

Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

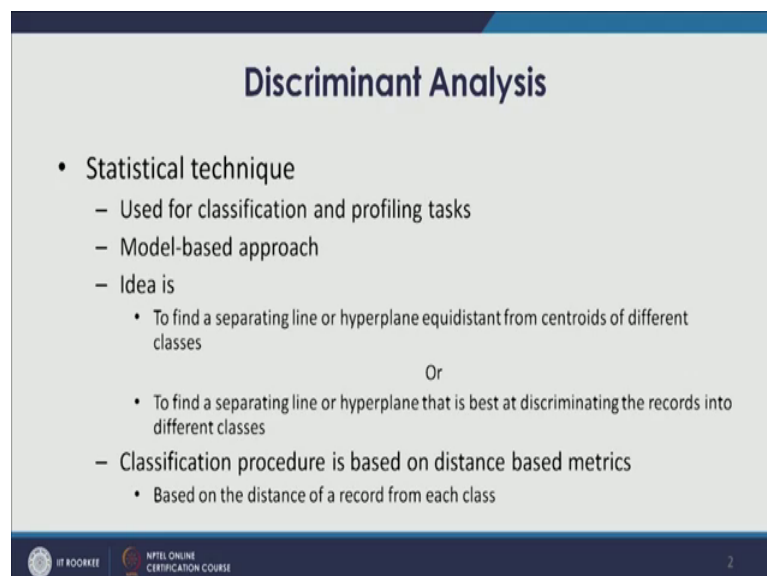
Lecture – 59
Discriminant Analysis-Part 1

Welcome to the course Business Analytics and Data Mining Modeling using R. So, we have come to our last technique that we want to discuss that is the discriminant analysis. So, let us start. So, discriminant analysis is also a statistical technique, typically used for classification or profiling tasks. So, application and the different types of tasks that it could be used are quite similar to what we discussed for a logistic regression.

So, this is also a model based approach. So, typically we make assumptions about the structure of relationship between outcome variable and set of predictors. So model based approach; if we look at the main idea behind discriminant analysis. So, we can see two points; so in two approaches have been used to conceptualize the idea of discriminant analysis.

First one is to find a separating line or hyperplane equidistance from centroids of different classes So, by this you can also understand main application mainly discriminant analysis is used for classification tasks. So, let us read it again to find a separating line or hyperplane equidistance from centroids or different classes.

(Refer Slide Time: 01:47)



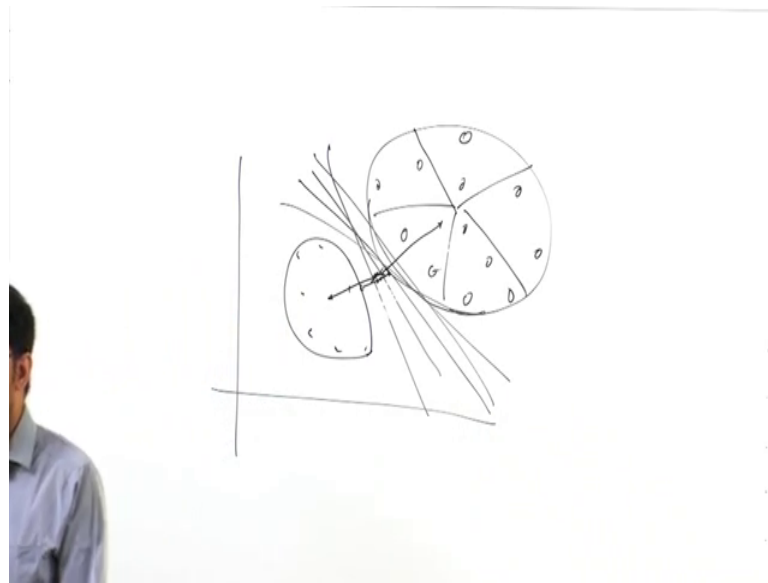
Discriminant Analysis

- Statistical technique
 - Used for classification and profiling tasks
 - Model-based approach
 - Idea is
 - To find a separating line or hyperplane equidistant from centroids of different classes
 - Or
 - To find a separating line or hyperplane that is best at discriminating the records into different classes
 - Classification procedure is based on distance based metrics
 - Based on the distance of a record from each class

 IIT ROORKEE  NPTEL ONLINE CERTIFICATION COURSE

2

(Refer Slide Time: 01:54)



So, from this we can understand that we are talking about when we discuss discriminant analysis we are essentially looking for separating line.

So, this particular technique is about finding a separating line which is equidistance from centroids of different classes. So, let us say centroid for this particular group is this one and centroid for this particular group is this one. So, probably we are looking for a line which is equidistant from the centroid of different classes.

So, sometimes this approach is used to implement discriminant analysis. The another approach is to find a separating line or hyperplane that is best at discriminating the records in two different classes. So, that is another approach. So, in terms of theoretical understanding, there is not much of a difference; however, when you go about implementing it using software, implementing the actual steps of algorithm then there is slight differences.

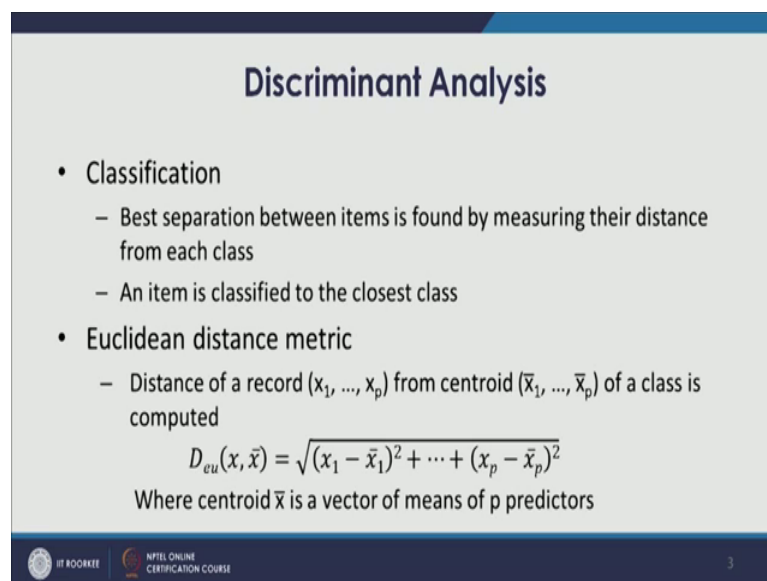
So, the second approach is to find a separating line or hyperplane that is best at discriminating the records in two different classes. So, in that sense we are looking for a particular line or hyperplane that might be best at discriminating these cloud of observations right into their respective groups right. So, these are the two approaches; first one finding equidistance or that finding an equidistant line or hyperplane.

So, that would of course, because that is equidistance from centroids of these groups. So, probably that would do a good job of classification, the second one is finding a line that would be best at discriminating the records in two different classes. So, these two approaches are popular in discriminant analysis and have been implemented.

So, the classification procedure that is used here is based on distance based matrix. So, few distance based matrix we have covered when we discussed (Refer Time: 04:21) and so this particular technique is also based on distance based matrix. So, the main idea is again the based on the distance of a record from each class. So, as you can understand from the two approaches that we discussed the underlying computation calculation that would be required and each of those approaches would be calculation of distance of a record from each class.

So, this calculation then becomes the basis for classifying observation. So, let us move forward, so classification few more points are there.

(Refer Slide Time: 05:00)



Discriminant Analysis

- Classification
 - Best separation between items is found by measuring their distance from each class
 - An item is classified to the closest class
- Euclidean distance metric
 - Distance of a record $(x_1, \dots, x_p)$ from centroid $(\bar{x}_1, \dots, \bar{x}_p)$ of a class is computed
$$D_{eu}(x, \bar{x}) = \sqrt{(x_1 - \bar{x}_1)^2 + \dots + (x_p - \bar{x}_p)^2}$$
Where centroid $\bar{x}$ is a vector of means of p predictors

IT ROOKIE | NPTEL ONLINE CERTIFICATION COURSE | 3

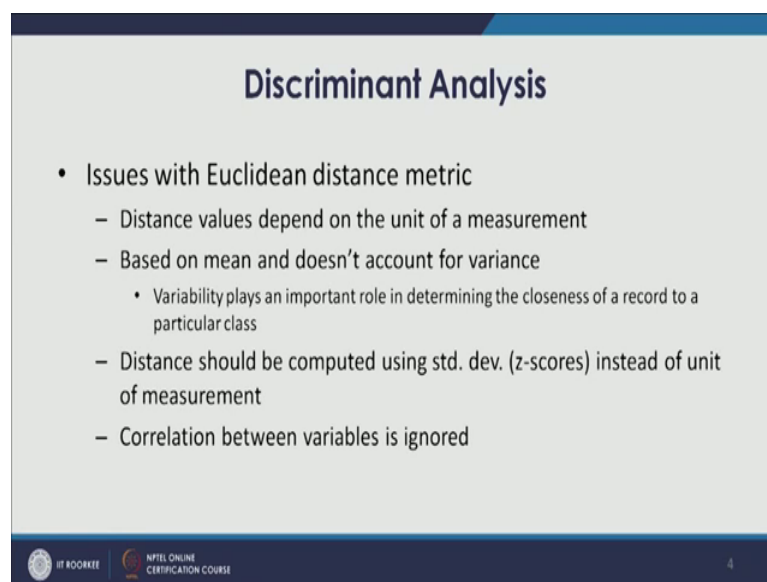
For example, best separation between items is found by measuring their distance from each class so for different items that are there. So, how we can separate them into their respective groups? So, that is found as we talked about measuring their distance from each class and particular item is classified to the closest class.

So, we measure the distance of a particular item from each class and then depending on its closeness to a particular class, so the particular item is accordingly classified. So, next important point is what are these distance metric that could be used; So, one alternative one option to use a Euclidean distance metric for discriminant analysis.

So, we are already familiar with this. This is the formula for a Euclidian distance metric. Distance of a record, let us say x_1 , to x_p . So, we have p predictors. So, we will have these p values and there for this p values, in this the distance of a record x_1 to x_p from centroid $\bar{x}_1$ to $\bar{x}_p$ of a class is computed in this fashion; So, this distance e_u ; that is for Euclidean and distance between x ; that is for an item for a record and distance from the centroid; that is $\bar{x}$. So, that is nothing, but vector of means of the predictors as you can see where centroid $\bar{x}$, centroid $\bar{x}$ is a vector of means of p predictors.

So, formula is quite familiar square root of x_1 minus $\bar{x}_1$ is square then up to x_p minus $\bar{x}_p$ whole square, so this is the formula and this can be used as a distance metric for those calculation. For calculation for each item and it is distance from each of those classes and then finally, classification based on those computations.

(Refer Slide Time: 07:21)



Discriminant Analysis

- Issues with Euclidean distance metric
 - Distance values depend on the unit of a measurement
 - Based on mean and doesn't account for variance
 - Variability plays an important role in determining the closeness of a record to a particular class
 - Distance should be computed using std. dev. (z-scores) instead of unit of measurement
 - Correlation between variables is ignored

IT ROOKIE | NPTEL ONLINE CERTIFICATION COURSE | 4

So, Euclidean distance metric is one that could be used, but as we have discussed before as well in (Refer Time: 07:25) as well that, there are a few issues with the Euclidean distance metric. So, let us understand them, so first one is that distance values depend on the unit of a measurement. So, as it is very clear from the formula itself. So, if we have

two variable one is being measured on square feet and the another one is being measured on hundreds of a square feet. So, of course, the actual values are going to be different, but those values when they are used in Euclidean distance metric one particular variable would dominate the distance.

So, therefore, distance values that are going to come, they are going to be dependent on unit of measurement. So, that is quite problematic in sometimes in some data sets. So, that is one problem, second one is based on mean and does not account for variance. So, we look at the Euclidean distance metric that we just saw. We are trying to compute the distance of a particular item, particular record from centroid of each of the classes. So, therefore, those now centroid is vector of means for all predictors then therefore, the distance computation is actually just accounting for mean values and we are not accounting for variance.

Now, a variability might play sometimes an important role in determining the closeness of a record to a particular class right because if the computation is based on just this, but the variance on the variability is not accounted one class would be one class might have a larger split and higher variability. So, because of that larger split also the new observation is likely to be closer to this class despite the distance between mean and that new observation might be on the higher side.

But because of the variability this split more likely that a new observation could belong to this class; however, even though the distance would be smaller from this lower variability group as per the Euclidean distance metric, that new observation might be assigned to these groups; however, higher split. So, probably there is a good chance that, that observation your objection might go to this group.

So, variability plays an important role in determining the closeness of a record to a particular class. Let us say a this is the observation new observation and we can see this is the distance from this centroid and this is the distance from this particular centroid. So, if we look at it this distance is smaller so therefore, as per the Euclidean distance metric probably this record is going to be allocated to this group; however, be look at the split of this group this is split is much wider. So, this point is quite close to this sphere; however, it is slightly more distant to this, this is sphere.

So, because of that we can say that there is again, so there is another argument could be there that this observation, if we will just look at the variability or spread then probably this observation can also belong to this particular group. So, as per the variability observation, this new observation can go to this group, but if we look at just the distance from the centroid that is accounting for mean this observation will go to this group.

So, therefore, a mean and a variance both should be accounted, so which is not the case with the Euclidean distance matrix. So, apart from the a scale dominance and the variance is also an issue on Euclidean distance matrix. So, to eliminate these two problems we can think of, a distance can be computed using standard deviation z score, that is z score instead of unit of measurement.

So, instead of using those actual, even if they are in different unit of measurement we can actually go for measuring them in terms of standard deviation So, therefore, that means, z scores. So, that would eliminate the first problem the scale dependence and the second problem also would be covered because this standard deviation which is also and indicator for the split would also be part of this process.

So, that is one way we can overcome these two problems; however, there is one more issue with Euclidean distance metric is that correlation between variables is ignored. So, these variables that are going to be the part of the process of this distance calculation. So, the correlation is also important. So, that is also ignored in Euclidean distance matrix.

So, how do we solve this one? Because correlation relationship between different because this one is more like 2 D space when we talking about p predictors. We are into p dimensional space. So, therefore, the correlation between two variables might also play an important role in terms of determining the closeness of a record to a particular group. So, this particular issue should also be resolved.

(Refer Slide Time: 13:38)

Discriminant Analysis

- “Statistical distance” (or Mahalanobis distance) can be used to overcome issues with Euclidean distance metric

$$D_{ml}(x, \bar{x}) = [x - \bar{x}]' S^{-1} [x - \bar{x}]$$

Where $[x - \bar{x}]'$ is transpose matrix of $[x - \bar{x}]$

- Column vectors are turned into row vectors
- and S^{-1} is inverse matrix of S (covariance matrix between p predictors)
- Can be considered as p -dimensional extension of division operation



So, the next distance metric that is called statistical distance or Mahalanobis distance, that can be used to overcome the issues that we have discussed that issues that with Euclidean distance metric. So, the statistical distance or Mahalanobis distance is typically defined as below as you can see in the slide; D_{ml} that is ml form Mahalanobis distance, then the distance of record x from centroid of a class $\bar{x}$ can be computed in this fashion, x minus $\bar{x}$ transpose and then we have S inverse then x minus $\bar{x}$.

So, you can see that x minus $\bar{x}$ transpose, this is transpose matrix of x minus $\bar{x}$. So, essentially the column vectors are turned into row vectors and then we have S inverse. So, this is inverse matrix of S .

So, this can be a thought of, this is where S is a covariance matrix between p predictors. So, see in the definition itself covariance metric, matrix is part of this calculation, the statistical distance calculation. So, therefore, correlation between predictors that is accounted and you can see that the S inverse, this can also be considered as p dimensional extension of division operation. So, in this fashion scaling is also part of the process. So, if we take this particular statistical distance formula to one dimensional space you would also realize that this formula would convert into z score for one dimensional space.

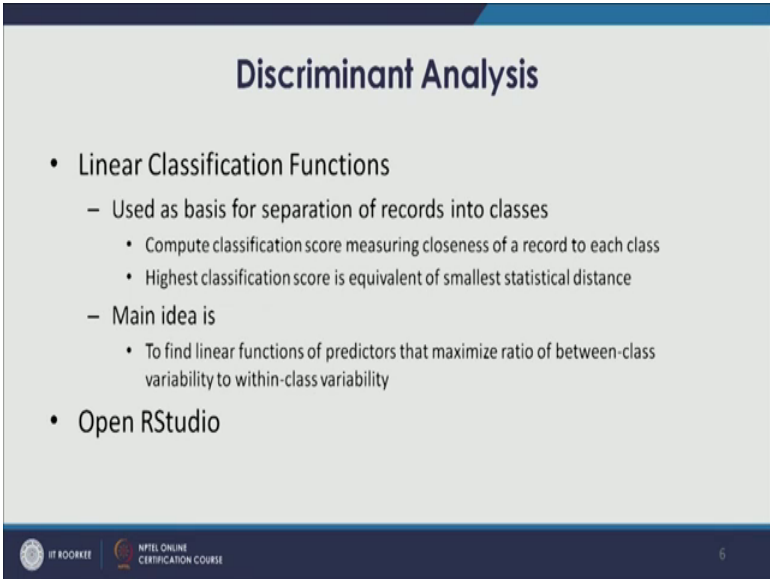
So, we would see that some of the two important issues that scale dependence and the variance, there that we could overcome using z scores are part of this formula because this is for n dimensional or specifically p dimensional space. So, this is also is

overcoming the taking the same advantage. It can say extension into a p dimensional space and because s is there, so which is covariance matrix.

So, correlation, another issue that was there with the Euclidean distance matrix that is also overcome using this statistical distance. So, correlation is accounted using the covariance matrix and since we are taking inverse of it. So, scaling is also accounted and since we are scaling also accounted, since we are taking the subtraction from mean value that is centroid so and then scaling is also accounted. So, scaled dependence is also considered variability and other things are also accounted for.

So, this statistical distance metric Mahalanobis distance metric that seems to be a much better metric for distance calculation in discriminant analysis.

(Refer Slide Time: 16:55)



Discriminant Analysis

- Linear Classification Functions
 - Used as basis for separation of records into classes
 - Compute classification score measuring closeness of a record to each class
 - Highest classification score is equivalent of smallest statistical distance
 - Main idea is
 - To find linear functions of predictors that maximize ratio of between-class variability to within-class variability
- Open RStudio

IT ROOKIE | NPTEL ONLINE CERTIFICATION COURSE | 6

So, next important point is about, how these some of these things are implemented? However, we have talked about the discussion that we had about, the discriminant analysis. It was mainly based on distance computing distance of new observation from each of the class and then assigning it to the closest class.

However the implementation is done using linear classification functions and that also brings about the similarity of discriminant analysis with the multiple linear regression. So, let us discuss this aspect. So, linear classification functions are used to implement some of these things that we talked about the distances and other things. So, used as

basis for separation of records into classes. So, these functions are used as basis for separations of records into classes and so this function compute classification score measuring closeness of a record to each class.

So, something that we discussed that distance metric are used and this is classification function, actually implement the same thing in a functional forms, so distance idea in a functional form is implemented. So, you can see first point here, compute classification score measuring closeness of a record to each class. The second point is highest classification score is equivalent of smallest statistical distance. So, you would see the main idea that was based on using the distance metric, it is actually captured here in linear classification functions.

So, but the implementation is now different. So, instead of measuring the distance, calculating the distance we actually use these classification functions to compute classification score; however, in terms of understanding and interpretation the main idea, the main underlying basis is same. So, instead of saying is smallest statistical distance, we would be saying is highest classification score.

So, again to understand what these functions do next point is so main idea behind these function is to find linear functions or predictors that maximize the ratio of between class variability to within class variability. So, you can see, when I said that linear classification functions being discriminant analysis closure to what we understood and multiple linear regression, you can see that here also we are looking for a linear functions of predictors.

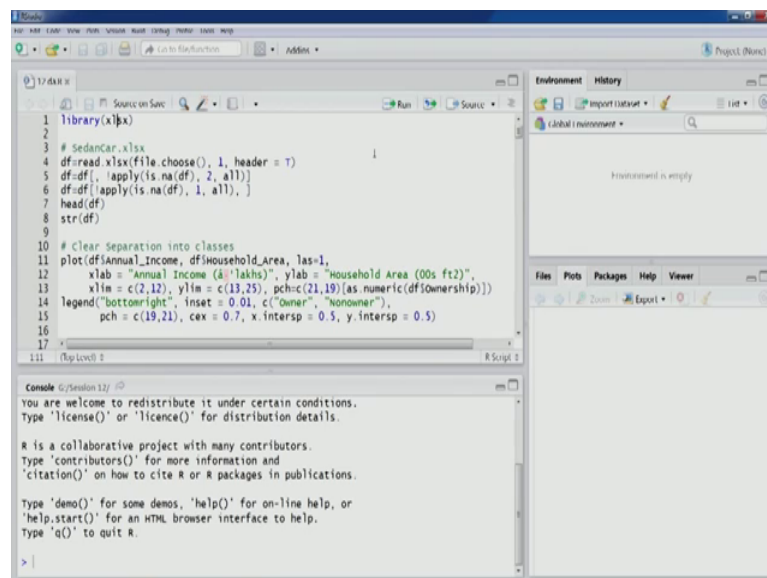
So, these classification functions, they are actually linear combinations of linear functions or predictors so in that sense, the similarity with the multiple linear regression. And later on will also discuss some of the application and performance of discriminant analysis are also quite similar to what we discussed in multiple linear regression. So, the idea to find this linear functions or predictors that maximize ratio of between class variability to within class variability. So, therefore, if the between class variability is between these two classes for example, in this two class case, so ratio of this. So, this variability and divided by within class variability.

So that means, we are trying to separate these groups. So, this ratio would indicate, so if this ratio is maximized, so therefore, we are trying to achieve maximum separation

between these groups. And we would be able to discriminate between observations and will be able to classify them to their respective groups.

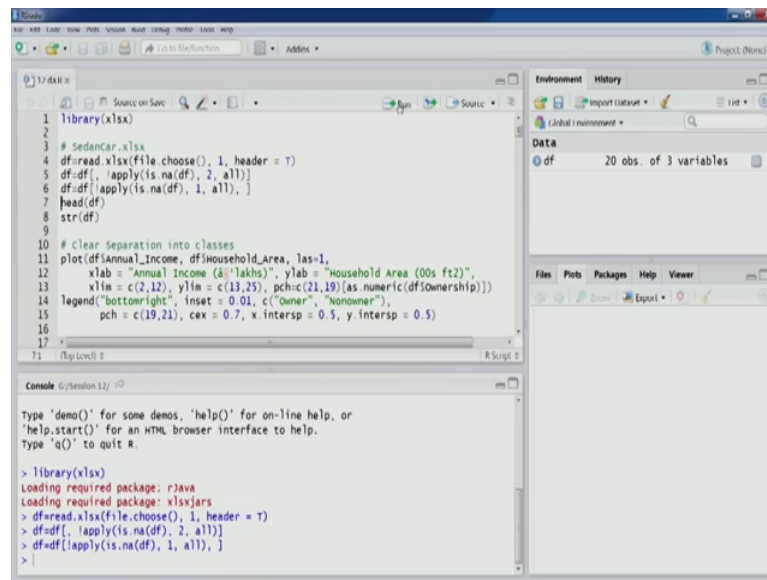
So, through maximization of this ratio these functions are determined. So, these linear functions or predictors they are determined using by maximizing this ratio and the underlying understanding is quite similar to what we discussed, the resistance based calculation. Instead of distance based calculation will have the classification scores, but the idea is coming from that. So, to understand few things what we have discussed till now, let us go back to R studio and through an exercise will try understand few of the points that we have discussed.

(Refer Slide Time: 21:37)



So, the data set that we are going to use right now is sedan car data set that we are already familiar with. So, let us load this package s plus x.

(Refer Slide Time: 21:46)



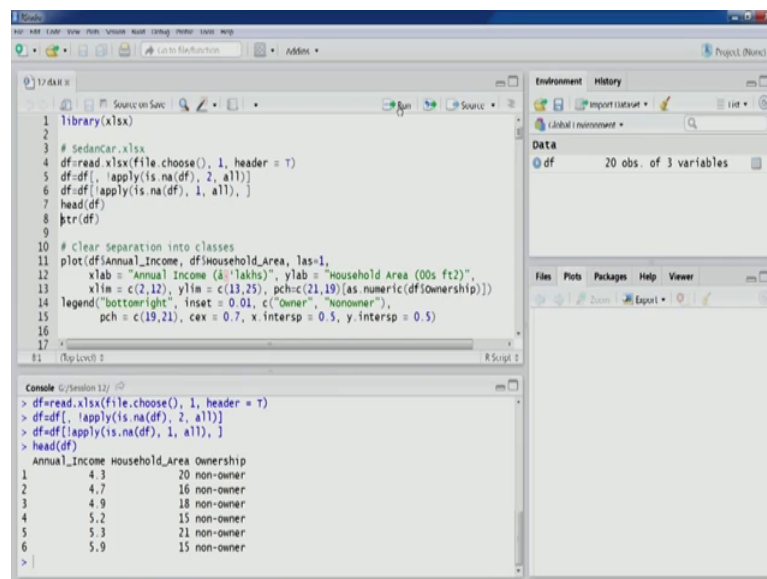
The screenshot shows the RStudio environment. The source editor contains the following R code:

```
1 library(xlsx)
2
3 # SedanCar.xlsx
4 df=read.xlsx(file.choose(), 1, header = T)
5 df=df[, !apply(is.na(df), 2, all)]
6 df=df[!apply(is.na(df), 1, all), ]
7 head(df)
8 str(df)
9
10 # Clear separation into classes
11 plot(df$Annual_Income, df$Household_Area, las=1,
12      xlab = "Annual Income (in Lakhs)", ylab = "Household Area (00s ft2)",
13      xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
14 legend("bottomright", inset = 0.01, c("Owner", "Nonowner"),
15      pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)
16
17
```

The console shows the execution of the code, including the loading of the 'xlsx' package and the execution of the read and plot functions.

So let us import the data set, let us to move array columns array rows.

(Refer Slide Time: 22:04)



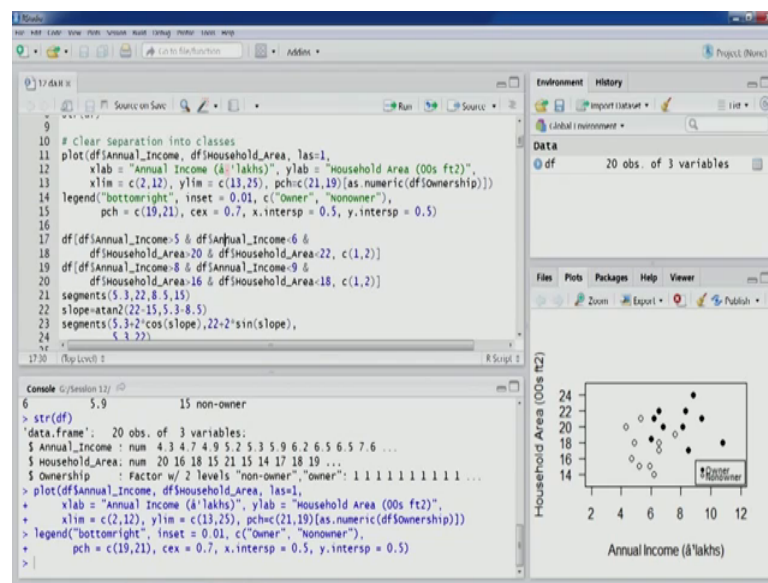
The screenshot shows the RStudio environment. The source editor contains the same R code as the previous screenshot. The console shows the execution of the code, including the loading of the 'xlsx' package and the execution of the read and plot functions. The output of the `head(df)` command is displayed in the console:

```
> df=read.xlsx(file.choose(), 1, header = T)
> df=df[, !apply(is.na(df), 2, all)]
> df=df[!apply(is.na(df), 1, all), ]
> head(df)
  Annual_Income Household_Area Ownership
1          4.3             20 non-owner
2          4.7             16 non-owner
3          4.9             18 non-owner
4          5.2             15 non-owner
5          5.3             21 non-owner
6          5.9             15 non-owner
```

So, these are the three variables that we already are familiar with annual income household area ownership and this is the structure of the data frame; annual income and household area are numeric and ownership is the factor variable, categorical variable.

So, that is also our outcome variable in this case. So, we have two groups as you can see non owner and owner.

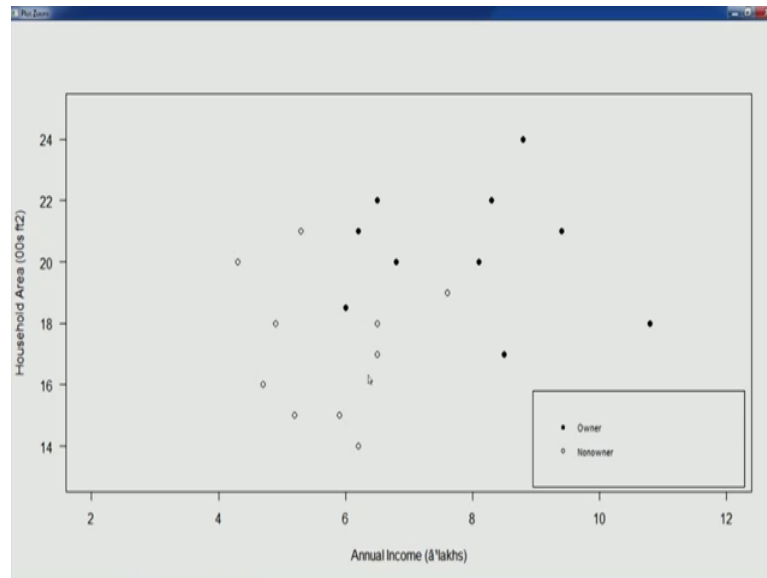
(Refer Slide Time: 22:24)



Now, first important aspect that we need to discuss is the class separation into classes. So, it is important for a particular classification model to do well in classifying observation that class separation is should be bit more clear. If class is separation is clear then a particular classification model would probably do a good job of classifying the observations.

However, if the class separation is not clear then in that case the model and modelling would be much more complicated and the performance would not be as expected. So, let us understand these two things using some plots here So, this plot which is quite similar to the example that we have shown here in the board, so let us plot this.

(Refer Slide Time: 23:23)



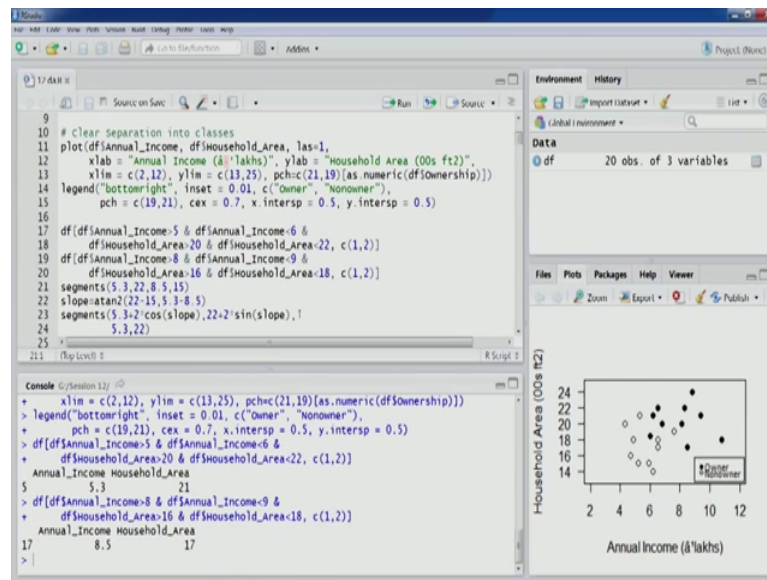
So, these are the observations this is the scatter plot. So, as you can see here this particular group is top right group, this is owners group and then bottom left group this is the non owners groups.

So, there is clear separation between these two groups and therefore, it is easier for any classification model to separate observation into their respective groups So, if the class separation is clear then, probably the classification would be much easier. So, we can find you know so discriminant analysis that main idea that we talked about that finding a line or hyperplane that is either equidistant. So, that was one approach or finding a best line or hyperplane that does the good job of discriminating the observations.

So, let us understand that, so try and find out a line. So, if we look at this scatter plot, if we draw a line somewhere from here to here you would see we will get a good enough separation where only one observation would be misclassified as we can see here. So, this is the point that we want to locate you can see, this is between 20 and 22 on y axis and between I guess 5 and 6 along x axis. So, this point we want to locate. So, our line would come above this particular point.

So, let us find out. So, this is one way to find out because we are doing some manual process here following manual process to find that line So, we can see 20 to 22 and 5 to 6. So, you can see annual income 5 to 6 and household area 20 to 22. So, we would be able to find probably this point.

(Refer Slide Time: 25:35)



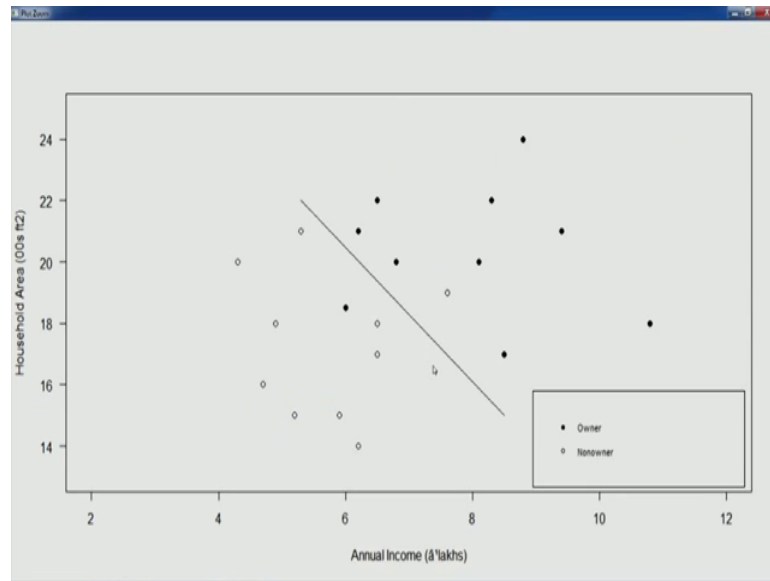
So, you can see this is the point 5.3 and 21. So, this seems to be the point 5.3 and 21. So, this is the point. So, let us find out another point in this particular zone.

So, this seems to the point. So, our line should be our discriminant lines should be below this particular line. So, as I suggested the line could go like this and therefore, it should be below this point. So, let us find out this point, this point seems to be between 8 and 9 along x axis and between 16 and 18 along y axis. So, you can see here 8 and 9 here annual income along x axis and 16 to 18 along y axis. So, let us find out this particular point. So, you can see 8.5 and 17. So, this seems to the point 8.5 about 8.5 and 17.

Now, we can now manually assign some coordinates to draw our line. You can see here we are plotting a line from 5.3 to 22. So, the point was 5.3, 21. So, we are moving up in the y direction. So, 22 value therefore, and the second point is 8.5 and the 17. So, again we are moving down along the y direction.

So, the x value is same at 0.5 and the value is say from 17 here to 15 now. So, will get a line a separating a discriminant line. So, let us plot this.

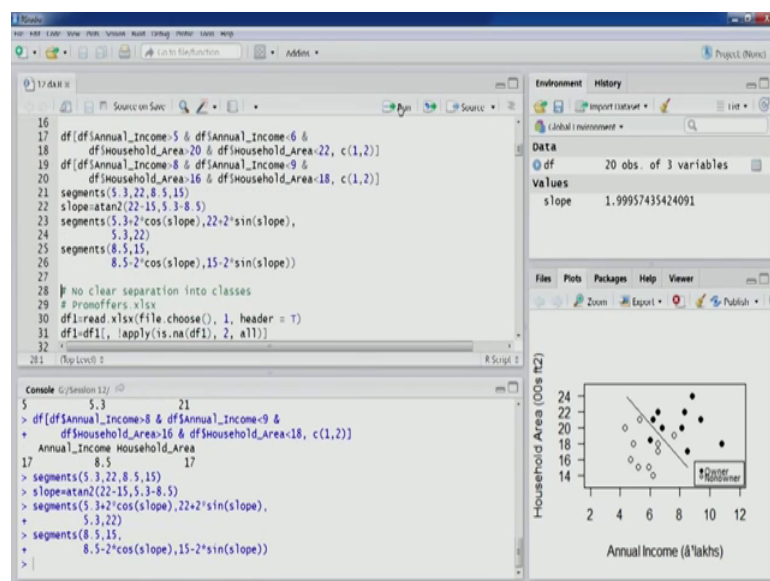
(Refer Slide Time: 27:09)



We can see, so probably this could be. So, through our manual process visual inspection we can see that this could be one line that would be able to discriminate the observations into their respective groups. So, this also seems to be equidistance line from the centroid of these two groups.

So, both these approaches typically the line that we get typically is going to be same. So, we can extend this line using few lines of codes here.

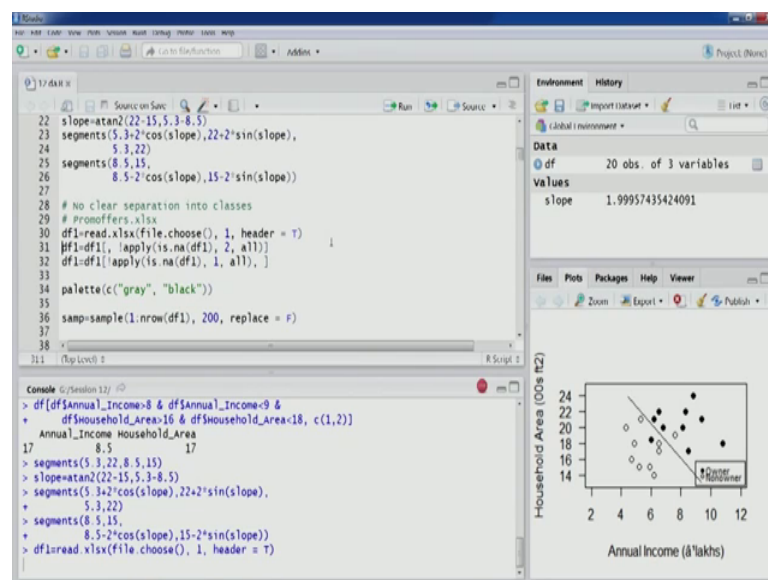
(Refer Slide Time: 27:45)



So, you can see using these two coordinates, these point coordinates that we have identified you can compute slope and we can use this slope to extend this line further as you would see the line is being extended in the plot. So, let us look at, you can see this line has been extended.

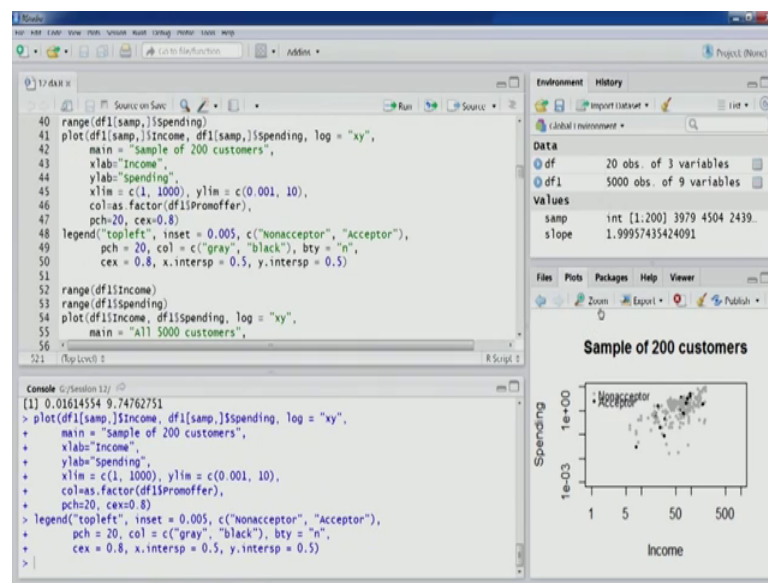
So, this is the line that we were talking about. So, the main idea about discriminant analysis is to find this line and that would be able to separate these observations. So, from this it is also clear that we would like to have the class separation. So, the class separation should be quite clear for discriminant analysis to probably work well. So, let us look at another example where this class separation is not quite clear.

(Refer Slide Time: 28:33)



So, this is promo offers data set. So, we are also familiar with this one as well. So, let us import this one and let us see, what is the scenario in this particular data set? So, in this particular data set as we would see that, the observations belonging to different groups there is no clear separation between those groups and that would actually complicate the performance of the model. So, let us you move array columns array rows. Let us change the palette and we are taking these 200 observations.

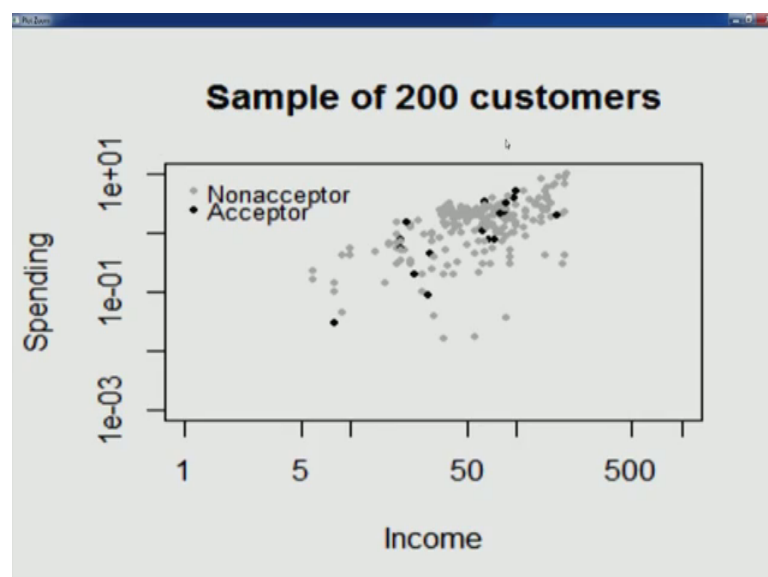
(Refer Slide Time: 29:14)



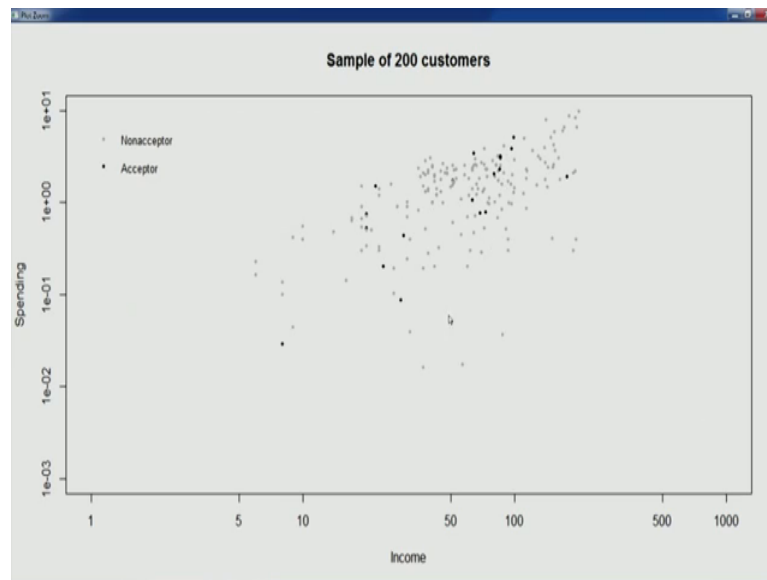
And using these two observation now we are again using log scale here, so because the there are too many observations and they are within one particular limited coordinate area. So, we would like to space them out.

So, that is why we are using log scale here. So these are the observations.

(Refer Slide Time: 29:46)

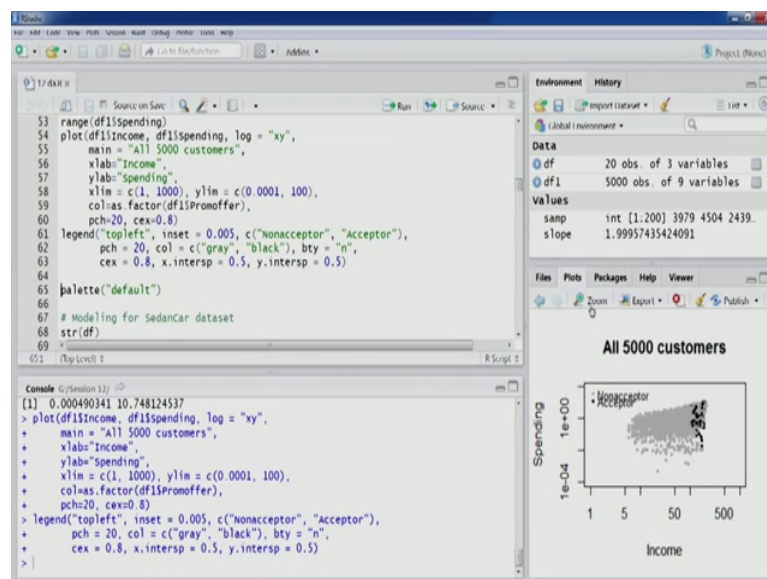


(Refer Slide Time: 29:47)



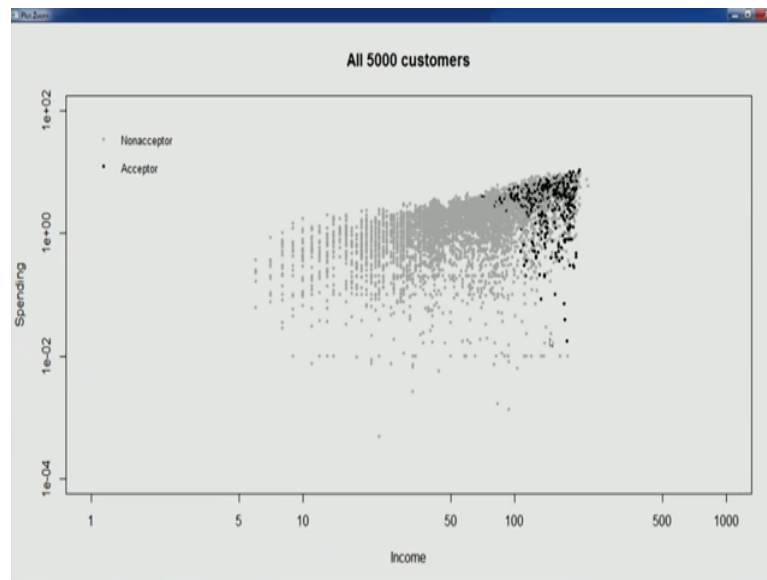
So, let us look at, so in this particular case we can see here that, even though we have used only 200 points we can see that a both the groups they are lying all around in the similar area and we can see that separation is quite difficult here.

(Refer Slide Time: 30:08)



Similarly, if we plot we use all the observations instead of this sample. So, let us look at all the observations. So, if we look at all the observations this situation is even more difficult, you can see here.

(Refer Slide Time: 30:20)



So, here you would see that this particular half top right half we see that observation belonging to both acceptor and non acceptor class are clubbed in there.

So, it is going to be quite those, the class separation is not quite clear in this particular region; however, this particular half the mid part and this left part, most of the observations belong to the non acceptor groups. So, these are low hanging fruits and model would be able to correctly classify these observations because typically belong to just one class one group; however, this top right group this is quite complicated the separation is not clear.

So, the separation is not clear it is going to be difficult for a model to give good performance. Now you would also see for any model that we apply on this particular data typically the performance of that what model would come out to be you know quite good that is because majority the observations are very easily going to be classified. You can see here so many observations are here and which are going to be easily classified as non acceptor. So, overall performance of the model would be quite good, but if we restrict our self to this part probably the performance of the model would be tested on this part because the class separation is not that clear.

So, with this we will stop here and in our next lecture on discriminant analysis, we will do modelling, we will build our discriminant analysis model using the sedan car data set.

Thank you.