Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 09

Lecture – 44

Introduction to Classification

Hello friends, welcome to the course Multivariate Procedure with R. So, from this lecture we are going to start a new topic which is Linear Discriminant Analysis LDA. So, the first question comes what is LDA, what is this linear discriminant analysis, why it is used and what are we going to learn in this topic. Suppose I give you a simple question to answer, there is one student who comes before you and you want to know whether this student has a background in mathematics or biology, what are you going to do. The simple solution is that you ask him some questions on mathematics, some questions on biology. And if you see that the person is able to answer more questions on mathematics and the number is significant larger than the correct answers given in the biology, then possibly you will think that okay this person has got a background in mathematics.

On the other hand, in case if the person is unable to answer the questions in mathematics, but the person or the student is giving more correct answer in the biology subject, then you would possibly consider the student as if the student has got a biology background. If you try to see in all this process, what are you going to do? You have some prior experience, prior experience that how a student of mathematics or how a student of biology should be. And based on that you have already collected some observation in your mind and based on that you have some idea about the probability that the student is going to be from biology background or mathematics background and based on that you have created a rule inside your mind. So, as soon as the student comes before you, you start asking the questions in each of the subject and based on that you try to fit this data into the rule which is inside your mind and you try to come up with the final conclusion that okay it appears that this student is from mathematics background or biology background, right.

Similar is the case when we try to see the report of a blood test analysis. Based on that the doctors always diagnose that the patient has got some disease or not or a particular type of infection or not. What they try to do? That first they try to take two groups of patients, one group of patient which has no disease and say another group of patient which has got some disease and then they try to take their blood samples and conduct the analysis based on certain values of the blood parameters. Then they try to see okay if this value is below then, below some threshold limit or this value is above some threshold limit, then there is a very high probability that the person has got certain disease or not. Now, a new person comes where the doctor does not know that the person has got the disease or not.

So, the blood test is conducted and based on the report of the blood test the doctor looks at the report and tries to classify whether the person has got a disease or not. So, this is basically the classification problem. And in this classification problem if you try to see there is a rule and whenever a new observation comes the values of those observations are substituted in that rule and we try to see whether the outcome is below or above certain threshold. And based on that we try to classify the observation into one of the population. Well, this can be extended to more than two populations also without any problem exactly on the same concepts.

So, in this chapter we are going to consider this classification problem. And we would like to see and we would like to understand that whatever we are thinking whatever we are doing in a very natural way the way our human mind is working, we would like to translate that rule in a mathematical sense. Why mathematical sense? Because once a rule is in the mathematical way then people do accept that decision very easily. And then all these concepts whatever we are thinking we will try to implement them through some mathematical analysis. For example, you would like to make a decision rule in such a way such that the probability of classification or the correct classification is as high as possible and the probability of misclassification is as minimum as possible that is what you are trying to do right.

So, all these things how to do them on the basis of a statistical tool is the objective in this chapter what we are going to achieve. So, we will try to take certain aspects, we will try to first create a mathematical framework, a statistical framework and then we will try to implement them in the R software. Well, I will try my best to keep the level of the mathematics as low as possible and there are many things which can be done more easily in the software. So, I will try my best to give you as much as possible the details and then I will try to illustrate the remaining aspect through the R software. So, with this objective let us begin our lecture and in this lecture I am just going to create a background and an overview of this linear discriminant analysis right ok.

So, let us begin our lecture. So now, first of all in this lecture we are going to talk about the introduction to classification which is going to be achieved by the linear discriminant analysis which is also called as LDA. L is coming from linear, D from discriminant and A from analysis right ok. So, the first question comes here why classification? So, try to consider a situation where we have large data sets and we want to dig out the information from such a data set and we need to classify and group the data into in the data set right. Then an investigator makes a number of measurement on an individual and the investigator wishes to classify the individual observation into one of the several categories on the basis of these observation.

Yeah means obviously, we are considering here the finite number of a categories right. So, now if you try to see when an individual comes and the observations are made then there is a chance that the person may belong to category 1, category 2 or category 3 likewise. So, we can consider the individual as a random observation from one of the population. The population is equivalent to different categories that means, this observation is coming from one of the categories right. For example, in the example of blood test we have created two populations, one population where the patients have no infection and other population where patients have got infection.

Now a new patient comes and that is supposed to be coming from one of these population and then based on the observation the person or the patient has to be classified as healthy or having some numerical issues right. So, the question is now that given an individual with certain measurements from which population did the observation arise? That means, given the result of the blood test we want to know whether the patient has to be classified into the patient having infection or patients not having infection. Similarly looking at the examination scores of mathematics and biology we would like to classify the student into one of the two population that one population is of student from biology background and another population is the population of student from mathematics background right. So, this classification scheme is a convenient method for organising a large data set so that the retrieval of information may be made easier. And if you try to understand this classification scheme actually describes the patterns of similarity and differences among the objects under investigation right.

By the course in the mathematics and biology subject we are trying to understand that how similar is the new observation to the students who are having a mathematical background or the students who are having the biological background right. And all the students they are suppose broadly classified into two categories and one category is has a student who have got a similar type of knowledge in the mathematics and another category where the similar type of student have got a biological biology background right. So, let me try to take one more example. Suppose there are two fields, fields means mining fields right where we get the ores for example, to dig out the iron ore there are certain fields to dig out the coal fields there are coal fields to dig out the coal etcetera. So, suppose there are two fields from where the iron ores are obtained and suppose these two fields are suppose here field number 1 and field number 2 where the ores for extracting the iron are obtained, but the ores have got different chemical properties right.

So, based on that we have taken some samples from F1 and F2 and we have conducted the analysis the chemical analysis that what is the chemical composition of the ores coming from the field F1 and what is the chemical compositions of the ore which are coming from field F2. Right and then we have some here results say results 1 and result 2. Now a new sample arrives right. Now this sample can be can belong to the field F1 or field F2 and we want to know that from where this sample is arriving right based on some criteria. So, this criteria is now that we try to do the chemical analysis and we try to obtain the values of those parameters in the chemical analysis which was done on the sample from field F1 and field F2 right.

So, we try to observe the chemical properties of the new sample and then we try to match that whether the chemical properties of the new sample they are matching with F1 or F2. And in case if they are matching to both in certain parameter then I would like to see the proportion of matching of the properties of the new sample with respect to F1 and with respect to F2. And then based on that wherever is the more proportion of matching we would like to classify this observation into one of that field and we will assume that the that this new sample is coming from that given field right. So, that is the same process which we have done in the case of students and the blood test also right means in the blood test also you had two patients which are healthy and which are not healthy. And then you have conducted the blood test report blood test on these two groups healthy and non healthy then you have got here some data that how the blood test report of healthy person will look like and how the blood test report of non healthy person will look like.

And now a new patient comes and you try to conduct the blood test. Now based on the properties of the blood test or the values of different parameters in the blood test of the new patient you would try to see whether this person has more matching with the values in the blood sample of a healthy person or a blood sample of a non healthy person and then you try to classify the new patient into one of the group healthy or not healthy right.

So, now in in all such cases if you try to see the decision is good if the sample is classified to the correct free or correct population. If it is classified to the wrong field for example, an ore is coming from field F1 and it is classified to F1 then it is a correct decision in case if the ore is coming from field F2 and it is classified to field F2 then it is a correct decision. But in case if the ore is coming from F1 and classified into F2 or vice versa that the more is coming from field F2 and classified into field F1 then it is a wrong decision and it and we call it as misclassified.

Similarly if a person is healthy and it is and a person is classified to the group of healthy patient it is correct and similarly unhealthy person is classified to the patient of unhealthy patients then it is correct, but if a healthy patient is classified to unhealthy population or unhealthy patient is classified to a healthy population of the patient then it is misclassification. Similarly, if a biology student is classified into mathematics or a mathematics student is classified into biology this is misclassification. Whereas, if the biology student is a categorised into the biology student population and if a mathematics student is classified into the mathematics student population then it is a correct decision. So, now how to identify and how to characterize these fields or this different type of population. So, each of this field or each of this category can be described by a probability distribution.

That is the most simple thing to do for a statistician that try to characterize the properties through a probability distribution, right. Now, from this probability distribution we can calculate the probabilities of classification and misclassification. So, a good classification scheme is one which provides a smaller probability of misclassification, right. So, now if you try to now consider the this problem of classification in a different way, then the problem of classification can also be considered as it in the framework of what testing of hypothesis, right. Mean each of the hypothesis is about the probability distribution of the observation given from a category.

For example, say here H0 is the some observation x belong to the PDF 1 and H1 is here x belongs to here PDF2, right. And then if H0 is accepted that mean the observation is belonging to population P1 and if H1 is accepted then the population kind of observation is belonging to the population P2, right. So, this is also achieved in the topic of classification analysis. Yeah, you can look into the actually books. I may not be covering it here, but definitely these things are available in all the standard textbooks.

So, our objective is now to find out a classification rule to accept the correct hypothesis and reject the other hypothesis. In the case when we are trying to view the classification problem from that testing of hypothesis framework point of view, right. Now, when we are trying to think about the classification approaches, then we have two types of analysis, one is called discriminant analysis and another is clustered analysis. Yes, after this idea we are going to study the topic of clustered analysis, right. So, what is the difference between the two? In both the cases we are trying to classify an observation.

In clustered analysis also you will try to see that we have a similar objective, but the main difference is that that in discriminant analysis we are going to develop a technique for grouping individual observations or the objects into some known group. That means, this population is known to us after IIT. For example, we knew that we have two population, one of a student coming from biology and say another student coming from mathematics. Similarly, we knew that we have two population, one population of patients who are healthy and those patients who are not healthy, right, with respect to certain disease. And similarly the poor fields, field number 1 and field number 2, they were known to us and we knew that this observation is going to belong to one of these categories.

So, these groups are known a priori in which the observation has to be classified in the case of discriminant analysis. And we will see that when we try to do the statistical analysis, then we are going to consider or find the linear discriminant function for classification, right. There can be a non-linear discriminant function also, but anyway we are going to consider here only the linear discriminant function. So, that is why it is called as linear discriminant analysis or say popularly it is known as LDA, right. On the other hand, when we consider the cluster analysis, then the cluster analysis is that it is technique for grouping the individual or objects into some unknown groups.

There is a huge population and when an individual comes, then we do not know that what are the different population which are existing inside the large particular place. So, the objective will be that first we have to create the population of the similar objects and then classify the observation into one of those clusters. So, a dataset that contains only data point is available and is without the class labels that we do not know whether this observation is going to belong to mathematics or biology. There can be more students also with different background, right. For example, if we have a big population of a student where we have students from the background of mathematics, biology, commerce, etcetera, then we do not know and we have no idea that whether there are this four categories of a student or only three categories of a student that is also even we do not know.

So, under those situations we try to go for cluster analysis, but anyway we will try to take up this topic in more detail later on, right. So, for example, in the case of classification approaches, let me try to take some more example. For example, these are some very popular things which are happening in a day-to-day basis. For example, we always try to classify the data on economic status in group. For example, lower income group, middle income group, upper income group or say lower class, middle class, upper class.

So, these are some known groups and we try to assign an individual into one of these groups. So, this is called as discriminant analysis, whereas the diseases which look similar can have different causes and conversely different diseases from the same cause can also be there and we do not know. So, to understand the treatment we try to classify the symptoms to diagnose the disease and these are our unknown groups that what are the different groups due to which a particular disease can be an upper we are we do not know. So, this is under the view of cluster analysis, right. So, the problem of classification arises when an investigator makes a number of measurement on an individual and wishes to classify the individual into one of the several categories on the basis of these measurements, right.

So, everything is based on the measurement which the investigator is going to make, right. So, we are classifying just the population with known people and then we are classifying an unknown person into one of those two population, right. And the problem is this that the investigator cannot identify the individual with the category directly, but the person has to use certain measurements to make a decision that whether the person belongs to the category 1, category 2 or like that, right. So, we are assuming here that we are going to deal only with the finite number of categories or there are only finite number of population from which the individual may have come. The only thing is this we do not know which population from which population the observation is coming and we are assuming that each population is characterized by a probability distribution of these measurements and we will consider an individual as a random observation from any of this population.

And now the question which we want to answer is that given an individual with certain measurements from which population did the observation arise, right or given an individual from which population can this observation be classified, right. In some instances the categories are specified beforehand in the sense that the probability distribution of the measurements are assumed to be completely known, right. And in some cases we do not have this complete information, the form of each distribution may be known for example, but the parameters of the distribution are not known and so in such cases they must be estimated from a sample which is arising from that population, right. For example, if I say if I assume that the population is normal mu sigma square and suppose if the value of sigma square is known suppose and we do not know the value of mu then we have to take a sample and then we have to estimate mu. Now you know that sample mean is an unbiased estimator of population mean, so then x bar sample mean can be used in the place of population mean and the same thing can happen to the case if mu and sigma square both are unknown, both can be estimated on the basis of given sample of data and they can be replaced back to the place where we want to use these values, right.

For example, in case if I try to illustrate that how are you going to do it, suppose there are some prospective students who want to apply for admission in a college and this has to be and this has to be based on an examination. So they are given a battery of test, battery of test you can say that certain question for example, whenever we go to appear in any entrance examination we are given a say about say this 100 objective question or 10 question like that and we have to answer them based on that we are going to be classified as selected or not selected, right. And then there can be various aspects on which these batteries can be given. For example, so finally what will happen that we will get here the vector of scores which are the set of measurements say for example x. And now those prospective students they may be a member of one population consisting of those students who will successfully complete college training or have potential for successful completing the training program or the student may be member of other population those who will not complete the college course successfully.

So that is really happening in a practice that why do we try to conduct the entrance examination in a college. We always want to test whether the student can complete the degree from that college or not and based on the level of education they try to classify the classify different sets of students as classified or say as classified into who are selected and classify them as who are not selected, right. So the problem is to classify a student applying for admission on the basis of his scores on the entrance examination, right. You always say that cut off is say this 90% that means any student who is getting more than 90% of marks is eligible to take the admission in the college or if the cut off is lower than 90 that means all those student who have got less than 90% marks they will not be given the admission in the college, right. So some in other cases the form of each distribution may be known but the parameters of the distribution marks we estimated from a sample from that population, right.

So in many cases you know that we do not know that what should be the cut off but we simply try to say okay means top say this 20% quantized student can be accommodated. So we always try to find out the quantile of that distribution and we say that okay all the students who have got the marks more than the 80th quantile they will be admitted and all those student who have got the marks like more than the 80th quantile they will not be admitted. So now this 80th quantile can be any number depending on that how many marks all the students have got during the examination or in the examination, right. So now in constructing a procedure for the classification it is desired to minimise the probability of misclassification or more specifically I can say it is desired to minimise the average on the average the bad effects of misclassification, right. And we want to maximize the probability of classification and also obviously the total probability can be divided into two parts, probability of classification and probability of misclassification.

So if probability of misclassification is minimized obviously the probability of classification will be maximized, right and so on. So some important assumptions in the discriminant analysis what we are going to understand here are the cases must belong to one and only one group or in other words the group must be mutually exclusive, right. You cannot say that an individual observation is belonging to population 1 as well as population 2. The number of cases for each group must not be greatly different, right. Means you cannot say that one population has only 5 observations and say another population has 500 observations.

Here the number of cases in all the population should nearly be the same or there should be not a say high differences, right. And the cases must be independent. Discriminant function performs better as sample size increases. If you try to increase the sample size you will get a better reply or more dependable reply. But that is a statistical process that if you try to increase the sample size we expect that in a good statistical process the decision will become better.

So a good guideline is that there should be at least 4 times as many as sample as there are independent variables, right. Because it is like that suppose if you say that you have got 5 variables on which you are trying to take the observation and then you have only suppose 6 observations. That will not actually work. Well, that is an empirical rule what I have told you that okay if you have more number of variables then obviously you should have more number of observations also. Otherwise the estimation of parameter becomes a challenge.

You can estimate them but the quality of the estimates may not be good, right. And this discriminant analysis what we are going to consider is highly specific to the outliers that if you have a very high extreme value in the dataset then this linear discriminant function will deviate, right, a lot. And each group should have the same variance for any independent variable. That we do not expect that the variances of different population are varying a lot, right. You can have different populations with different variability but then there will be a loss in efficiency and but if you are getting the sample from different population which have got the same variability then the results will be good.

So the first option is this. If you are getting different variances try to use some transformation to make them equal. Otherwise we have the procedures which can be used when we have different variances, right. So the and the independent variable should be multivariate normal. This is a good assumption but I will try to show you that we have procedures here which do not depend on the multivariate normal distribution also. But if you want to extend this discriminant analysis to a test of hypothesis other aspect then you will need this multivariate normal distribution.

So it is a better option that if you can assure that the normality is achieved in the observations. But if not then there is absolutely no issue. We have Bayes procedure rule which does not depend on the form of the distribution also, right. But the good part is this. The form of the linear discriminant function will come out to be the same whether the normality is assumed or not, right.

And I can say it was similar to the ordinary least square estimator in the case of linear regression modelling where the ordinary least square estimator and maximum likelihood estimator had the same form beta hat is equal to x transpose x whole inverse x transpose y. But later on you said that okay if you want to go for confidence interval estimation and test of hypothesis then you need to assume a probability distribution. So in the case of least square estimation although in the beginning we have not assumed any probability distribution with the random errors but in the later stage we assume them to be normally distributed. So similar is the story here also, right. So now with this we come to an end to this lecture and you can see here that this was basically a story telling type of a lecture but it was important in my opinion because when we are trying to start a new topic it is very important for us to have an overview so that from the next lecture when we are trying to do anything from the statistical or mathematical framework then we know why are we doing it.

Now if I try to show you that how are you going to calculate the probability of classification then at least you know that why you are doing it. You are doing it because you want to minimize the probability of misclassification and then whatever analysis I am going to do you will understand that why I am doing it because I want to make here a rule. I want to formulate here a mathematical formula based on which we can do the classification and what are we going to achieve that is also known to you that first we will have population with some known features. We will try to create a rule and then we will try to make a decision rule that if an unknown observation comes then how it is going to be classified into one of these population. So that is why this lecture was important but definitely we are going to use different concepts like matrix theory, multivariate, normal distribution etc. their estimator.

So it is good if you have a quick review of those lectures which we have done in the past. So you come prepared with these things and I will see you in the next lecture and I will continue with the linear discriminant analysis. Till then goodbye. Thank you.