the same goal and it is important, because you will realize that a lot of other machine learning techniques, take inspiration from this approach.

(Refer Slide Time: 01:26)




So, this dichotomy if you will was something that was first pointed out by Leo Breiman, a very famous statistician, where he said there are two very prominent cultures to statistical modeling and out here, he primarily talking about supervised learning approaches and he highlights and you know, there is some amount of overlaps and so on. And I would not really say these two cultures and two different, it is not like a classification as much as two very different styles.

And the idea is that, he says look there is the data modeling approach, which is what you for instance in the standard regression, where you have some you know output variable y and you kind of envision this output variable has some function of your input variable or variables x. So, for instance you would say y = f(x) + noise. In the case of linear regression and here a kind of shown you an examples of multiple regression, this f(x) becomes fairly straight forward, it intercept + $\beta_1$ $x_1$, $\beta_2$ $x_2$ for how many ever input variables you have.

Suppose, you just have one input variable that would be $\beta_0$ + $\beta_1$ $x_1$, it is fairly straight forward. So, the point he makes and it is generalization, because he says us about many other statistical methods which he says, you have these whole breed of approaches to supervised learning which capture a functional relationship between y and x and that function is in some sense cast in stone and the only job you are left with is just go, get the data and figure out what the parameters are of the function.

These βs, the job is to take some data and figure out the βs, this functional form itself

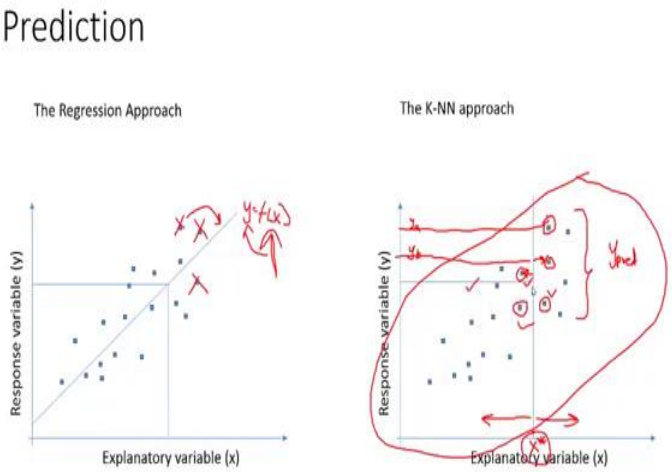
which is said in this case of the linear regression, that it is a linear combination of the different inputs + some Gaussian noise that is for instance cast and stone. And he says that you have those approaches and then, you have an alternative breed of approaches and he calls them the algorithmic modeling culture. So, the first one he calls is the data modeling culture, which is your and the best examples of the data modeling culture could be is the multiple regression.

So, we covered that in good detail in the previous lectures. So, he says now look there is another approach, another approach is what he calls the algorithmic modeling and there he says essentially, the focus is not as much on a rigid mathematical model that you have presupposed that have you apriori and that creates it is relationship between y and x, which is already cast and stone. But, instead you really have a set of algorithmic instructions or algorithmic ideas that relate the independent and dependent variables.

So, y is in the loose sense of the word of function of x, but that is really captured by algorithms rather than rigid mathematical models. Now, these two, there is a reason why in the start I said these are not like two classifications, but there are more two styles only because you can describe any algorithm in a mathematical form and perhaps you can describe the math in an algorithmic form.

But, what I am going do is I am going to explain to you the k nearest neighbors approach and we are going to talk about the k nearest neighbors approach as an example of the algorithmic modeling and hopefully at that point, it becomes really clear us to what the stylistic difference is between these two approaches.

(Refer Slide Time: 05:12)




So, let us look at how a prediction task happens with your linear regression approach. The way it happens is you have some data and right now I am focusing on the graph on the left hand side of the screen, you have some data, you do a regression analysis and what is the regression analysis do, it fix the line through the data, it fix a line and if you are using ordinary least square regression, if it is a line that minimizes the square deviation between the data points to the line and so it chooses the line that achieves this target.

So, you have you fit this line. Now, what you do when you need to make a prediction? And someone comes along and says, oh great, so you done a regression analysis. Could you tell me what, y I should expect for a given x? So, they come and give you an x, so they give you this x and let us call it x, let us just call it $x_1$. So, they give you an $x_1$ and say can you tell me what why I should expect and what you do is you draw this straight line up like it is already there and you say, let me see where my what value of y I would get based on my fitted model.

So, I basically take this x value, draw a line up to my regression fitted line and then see what the height of that point is. So, this is my predicted y you know, so may be $\hat{y}_1$. So, if somebody comes and gives me this x all I do, I do not really I have used all of this data to create a line, but then after that I can lose these data points, this data points can be erased from my memory. All I need to do to make a prediction at this point is I need to know this line and this line I have already created with the regression.

So, I have created this line I have it my regression and someone comes and ask me a question is to what y would you see with particular x, as you know this line has a form y = $b_0$ + $b_1$x. So, someone comes and gives me a particular x. So, call it $x_1$ I will just substitute this $x_1$ out here, I know the values of b 1 and $\beta_0$ from the regression, that is how I was able to plot the line. I will then just get a y, because I know all the three terms on the right hand side.

So, I will get a predicted y and that is my prediction of what why I will see for a given x. So, that is how a prediction task takes place of the regression approach. Now, let us take a look how the k nearest neighbors approach works, the way the k nearest neighbors approach works is that I basically I do not fit a line. So, I do not have line or I do not have mathematical form, the idea behind k nearest neighbors is that if you come to me with a question as to hey can you predict for me what y I will see for a given x.

I will take the given x and I will ask myself who it is nearest neighbors are. So, somebody walks up me and says can you give me a prediction of y for this $x_1$ let us call it $x_1$, let us call it $x_*$ maybe. So, $x_*$ for this $x_*$, because $x_1$ tends to have the connotation that it is the first data points. So, I going to call it $x_*$, so for this $x_*$ can you tell me what my predicted output would be. Now, remember I do not have a fitted line, this is a completely different approach to making predictions.

But, what I do is I go to this line and on my x axis, on my input variable axis I try to find the nearest neighbors of neighboring data points to the x under question. So, clearly this data point is kind of close to this line and this data point is may be close to this line and so on. And the idea is that this K-NN approach has a parameter which is k. So, let us say I have chosen five as a parameter and we will understand what the five means, it is means I am looking for the five nearest neighbors defined by the distance from my point under question, so this distance.

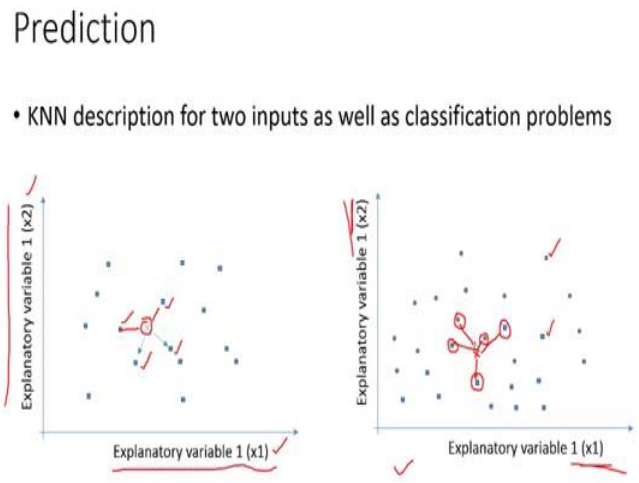
So, I am going to see the five closest data points and if I do that for instance, I see that these five data points are the ones that are closest. So, what do I do once I have identified the data points I take their average wise to make a predictions. So, I take the average of this y. So, let us call that $y_a$ this y, $y_b$ and so on. So, I do the same thing for this data point, this data point and I take the average of all those five specific wise and that is my predicted y, the predicted wise the average and there are many modifications to this, sometimes you do not take the average, you might do like the localized regression there,

there are other ways, where you do not just take the arithmetic mean you might take the median you might do other things.

But, the core approach with k nearest neighbors is that you have not fit a line, you not created any mathematical abstraction of the data, you not abstracted away from the data. For instance, in the regression remember I told you if I need to do a prediction task, I do not even need to remember the data points I can throw all my data points wave. Once I used the data points to create this line, I now only need the line to make a prediction I now have a y = f(x). So, I have this functional form and I just need to plug in my values of x that I want to use to predictions and now I will automatically get a predicted y.

Here we are not doing that abstraction, we are retaining the entire data set of points in the k nearest neighbors approach, we need all our data points. Now, you come and ask me a question about a particular x, I am going to go and look at the nearest neighbors of that x and if it is five nearest neighbors I am going to look at the five nearest neighbors, if it is ten nearest neighbors I am going to look at the ten nearest neighbors in all these cases I am going to look at nearest neighbors and you know either take a vote if it is a classification problem or if it is a regression problem am I choose to do choose to take something like an average. So, that is fairly clear in this line that I show is kind of what is approximately the arithmetic mean and we use that and finally, that is the output that we going to predict.

(Refer Slide Time: 12:20)




Now, that should kind of illustrate k nearest neighbors for you it is a fairly simple

technique a lot of the focus in using this approach to solve problems is based on how, because it is computationally very hard. Because, it is storing all the data and you need shift through all the data to find the nearest neighbors good amount of the focus is on the data mining aspect or the computational aspects using something like that, but it is also very convenient when you have no ideas to what the functional form is.

So, suppose the linear regression approach can be very useful if you believe the relationship between y and x is the straight line. But, if you do not want to make that assumptions at all k nearest neighbors approach is fairly flexible to any function of form that y and x could have. And there are two points to note here just an addition to what we have discussed which is this k nearest neighbors approach. It obviously works when you have multiple input variables. So, the examples that we took in the previous case there was one input variable and then there was this output variable, this is the output variable.

Now, what if had multiple input variables, the same idea? So, here we have one input variable in the x axis, one input variables in the y axis. So, where is the y, where is your output variables, one way to think of it is that it is coming out the screen essentially we are not able to in a two dimensional screen show you the three variables. But, if you assume that y is kind of coming out of the screen, you could graphically represented that way.

But, the important point that I wanted to make here is that if you needed to take like a five nearest neighbors approaches on two input variables, you can still do that let us say you are interested in this particular point marked with the x then your nearest neighbors to this x again get defined on the two dimensions. So, it could be this point, this point, perhaps this point and what you can see is we are taking some form of like may be Euclidean distance, because the distance itself is not just on the $x_1$ axis, it is not just on the $x_2$ axis, it is on both $x_1$ and $x_2$ axis.

So, you can still use the concept of the input space and once you have more than three year of when you have four or more input variables you are talking about hyper space. But, just you could still use some simple measures of Euclidean distance or some measure or some other distances some famous once I call them and Manhattan distance some of them are called the Mahalanobis distances, Mahalanobis distance as well.

Now, the good thing with this approach is that you can even use it for classification problems, if you briefly remember in the previous lecture we distinguish between

supervised learning tasks, which basically just means you have an output variable you need trying to predict that output variable from the input variables. But, this output variable can be continuous quantitative which is where we primarily talked about regression and so on. But, it could also a categorical variable.

o, if taken an example out here on the right hand side, where you have two input variables $x_1$ and $x_2$ are two input variables and your output variables is categorical variables with two classes. So, think of it as male or female or buyer or non buyer in marketing contexts, defective product, not defective product in a manufacturing context. So, let us say this output variables which is a categorical variable is represented by either square circles or squares.

So, the orange circle are one class of the output, the blue squares are another class of the output and $x_1$ and $x_2$ are just your two input variables again out here you might be interested in for instances making an prediction out here and you might take the five nearest neighbors perhaps it will be this and may be this. So, these might be the five nearest neighbors and because you need to predict, whether it will be class A or class B you might want to take a voting approach or if your approach is to predict the probability of it being circles or squares that is what I am call class A and class B then you might just take the ratio of the circles to squares in your nearest neighbors.

But, the idea is that this is also works perfectly well, you do not need to just take an average, you can take you know ratios or you can just make them all vote essentially the majority win. So, you have three squares and two circles, so I am going to predict this is going to be a square, you can using voting approach to say belongs to a particular class. So, I hope that gave you an idea of k nearest neighbors, but also more importantly motivated to you the idea that you have this regression style approaches, where you got this explicit data model.

So, this is your regression style approaches by you have a functional form and then you try and figure out the parameters. But, you can also take on very different approach to the process of predictive analytics, which is through the process of prediction, where the importance is not as much on the functional form which is cast and stone which is really an assumption you are making about the relationship between your input and output variables. But, it really goes beyond that right like you do not want to make those assumptions and you just want to take more algorithmic approach to this entire process.

You will get encounter a lot of machine learning techniques that we are going to be discussing later in this course really belonging to that class. So, I hope that gave you an idea both k nearest neighbors and this dichotomy that you be kind of seen machine learning.

Thank you.