Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 10

Lecture – 45

Bayes Procedure for Classification

Hello friends, welcome to the course Multivariate Procedure with R. So, you can recall that in the last lecture we started a new topic Linear Discriminant Analysis LDA and we had discussed an overview of what are we going to really achieve and what do we want to do in this topic and we had introduced the concept of classification problem. So, we are essentially going to classify an unknown observation into one of the known populations and based on different types of criteria. So, we have discussed different issues like as probability of classification, probability of misclassification and now I believe that you have a fair idea in your mind that what are we going to do and what we want to achieve, right. Means I always say that it is easier in life to work when we know what we have to achieve, right. So, now we know that we have to develop the statistical methodology using the mathematical tools to devise a classification rule.

So, in this lecture we are going to talk about the base procedure. So, base procedure will also give us a mathematical rule for classifying an observation into one of that population. And in this procedure you will observe that we are not assuming any particular form of the probability distribution of the populations, but definitely means there has to be some probability distribution, but I am not saying whether it is normal binomial position or something else. Then in the next lecture I will try to take the assumption that my samples or my populations are multivariate normal and my samples are coming from a multivariate normal population.

So, let us begin to understand this base procedure and we try to understand that how are we going to compute different types of probabilities and based on that we will try to classify the observation into one of the known population. And I believe that you have the primary concept that how to compute different types of probabilities using probability density function and the conditional probability. So, these are very elementary concept which are required to understand this whole methodology. So, let us begin our lecture and I will try to explain you how you can develop the base procedure, okay. So, now in this lecture we are going to talk about base procedure for classifications, right.

So, just for the sake of understanding and convenience we shall now consider the case of only two categories, right. This can be extended to more than two categories also. So, suppose we have an observation on say p variables, say X1, X2, X3 and this observation is obtained here as say x1, x2, x3 like this which is a multivariate in nature, multivariate observation on p variables, right. For example, if I say the same example which I considered in the last lecture that a person has got a black test where there are suppose five parameters on which those values have been obtained and but there is only one observation, one multivariate observation on that particular issue. And this has to be classified into one of the two population whether the person is healthy or not, right.

So, for example, in such a cases the first observation x1 can be on height, second observation x2 can be on weight and so on up to suppose the pth observation on xp is on the h, right. So, because we are now considering only two populations, suppose those two populations are denoted by capital P1 and capital P2 and we assume that this observation is arising from any of this population P1 or P2. And because we are getting the observation on p variables, so we can think that the observation as a point in a p dimensional space and so we would like to divide this p dimensional space into two regions that if the observation falls in the region R1, then we classify it coming from population P1. And if the observation falls in the regions R2, then we classify it as coming from the population P2. So, this region is now divided into only two mutually exclusive regions R1 and R2 and the observation has to belong either to this R1 or to R2, right.

So, now when we are trying to observe this process here, you can see here in case if the observation is coming from R1, it is classified to R1, this is correct. If the observation is coming from this R2 and it is classified to R2, then it is correct. But in case if observation is coming from R1 and classified to here R2, then it is incorrect. And if the observation is coming from this R2 and classified to R1, that is incorrect. So, these are the four possibilities.

So, based on that, there are two possibilities for the correct classification and two possibilities for the incorrect classification. So, if we consider the correct classification, then the first possibility is that the observation is coming from the population P1, that means the observation belong to the region R1 and it is classified within the R1. And similarly, if the observation is coming from the population P2 and it is classified into R2 region, so this is correct. Now there are two errors that will arise due to the wrong classification. So, these two types of errors can happen in case the observation is coming from the population P1 and it is classified into R2 or the observation is coming from population P2 and it is classified it in R1.

So, there are two options or two possibilities for the correct classification and two possibilities for the incorrect classification. Now the question is that how to quantify them. So, we try to introduce here the concept of now this probability. And we try to see that, we try to compute different types of probabilities on the basis of the given data. It is like that suppose if a patient enters into the room of the doctor and the physical condition of the patient is not good, the patient is coughing a lot, sneezing a lot and having problem in breathing, would you classify the patient as healthy? Certainly not.

Because the doctor has calculated the probability whether the person is healthy or not based on the past experience and based on that the doctor has already classified, okay this person is not healthy and the doctor has to attend the as soon as possible. So, now we are going to this idea and we are going to implement it in steps. So, suppose we indicate by this small q1 the probability that the observation is coming from P1, right. And q2 is the probability that the observation is coming from the second population P2. Right, so we can write it that probability that x is belonging to P1 is q1 and probability that observation x belongs to P2 is q2.

So obviously q1 + q2 is equal to 1 because we have only here two possibilities. So, it is something like this that when the patient with all these symptoms is entering into the room of the doctor, then doctor tries to see what is the probability that the person belongs to a healthy patient group or a non-healthy patient group. So, these probabilities are like q1 and q2. So we are assuming in the beginning that these prior probabilities q1 and q2 are known to us, right. And these probabilities are described by the probability density function.

So definitely when you are talking of probability then how the probabilities are distributed over different ranges of the values of random variable that is described by the probability density function. And then obviously when you are trying to misclassify an observation into any group there are always some losses and losses are indicated by the costs associated with both types of misclassification, right. For example, in case if I say that there is a student and there are two possibilities that the student is suppose here good and student is bad. Now the student appears in the entrance examination for a medical test, right. And this is the result outcome of the medical examination that the student gets collected and student does not get collected.

So if the student is good and if the student gets collected there is no issue, that is a good decision rule. If the student is bad and does not get selected this is also a correct decision. But now if you try to see if the student is bad and the student gets selected this is here an error. And on the other hand if a student is good and not getting selected this is also an error. But if you try to see the consequences that a bad student becoming a doctor and a good student becoming not a doctor do you think the consequences are same? No, they are different.

Possibly in my personal opinion a bad student becoming a doctor is more serious than good student not becoming a doctor because the good student can become something else also and he can have a, he or she can have a good career. So now in this case the cost or the loss due to misclassification do you think are they equal? No, they are certainly different, right. So now I will try to give you here two aspects that just for the sake of simplicity so that I can explain you how to develop the classification rule easily I am assuming this cost of misclassification or the loss due to misclassification to be equal and at the end I will try to extend this analysis to the case when the cost or the losses are unequal, right. So now in the beginning I am assuming that the cost associated with the both types of misclassification are the same or equal. So now we have to decide that how are we going to define the good classification procedure.

So a good classification procedure is the one which minimizes these probabilities of misclassification. And the basic idea about this classification is that we try to compute the probability that X is coming from the population P1 and population P2. Now the rule is very simple. If probability that X is coming from the population P1 is higher than the probability of X coming from the second population P2 then we are going to classifying the unknown observation X to P1 otherwise to P2, right. So that is what we are trying to do that if you get a student to whom you want to classify as having a mathematics background or a biology background you simply try to take some test on mathematics and biology and based on that you see if the scores in mathematics are much much higher than the scores in the biology then you decide that okay probability that this student belong to the mathematics background is much much higher than the probability that the student belongs to the biology group, right.

And then you try to classify the student into the mathematics group and vice versa, right. So now let us try to understand and try to formulate this probability in a probability density framework. So first let us try to specify the probabilities of correct classifications. So probability that an observation is classified in P1 that is actually drawn from P1 is indicated by this symbol, probability inside parenthesis 1 given 1 then here R. And suppose if P1 X is the PDF then this probability can be computed over the region R1 with the PDF of this P1, right.

$$\boldsymbol{P_1 \text{ that is drawn from } P_1 = P(1|1,R) = \int_{R_1} P_1(\underline{x})d\underline{x}}$$

$$\boldsymbol{P_2 \text{ that is drawn from } P_2 = P(2|2,R) = \int_{R_2} P_2(\underline{x})d\underline{x}}$$

That is how we try to compute the probability if you try to recall in the lectures in the beginning I had introduced it. Similarly, probability that an observation is classified into the second population P2 that is actually drawn from the same population P2 is indicated by probability 2 given 2, R. So this is my symbolic representation actually so that our this analysis will look simpler and easier, right. And this probability can be computed by using the probability density function of P2 and integrating it over the region R2. Well, I am assuming here that my random variables are continuous but without loss of generality they can be discrete also.

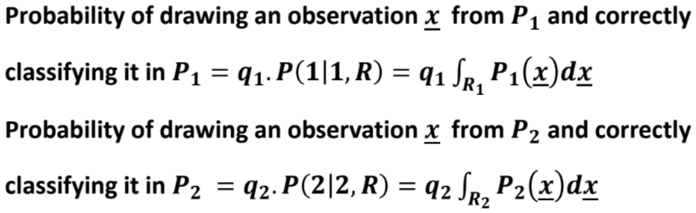
But doing the same thing for discrete and continuous is a time consuming. So that is why I am considering here this continuous case only, right. Similarly we can also find out the probabilities of incorrect classification. So the probability that an observation is classified into P1 and that is actually coming from the population P2 is indicated by this symbol P1 given 2, R which is computed from the PDF of P2 and integrating over the region R1. R1 is the region for P1, right.

$$\boldsymbol{P_1 \text{ that is drawn from } P_2 = P(1|2,R) = \int_{R_1} P_2(\underline{x})d\underline{x}}$$

$$\boldsymbol{P_2 \text{ that is drawn from } P_1 = P(2|1,R) = \int_{R_2} P_1(\underline{x})d\underline{x}}$$

And similarly the probability of misclassifying an observation into P2 that is coming from P1 is similarly indicated by P2 given 1, R which is obtained as a integration over the PDF of say P1 and integrated over the region R2, right. Now we would like to extend this definition of the probabilities and we would like to compute the probability that if we know the prior probability and based on that we would try to compute the probability that an observation is coming from particular population and it is correctly or when correctly classified into a population. So we have here 4 different types of probabilities which are based on these 2 correct classification and 2 incorrect classification. So they are like this, probability of drawing an observation x from P1 and correctly classifying it into P1 that can be written as say Q1 into P1 given 1, R and this can be computed here like this. I simply have just substituted this P11, R here, right.

And similarly the probability of drawing an observation x from P2 and correctly classifying it into P2 is simply here Q2 into P2 given 2, R which is here something Q2 into this quantity is P2 given 2, R that I just showed you in the earlier slide. Similarly when we go to the incorrect classification, the probability of drawing an observation x from P1 and incorrectly classifying it into P2 that is Q1 into P2 given 1, R and which is computed as Q1 into integral over R to P2 x dx. And similarly the probability of drawing an observation x from P2 and incorrectly classifying it into P1 is Q2 into P1 given 2, R which is here Q2 into this probability of P1 x and integrate over the region R1. So this is how you can compute these different types of probabilities, right. Now what is your problem? Your problem is that an unknown observation comes and then you have to classify it into one of these region.



Probability of drawing an observation $\underline{x}$ from P_1 and correctly classifying it in $P_1 = q_1 . P(1|1, R) = q_1 \int_{R_1} P_1(\underline{x}) d\underline{x}$

Probability of drawing an observation $\underline{x}$ from P_2 and correctly classifying it in $P_2 = q_2 . P(2|2, R) = q_2 \int_{R_2} P_2(\underline{x}) d\underline{x}$

So whenever you are doing it, there is always some cost of misclassification and based on this cost of misclassification, there is going to be some loss. So this expected loss is given here like this that assuming that the cost due to the misclassification are equal, we divide the region R into two parts R1 and R2 such that the expected loss is given by this Q1 into P2 given 1, R + Q2 into P1 given 2, R that is the expected loss and we would like to divide this region R into R1 and R2 such that this expected loss is as small as possible, right. And this is also the probability of misclassification which is to be minimized because if you try to see here, these are here like this. These are the two probabilities of misclassification, right. There will be some cost factor here because we are assuming the cost due to misclassification to be equal, so that is why they are not appearing here.

At the end, I will try to extend this loss function to the case when we have different cost due to misclassification, right. So now then we have certain mathematical procedures and we try to conduct a statistical analysis and based on that the procedure that minimize this expected loss, this expected loss is called the Bayes procedure and the average expected loss is called the Bayes risk, right. So the procedure which will minimize this expected loss that is called as Bayes procedure and the resulting expected loss is called as Bayes risk. So here we are assuming that Q1 and Q2 are known to us but in case if Q1 and Q2 are known to us, then these they can be obtained and then we use the minimax principle also. But in the software you will see later on that the software computes this probabilities, right.

So now our issue is we have got an observation X and we would like to classify it into either R1 or into R2 knowing the prior probabilities Q1 and Q2 such that this expected loss is minimax. So now for that we need to find out suppose the probabilities that probabilities is that the observation is coming from say population P1 or P2. So for that we use here the concept of conditional probability and we try to use the Bayes theorem. You may recall that there is a way to compute the Bayes probabilities also. So we are going to find out the conditional probability using the Bayes theorem.

So the conditional probability that an observation arises from the population P1 given an observation X is obtained using the Bayes rule of probability which is here like this. Probability that X belong to P1 given X equal to X, this can be written as the joint probability of X equal to X and X belong to the population P1 divided by total probability that X is equal to X. Now this probability in the denominator can be written as say probability of joint probability of X equal to X and X belong to P1 into probability that X belongs to P1 + joint probability that probability that X is equal to X given P2 and probability that X belongs to P2. And then the probability in the numerator can also be written here like this that the conditional probability X equal to X given X belongs to P1 into probability that X belongs to P1. Now if you try to substitute these values over here then it will come out to be here that now this probability can be expressed here like this.

P1 X into Q1 divided by P1 X into Q1 + P2 X into Q2. Right, so this is the probability that an observation is arising from P1 given that there are two possibilities that it can come from P1 or P2. So the same concept if I try to repeat and I try to find out the conditional probability that an observation is arising from the second population P2 given observation X is again obtained using the base rule of probability as probability that X belongs to P2 given X equal to X is equal to P2 X into Q2 divided by the same quantity P1 X into Q1 + P2 X into Q2. So now if you try to see you have got here the probability of X that is belonging to P1 or P2. Right, so now what you have to do? You need to assign this observation to a population wherever this probability is higher.

$$\frac{\boldsymbol{P_1(\underline{x}).q_1}}{\boldsymbol{P_1(\underline{x}).q_1 + P_2(\underline{x}).q_2}} \geq \frac{\boldsymbol{P_2(\underline{x}).q_2}}{\boldsymbol{P_1(\underline{x}).q_1 + P_2(\underline{x}).q_2}}$$

For example, the doctor knows from the past experience that if a person is healthy usually they will not be sneezing too much, they will not be coughing too much etc. And now when a patient enters into the room of the doctor and if the patient is sneezing a lot, coughing a lot then from the past experience doctor decide without even examining the patient that this person, that this patient belongs to the category of unhealthy patients. So similar rule we try to do here, we try to compute both the probabilities for this given observation X and then we try to minimize the probability of misclassification by assigning the observation to a population that has higher conditional probability. So if you try to assign it to a population where the probability of misclassification is lower then obviously the loss will be more. So now I can say that here I have computed here both the conditional probability that X belongs to P1 and X belongs to P2 and we try to compare these two probabilities.

$$\underline{x} \in \boldsymbol{P_1} \text{ if } \boldsymbol{R_1}: \boldsymbol{q_1}.\boldsymbol{P_1}(\underline{x}) \geq \boldsymbol{q_2}.\boldsymbol{P_2}(\underline{x}) \quad \text{or} \quad \frac{q_1}{q_2} \geq \frac{P_2(\underline{x})}{P_1(\underline{x})}$$

$$\underline{x} \in \boldsymbol{P_2} \text{ if } \boldsymbol{R_2}: \boldsymbol{q_1}.\boldsymbol{P_1}(\underline{x}) < \boldsymbol{q_2}.\boldsymbol{P_2}(\underline{x}) \quad \text{or} \quad \frac{q_1}{q_2} < \frac{P_2(\underline{x})}{P_1(\underline{x})}$$

Suppose if this probability is greater than or equal to this probability, so you can see here this denominator is the same in both the cases, so I can write down here very simply that P1 X into Q1 is greater than or equal to P2 X into Q2. If this is happening then we try to assign the observation X to population P1, otherwise we assign it to the population P2. So you can see here it is a very simple rule, right. And now I can formulate my here this phase rule to classify an observation X when the prior probabilities Q1 and Q2 are known like this. That this observation belong to P1 if the region R1 is like that, that Q1 into P1 X is greater than or equal to Q2 into P2 X or this can be written here as say if you try to bring this Q2 here and P1 X on the right hand side, so I can write down here more simply that Q1 upon Q2 is greater than or equal to P2 X upon P1 X.

And X belongs to here say P2, if the region R2 is like that Q1 into P1 X is less than Q2 into P2 X or I will say that here Q1 upon Q2 is less than P2 X upon P1 X, right. And you have to observe that, you have to notice that R2 is a complement of R1. And in case if Q1 P1 X is equal to Q2 P2 X, then the observation X can be classified to either of the population P1 or P2 without any problem. And statistically it can be proved that the procedure is the best procedure which minimizes the average expected loss this thing. We are not going to give you here the proof but it is available in the books and but I would like to inform you that this decision rule what you have obtained is this is the best decision rule in the sense that it is trying to minimize the average expected loss, right.

Okay, now I try to give you with idea that if the cost of misclassifications are different. So we indicate here the cost of misclassification as C1 given 2 which is the cost of misclassifying the observation from population P1 into population P2. And obviously this is going to be greater than 0 and similarly C2 given 1 is the cost of misclassifying the observation from second population P2 into the first population P1 and it is also greater than 0, cost is always greater than 0. So if the cost of misclassification are unequal then the base rule to classify an observation x when the prior probabilities Q1 and Q2 are known is based on the minimization of the expected loss function which is now here modified here like this. You can see here now this C21 and C12 are introduced in the earlier defined loss function.

So, this is now here the modified loss function where this cost of misclassification are unequal. And if you try to compute similar type of probabilities and finally the base rule will come out to be here like this that if the cost of misclassification are unequal then the base rule to classify an observation x when the prior probabilities Q1 and Q2 are known is modified as this x belong to P1. If R1 is modified as you have Q1 x into C2 given 1, Q2 is greater than or equal to Q2 into P2 x T1 given 2 and this can be modified here as Q1 upon T2 is greater than or equal to P2 x into C12 divided by P1 x into C21. So, you can see here that if this cost are equal then they will cancel, right. And similarly if x belongs to the second population P2 then the region R2 is obtained here say Q1 into P1 x into C21 is less than Q2 into P2 x into C12 or Q1 upon Q2 is less than P2 x into C12 upon P1 x into C21, right.

So, you can see here also it is this cost are equal to the right. So, now we come to an end to this lecture here and you can see here in this lecture we have given you a statistical procedure that how this classification rules are obtained, right. And we have used the simple concept which you use to say use inside your mind now we have transformed into a mathematical framework and using our statistical analysis now we have a clear cut rule and which is based on the basic assumption that our prior probabilities are known to us. We have not assumed any particular form of the probability distribution but we have given the result in terms of they say P1 and P2. Now this P1 and P2 can be normal or can be binomial, it can be multinomial or anything else.

So, then based on that you can always compute such probabilities and then you can devise the classification rule, right. Now, the next question comes here that how to means extend it for a particular distribution. So, now we have seen that multivariate normal distribution is a very popular distribution. So, in the next lecture we will try to extend this thing these details to a multivariate normal population. And the way I have done it this is called the base procedure but even if you try to go for a likelihood principle also then also similar type of result is obtained.

So, all those things we will try to discuss in the forthcoming lecture and after that I will try to show you that how you can implement these procedures in the R software. So, it is now your turn please try to look into some books and if you have a reasonable background in mathematics please try to spend some time and try to see that how these procedures have been derived and how they have been proved that they are the best procedures. So, you try to practice it and I will see you in the next lecture till then goodbye. Thank you.