Machine Learning for Soil and Crop Management
Professor. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology, Kharagpur
Lecture 06
Basics of Multivariate Data Analytics

Welcome friends to this week 2 of lectures of this NPTEL online certification course of Machine Learning for soil and crop management. And in this week 2, we are going to learn the basics of multivariate data analytics. And this is lecture 6.

And in our first week we have discussed, the basics of machine learning applications in agriculture. And we have seen what is Machine Learning and what is Artificial Intelligence and also what is the difference between Machine Learning, Artificial Intelligence and Deep Learning. We have seen their examples, their applications and some broad applications in agriculture crop and soil management.

Now, to understand the application of different machine learning algorithms and deep learning algorithms in agricultural problems, first you need to understand the principle. Now also for understanding the principle of different machine learning, specifically different classification clustering methods as well as the calibration or prediction methods, we are going to need the basic understanding of multivariate data handling.

So, in this week we are going to focus mainly on statistical approaches, and we will recall some of the some of our earlier knowledges, which we have already which we already know like basic data structure as well as how to deal what is the simple linear regression. And then we will go from there to multiple multivariate data and also we will discuss multiple linear regression and their diagnostic features.

So, in this, in this first a lecture of week 2, or in other words this is lecture 6 we are going to learn these following concepts. We are going to see what is multivariate data and also we are going to see the data matrix and scatterplot, what is scatterplot? And then we are going to recall some of the aspects of covariance and also we are going to discuss the simple linear regression and their different matrices.

So, these are the keywords for this lecture 6, we are going to see multivariate data covariance correlation, simple linear regression and also outlier.

So, let us start we know that in case of multivariate data that means there are multiple variables or multiple features. So, if you can see in this table it contains in the first column it contains the number of samples, that is the sample identification number sample 1 to sample 13. And also you can see there are 6 variables X1, X2, X3, X4, X5 and X6 and these are called the features, or variables. And we have the reading for, or values for all these 6 variables for all these 14 observations.

So, this is called a multivariate, this is how a multivariate data matrix looks like. So, you can see there is a matrix from in both row and columns and that is what is called a multivariate data matrix. And in multivariate data, we generally have more than one feature, so that is why it is called multivariate data.

Similarly, the basic difference between simple linear regression, multiple linear regression is in the simple linear regression there is one only one predictor, where whereas in case of multiple linear regression there are more than one predictors, we are going to discuss them in details.

So, this is how the multivariate data looks like and in soil science and crop science also the application of multivariate data is in a huge, because you know when we are talking about collecting the data from different sources, different sensor sources, they are not a single variable.

They are you know when we are trying to develop the crop yield model, we remember that in our first week we have discussed, that scientists are developing the crop yield prediction model based on the, based on the climatic data and the climatic data contains different types of environmental variables and these environmental variables more than one environmental variable. So, that is why they are multivariate data.

Similarly, for soil management also, we depend on different types of multivariate data we are going to discuss them in details.

Now, what is a multivariate data? If you see, if you see that data matrix it is a rectangular basically table, which is also called matrix. So, sometime we call it a spreadsheet. And it consists of n number of rows, and m number of columns, and each cell containing a numerical value as you can see here, there are n number of rows here the n is 14 and m is the number of columns. And here the m is 6 and each cell containing the you know some numerical value.

So, each row corresponds to a sample we know that we have already I have already told you and each column corresponds to a particular feature of the object, which we call variable. For instance you know a measurement on the object. So, variable, or feature is the measurement on the object.

So, we call this data matrix as a matrix X suppose and with element xij in row i and column j. So, if we consider this is row i, I can take any value from 1 to 13, actually there are 13 number of

samples starting from 2 to 14. So, I can take any value from a 1 to 13 and here also this j can take any value from 1 to 6, because there are 6 features.

So, any element in any of this cell can be represented by xij and that is in row i and column j. So, a column vector xj contains the values of variables j for all the objects. So, is this column vector you can see X1, X2, X3, X4, these are the column vectors. So, they can contain the values of variable j for all the objects. And a row vector, which is x transpose i is a transpose vector and contains all features of object i. Here i, generally varies from 1 to 13, whereas j varies from 1 to 6.

Now, in case of multivariate data, since it is multi-dimensional, multi-dimensional means when there are multiple features, in the in a data set that is called multi-dimensional data. So, multiple variables are multi feature, or multi-dimensional data can be represented in different dimensions. So, here you can see that x, y, and z, they are orthogonal to each other, and we are, these data can be projected in this multivariate data can be projected in this different dimension. These are individual dimension. So, multivariate data can be projected in this individual directions

Now, once we know, we know that what is multivariate data and what is the scatterplot and what is what is the matrix. So, and spreadsheet, now what is the scatterplot? Scatterplot is a very basic representation of the data. So, whenever you want to see any the behavior of any data set, it is always advisable to plot the data, be it a 2-dimensional be it a you know more than 2-dimensional, it is always recommended that you should plot the data to identify the basic linearity, or non-linearity among the data and also their interaction.

So, it is advisable for and it is, it is the first and foremost thing a data analyst should do that he or she should first plot the data in the scatterplot to see the feature of which is which is contained within the data.

So, now we have you know, we know that what is spreadsheet, what is multivariate data, what is also multi-dimensional data. Now, let us recall some of the concepts, because these concepts are very much important to understand the multivariate statistics, multivariate data analytics also. So, let us first recall the concept of covariance, you all know that covariance, you know has this formula of covariance of x and y.

So, in probability theory and statistics covariance is a measure of the joint variability of two random variables. So, here you can see there are two random variables x and y and the covariance calculation can be done by using this formula, which is summation of 1 to n xi minus X bar into yi minus Y bar.

So, if you subtract the mean of a variable from their individual values and multiply with the in you know this multiply with the difference of the individual values of the second variable and also they are mean and then take a sum and then you divide by the number of sample minus 1, then you will get the covariance.

So, in probability theory and statistics, basically it says the measures of joint variability you know where variance is a measure of variability of a single variable and covariance is a measure of variability of two random variable. Now, there are three types of values you can get in case of covariance measurement, one is covariance of x, y, if it is greater than 0. Then we can say that x and y are positively correlated and if the covariance of x, y is less than 0, then we can say that x and y are inversely correlated. And covariance of x, y, is 0, then they are called they are they are said that x and y are independent.

So, when the covariance of x and y is greater than 0, that means positive, then we can say that a positive change in x variable shows the positive change in y variable. Also in case of variance of x, y, if it is if it is negative that means a positive change of x variable can shows a negative

change in y variable. And when there is no relationship at all, then we can call that x and y are independent to each other. So, we know this thing.

So, now another important aspects, which we have to remember, or recall is the correlation. Now, you know correlation is basically the linear association between two quantity variables, when one variable increases or decreases at a fixed amount for a unit increase, or decreases in the other. So, basically the correlation indicates, the linear association, it is a type of association. So, in case of the linear association between two variables.

Now, what is the difference between correlation and regression? The difference between correlation regression is in case of correlation we want to see the linear association between two quantitative variables without any levels. However, in case of regression we want to estimate the best straight line to summarize the association.

So, that means in case of regression, there is an outcome and there is a input. So, we can clearly label, we should clearly label, what is the output, or dependent variable, or and what is the input, or independent variable. We will discuss about these in more in our coming slides, but at this point of time just remember, that in case of correlation there is no label, but in case of regression there are labels of output and input.

And so correlation coefficient, so how we measure the strength of correlation between two quantitative variables? It is measured through correlation coefficient, which also measures the degree of association. Now, it is generally denoted by this small r, sometime you will see that people are using capital R also, but generally it is small r, small r is denoted where small r generally indicates the correlation coefficient.

Now, Pearson correlation coefficient, which is a universally accepted correlation coefficient. It is basically standardized variance and also it is unit less. So, the formula of covariance of formula of correlation coefficient as you can see it is covariance of x, y divided by root of r of variance of x and variance of i.

So, if we expand this we can we know that the formula of covariance is xi minus X bar yi i minus Y bar summation of 1 to n by n minus 1, and the variance of x and variance of y, you know and the formula we have put. So, if we simplify it, we will get this nullify and ultimately we will get this is ultimately we will get this nullify. And so ultimately it will be 1, to equal to 1 to n xi minus x bar, yi minus y bar divided by n xi minus x bar whole squared and n yi minus y bar whole square.

So, if we simplify this thing, it will be sum square of x, y divided by root over of sum square of x into root over of sum square of y. So, this is the simplified form of this are you know the correlation coefficient. And so and this correlation coefficient helps in identifying, or measuring the degree of association.

Now, what are the different features of correlation coefficient? There are different features of the correlation coefficient, first of all it ranges between minus 1 to plus 1, it values ranges between minus 1 to plus 1. The closer to minus 1, the stronger the negative linear relationship. And the closer to plus 1, the stronger the positive linear relationship. And the closer to 0, the weaker any positive linear relationship.

So, you can see here, when the correlation coefficient is minus 1, then you can see that there is a strong linear negative linear relationship between x and y. So, when x is increasing, y is clearly decreasing. And also another extreme is when there is a positive perfect positive linear relationship, you can see here r value is plus 1. And when there is no relationship at all when

they are in the straight line, which are horizontal to the x axis, then you can see then it is also r equal to 0.

Another version of no correlation is given here, you can see r equal to 0, and the weak negative relationship and weak positive correlation are given, when the data are more scattered, but still showing some kind of negative and positive relationship. But you can see they are more scattered as compared to this, and here also this negative relationship is more scattered than this strong negative relationship. So, that indicates the strength of the positive, or negative relationship in the data set.

So, what are the other features of correlation coefficient? So, correlation coefficient does not depend on the labeling of x and y, of course I have already mentioned that, the labeling of x and y are not required for measuring the correlation coefficient. And r is independent of the units of measurement, the units of measurement does not play any important role in case of r. And r is sensitive to outliers. So, it is not resistant.

So, what is outlier? Outlier is an abnormally high, or low value in a data set and which sometime creates problem. So, r is an, so r is sensitive to the outlet. So, you can see here different types of correlation relationships are given. Here r equal to minus 0.95, minus 0.8, minus 0.5, as you can see from strong negative linear relation to the weak negative linear relationship, you can see the data is getting more and more scattered.

This is also here it is more scattered and you are getting you know less, or weak negative relationship. And here there is no relationship at all. So, r equal to 0. And then you can see gradually, how they are changing from negative relationship to positive relationships. So, here this is weak negative positive relationship, then it is more stronger positive relationship, then 0 point, further strong and positive relationship.

And then you can see very strong positive relationship. So, this is how we can see the gradual changes from strong negative relationship to the strong positive relationship.

So, whatever we are going to discuss, I mean we are going to now discuss the correlation and regression matrices and simple linear regression. So, we will be discussing based on a data set that is called Forbes boiling point of water data set. So, this data set basically shows, this is, this data set was created by James David Forbes in 1809, who was present in 18 from 1809 to 1868. And he was a Scottish physicist, who worked extensively on the conduction of heat, seismology and glaciology.

So, he collected data, that would help him to estimate the altitude, that is basically the height above the sea level from a simple measurement on the boiling point of water. So, from this data Forbes main interest was in the connection between the boiling point of the water and the barometric pressure. So, he wanted to make a relationship to between the barometric pressure and also the boiling point of the water.

So, Forbes theory suggested that, the logarithm of barometric pressure would be an approximately linear function of the boiling point of water over the range of boiling points, that he observed. So, this is the basic background of Forbes data set, which we are going to explore using correlation and regression.

So, if we define some of the important terms, like association. So, you can see two variables, measured on the same group of individuals are associated, if certain values of the first variable tend to occur at the same time as the certain value of the second. So, when there are relationship between the variable values, then they we recall that there are association. So, a response variable generally measures the outcome of the study, which is Y variable, outcome of the study, which is also known as the Y variable, which is also known as the dependent or endogenous variable.

Now, in case of Forbes data set of course the dependent or endogenous variable is what? In case of Forbes data set, the indigenous variable is the logarithmic of barometric pressure. So, and what is an explanatory variable? The explanatory variable is used to explain the causes changes in the response variable, that is X variable, or in other words the explanatory variable is an independent or exogenous or predictor variable. So, in any relationship if we want to predict y based on x, so here x is the explanatory variable or exogenous variable or indigenous independent variable, whereas y is the dependent or endogenous variable.

Now, in the Forbes data set, the explanatory variable is the boiling point of water. So, but sometime you know cannot classify this association cannot classify one variable as the response and the other as the explanatory variable, it is difficult sometimes. So, you have to have a clear knowledge of your objective and what you want to achieve, so based on that you can fix your endogenous and exogenous variable.

So, in the, as I have already mentioned in case of Forbes. Forbes data set the logarithmic barometric pressure is the dependent variable, whereas the boiling point of water is the independent or predictor or exogenous variable.

Now, suppose we have observation on two variable at a time, which are denoted by X1, Y1, X2, Y2, and then Xy, Xn, Yn. So, we already know that as scatterplot is used to investigating the possible relationship between two variables.

So, as I have told you that, if you do if you first see the data set in the scatterplot, then you will have a clear idea about the distribution of the data, whether there are some positive relationship, whether there is a positive, or negative correlation, you will have a basic idea. And then you can explore the data in more details.

But scatterplot is the first, I would say the starting point of any statistical analysis. And the correlation measures in some numerical the correlation, correlation generally measures in some numerical way, how two variables are linearly related. So, we have already know that and regression is used for explaining, or predicting one variable in terms of another. So, here in case of regression, there is a dependent variable and there is an independent variable and we want to predict the dependent variable based on the independent variable.

Now, so there is a software called SAS, which is a very widely used statistical software. Here these are the correlation, correlation results from the Forbes data set and you can see that first you know here it says that the L pre is basically the logarithm of the barometric pressure and these are the data lines. So, here the inputs in the data are where boiling point and also the pressure. And here we have they have converted into logarithm of the pressure, because there is a linear relationship.

So, once and then the data set was incorporated and then we are using this proc corr command to see the Pearson correlation coefficient. And here the variables are boiling point and also the logarithm of the pressure. So, in the Pearson correlation coefficient, you can see there are total number of 17 observations.

So, here you can see this is the correlation matrix and we can see of course bp versus bp, this is has to be this is this is the correlation Pearson correlation coefficient r equal to 1. Here the relationship between the logarithm of barometric pressure and bp, L pre and bp is 0.99, which is which is very significant p value is less than 0.001.

And here also L pre versus bp is 0.99 and here you can see that is L pre versus L pre equal to 1. So, that shows the that shows the correlation between correlation matrix or correlation between two variables.

So, scatterplot are, as the names as I have told couple of times, scatterplots are excellent for picking out the patterns in the data. So, if you see the scatterplot between the log of pressure and boiling point you can see there is a straight or linear relationship. So, scatterplot are excellent for picking out the patterns in the data. So, is there a, say and scatterplot gives us the answer for this type of question. Is there exist any simple linear relationship between a y and x? Linear is often the simplest. And when you go for non-linear that becomes complex.

So, are there gaps in the data, we can get the answer for there. Are there outliers in the data or not? Do we need to transform the data? Which variable goes on which axis? How strong is the relationship in the data? How close do the points follow the pattern in the data? And what direction is the association? Whether there is a positive association or negative association? So, we can we can get all the answer from the scatterplot. So, that is why I am again focusing on the importance of the scatterplot at the starting point of data analysis.

And what is outlier? Outlier is the in statistics an outlier is a data point that differs significantly from other observation. So, an outlier may be due to variability in the measurement, or it may indicate experimental error. So, here you can see in the first box plot, this is a box plot and here you can see that, this is an outlier and this box plot is showing this the third box is showing 4 outlets.

So, these are the data points that differ significantly from the other observations. And outlier generally occurred due to variability in the measurement, or maybe due to experimental error. And the latter are sometime excluded for the data set. So, if we I can identify these are the outlier due to experimental error, then you can then we can remove them from the data set for doing the further data analysis. And outlier can cause serious problem in statistical analysis.

So, this is very important an outlier, the presence of an outlier sometime causes very problems in the statistical analysis and influences the results. So, this is why it is very important to identify the outlier before doing any statistical analysis.

Guys let us wrap up our lecture here. And so we have learned the basic associations correlation regression and also we have learned about the types of correlation, positive correlation, negative correlation, Pearson's correlation. We have also learned about the outlier. And how outlier generally arise. So, we know we have learned about these things.

In the next lecture, we are going to discuss we will start from here and we are going to discuss the regression in more details. We will start with the linear regression, simple linear regression and will identify their perform their different features, like their slope and then what is the offset and how to calculate the slope and offset, and what are the performance metrics of a simple linear regression and so on so forth.

So, let us wrap up our lecture here and thank you for joining in this lecture. And we will learn more about this associations in our next lecture. Thank you.