Introduction to Proteogenomics

Dr. Sanjeeva Srivastava Dr. David Fenyo Department of Biosciences and Bioengineering New York University Indian Institute of Technology, Bombay

Supplementary - 13 Predictive Analysis

Welcome to MOOC course on Introduction to Proteogenomics. Today we have a hands on session by Doctor David Fenyo. In this session he will talk to you about some basic informations like how to open a file in R, how to transpose your data and how to run a command. A background knowledge in R will help you to understand these things must quickly next he will also discuss about how we can generate different kinds of plot and correlation study using R.

Finally, he will discuss about binary classification model and how to plot a binarized copy number. So, let us welcome Doctor David Fenyo for today's hands on session.

You should start our RStudio everyone should have that installed from previously.

(Refer Slide Time: 01:25)

Import	data				
	n in RStudio mport Datas			n Text (base)"	
	RStudio			- 0	×
	File Edit Code View Plots	Session Build Debu	Profile To		ionel •
	New Project		6		
	Open File	Ctrl+O		💣 🔄 🐨 import Dataset - 💉 👘 List -	
	Recent Files	•	ray" tistical		
	Open Project Open Project in New Sessic	20		Environment is empty	
	Recent Projects	•	Y NO WARR		
	Import Dataset	,			
	Save	Ctrl+S	ution det		==
	Save As Save All	Ctrl+Alt+S	ributors.	Trees Plaport . 0	
	Print		in publi		
	Close	Ctrl+W	on-line h		
	Close All	Ctrl+Shift+W	e to help		
	Close All Except Current	Ctrl+Alt+Shift+W			
	Close Project	Ctrl+Q			
	Quit Session				

And then you need to do two things one is to import the data and if you can look at the PowerPoint yourself allow I will show it quickly here but and but you should have the PowerPoint also. So, you can look at it how you import data.

(Refer Slide Time: 01:35)

 ← → ↑ ♠ > This PC > Desktop ✓ ♥ ♥ Search Desktop ✓ ♥ ✓ Organize • New folder ✓ Quick access ✓ Presentations # Fread # Grants # Grants # Grants # Grants # Soc Sync # 20181206-11 IIT 	Select File to Import				×
Organize New folder Image: Desktop Date modified Image: Desktop Date modified Image: Desktop Difference Image: Desktop <th></th> <th>> Desktop</th> <th>نة v</th> <th>Search Deskton</th> <th>Q</th>		> Desktop	نة v	Search Deskton	Q
Polick access Date modified Desktop prediction.tutorial.data.bt 2018-12-08 12:46 Presentations # prediction.tutorial.R 2018-12-08 18:08 Polick # Grants # Google Drive # Google Drive # Box Sync # 20181206-11 IIT					0
Desktop Presentations Presentations Publications Publications Crant Prepara Grants Google Drive # Desk Sync # 20181206-11 IIT	^ N	ame			Ту
	Presentations # Publications # Feaching # Peard # Grants # Grants # Grant Prepara # Google Drive #				R
data				_	

And it is the dot txt file that is the data.

(Refer Slide Time: 01:41)

RStudio							-	
File Edit Co	de View	Plots Session Build	Debug Profile T	ools Help				
		Go to file/fu					3	Project: (None)
Console Term	Impact Data	and a						
-1 -								E List - C
0	Name		Input File					
R verst Copyright	-	-	gene = A2-A0CM ER882.CNV = -0.2 ER882.CNV.binary	710.07	80.003	- 0.858 - 1	0.033 -	
Computing	Encoding	Automatic	ER8824.41911 ER8842.3362					
Platform:	-	• Yes O No	GR873.00480 TBCB - 1.241778	7174.0 49 - 0.796	5385342 058667 - 0	902759321 -	-3.165	
R is free			TRIM65 - 2.3511	0338 - 0.2	27241821	0.101326284	- 1.581	
ANTY. You are we								
ditions.	Separator	Tab						
Type 'lice ails.	Decimal	Period						-0
	Quote	Double quote (*)	· <				>	
R is a col Type 'cont	Comment	None					-	
'citation(cations.	na.strings	NA	Data Frame					
	Strings	as factors	gene ERB82.CNV ERB82.CNV.binary	A2. A0CM -0.271000 0.000000	A2. A0D2 -0. 0780000 0. 0000000	A2. A0EQ -0.0030000 0.0000000	A2.A01 0.858 0.000	
Type 'demo elp, or			ERBB2 ERBB4	-4.419112	-5.4210097	-2.7956007 4.9580924	-3.252	
'help.star			GR87 TBCB	-3.004808	-4.0538534 0.7960587	-2.9027593 0.1359536	-3.165	
Type 'q()'			TRIM65	2.351180	0.2272418	-0.1013263	1.581	
>								
			<				>	
					In	nport	Cancel	

And the only thing that is important to do when you import open the data file is that, you change the name and call it x just for so, that it is short and you should click yes here on the heading.

So, then it will take the sample names as the column headings and you should select use the first column as the row name. So, these are the gene names. So, so the data is a small data set. So, it is the 77 breast cancer samples from CPTAC that you have seen already before and I have all just done a small selection of genes.

So, these are from the bottom here there are 5 genes and these are the protein levels that we measured and it is the log 2 protein levels. So, it is ERBB 2, ERBB 4, GRB 7, TBCB and TRIM 65. So, there was are the 5 protein and then they have a value for each sample here and then the other thing we have is copy number. So, we have copy number only for ERBB 2 and we also I have binarized the copy number.

So, divided it into low and high. So, that is that is all the data and then you just click on import but if you do not for example, if you change do not change the name to x, it is going to be complicated than later for you and.

(Refer Slide Time: 03:35)

RStudio						- 🗆 X
Eile Edit Code View Plots Session Build Debug Profile Iool						ols <u>H</u> elp
• 🤏 💣	- 1 H M	📥 🏘 G			- Addin	s • 🚯 Project: (None) •
x×						Environment History Connections
01	🖓 Filter			2,		😅 🔚 🖙 Import Dataset • 🔏 👘 List • 🎯
-	A2.A0CM	A2.A0D2	A2.A0EQ	A2.A0EV	A2.A0E	🐴 Global Environment 👻 🔍
ERBB2.CNV	-0.271000	-0.0780000	-0.0030000	0.858000	0.033	Data
ERBB2.CNV.binary	0.000000	0.0000000	0.0000000	0.000000	0.000	🔿 x 7 obs. of 77 variables 📃
ERB82	-4.419112	-5.4210097	-2.7956007	-3.252209	-1.643	
ERB84	-2.336228	-5.6571851	4.9580924	-2.468041	-0.236	
GR87	-3.004808	-4.0538534	-2.9027593	-3.165390	-2.457	
TBCB	1.241778	0.7960587	0.1359536	-3.044965	1.087	
TRIM65	2.351180	0.2272418	-0.1013263	1.581628	0.014	Files Plots Packages Help Viewer
						🐖 🚳 🎤 Zoom 🛛 🚈 Export • 🔍 🧹
<					>	
Showing 1 to 7 of 7	antrian				,	
showing 1 to 7 of 7	comits					
Console Termin	al ×					
-/ 🐟						
lications.					^	
Type 'demo() help, or 'help.start lp. Type 'q()'	()' for an	HTML bro				
> x <- read on.tutorial > View(x)				top/predi	cti v	

So, that is what you should see if after you import it is you should see x here there should be 7; 77 columns and 7 rows and this is a showing a small part of it is.

Now, the next step is to open there the code which is the dot R file.

(Refer Slide Time: 04:03)

Image: Construction Source < P Image: Construction Image: Construction Image: Construction Image: Construction Image: Construction Image: Construction Image: Constretion	🕑 RStu	udio			-	>
<pre>time is a constant of the import Dataset is a constant of the</pre>	-					Project: (None)
<pre>> def i @mpoot Dasse + @ iut + @</pre>	-x×	prediction.tutorial.R × xt ×		Environment	History Connections	-1
<pre>2 low.xt[xt]_"EBB2_CNV"]>2,1 high.xt[_"EBB2_CNV"]>2,1 d genel="EBB2_CNV"]>2,1 d genel="EBB2_CNV"]>2,1 d genel="EBB2_CNV"]>2,1 d genel="EBB2_CNV"]>2,1 d genel="EBB2_CNV"]>2,1 d genel="EBB2_CNV"]>2,1 d threshold=-0.7 d high_below=high[high],"EEBB2_CNV_binary"]) fhreshold=-0.7 d high_below=high[high],genel]>=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]>=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high[high],genel]=threshold,1 d high_below=high_below</pre>		Source Q / -		-	import Dataset •	
<pre>Data high_titk1_tsttstBB2_cNV"]_2_1 high_titk1_tsttstBB2_cNV"]_2_1 points(high[genel],high["ERBB2_CNV"],col=" ippoints(high[genel],high["ERBB2_CNV",icol=" ippoints(high[genel],high["ERBB2_CNV.binary"]) ippoints(high[genel],high["ERBB2_CNV.binary"]) ippoints(high[genel],tt["EBB2_CNV.binary"]) ippoints(high[genel],tt["EBB2_CNV.binary"]) ippoints(high[genel],tt["EBB2_CNV.binary"]) ippoints(high[genel],tt["EBB2_CNV.binary"]) ippoints(high[genel],tt["EBB2_CNV.binary"]) ippoints(high[genel],tt["EBB2_CNV.binary"]) ispoints(high[genel],tt["EBB2_CNV.binary"]]) ispoints(high[genel],tt["EBB2_CNV.binary"]]) ispoints(high[genel],tt["EBB2_CNV.binary"]]) ispoints(high[genel],tt["EBB2_CNV.binary"]]) ispoints(high[genel],tt["EBB2_CNV.binary"]]) ispoints(high[genel],tt["EBB2_CNV.binary"]]]) ispoints(high[genel],tt["EBB2_CNV.binary"]]]) ispoints(high[genel</pre>		xt=t(x)	_			9
<pre>4 5 gene1="ERB84" 6 plot(xt(gene1],xt("ERB2.CNV")) 7 6 plot(xt(gene1],xt("ERB2.CNV"),col=" 1 10 0 plot(xt(gene1],xt("ERB2.CNV"),col=" 1 10 0 11 1 1 1 1 1 1 1 1 1 1 1 1 1 1</pre>				Data		
<pre>s genel="ERB84" plotxt;genel],rt;"ERB82.CNV"]) 7 points(high[genel],high["ERB82.CNV"]) 10 plot(xt[genel],xt["EBB2.CNV"]) 11 points(high[genel],tr["EBB2.CNV"]) 12 threshold=0.7 13 threshold=0.7 14 high_above_high[high[,genel],threshold] 14 or [bow herd][high[,genel],threshold] 15 cmster Termind * 15 Cmster Termind * 15 Cmster Termind * 16 cmster Termind * 17 help.start()' for an HTML browser interface to he for p. 7 (0) 'to quit R. 5 x < read.delim("c:/ksers/Fermyo/Desktop/predicti on.tutorial.data.txt", row.names_1) 5 View(Ch) 5 lowst(kt["EBB82.CNV"]-2,]</pre>	4	nign=xt[xt[, ERBB2.CNV"]>=4	2,1	high	num [1:16, 1:7]	3.66 3.66
<pre>7 8 points(high(,genel),high(,"EBBB2.CNV"),col=", 9 10 plot(xt[,genel],xt[,"EBBB2.CNV.binary"]) 11 points(high(,genel),high(,"EBBB2.CNV.binary"]) 12 brints(high(,genel),high(,genel)==threshold,] 15 high_blove=high(high(,genel)==threshold,] 15 high_blove=high(high(,genel)=threshold,] 15 high_blove=high(high(,genel)=threshold,] 16 consels Temmal -/- // Consels</pre>				low		
<pre>8 points(highl,genel],highl,TEBB2.cNv],col=""" 10 plot(t;(genel],highl,TEBB2.cNv.binary") 11 points(highl,genel],highl,TEBB2.cNv.binary") 12 threshold0.7 13 threshold0.7 14 high_betw=high[high],genel]=threshold,] 15 high_betw=high[high],genel]=threshold,] 16 (</pre>	6	<pre>plot(xt[,gene1],xt[,"ERBB2.</pre>	.CNV"])	🔘 x		
11 points(high[.genel],high[.fERB2.2(NV.binary" 13 threshold=0.7 14 high_betow-high[high].genel]=threshold,] 15 high_betow-high[high].genel]=threshold,] 16 c 18 c 10 c	9			xt	num [1:77, 1:7]	-0.271 -0
14 high_above-high(high[_gene1]==threshold,] 15 high_below-high(high[_gene1]==threshold,] 16 c 10 c	11			Files Plots	Packages Help Viewer	-
Torester Terminal a Conset Terminal a Con	13 14 15	high_above=high[high[,gene] high_below=high[high[,gene]	1] <threshold,]< td=""><td></td><td>Zoom 🛛 🐣 Export 👻 🔍 📄</td><td></td></threshold,]<>		Zoom 🛛 🐣 Export 👻 🔍 📄	
Conset Terminal Conset Termina	47					
<pre>//indip.start()' for an HTML browser interface to he / ip. Type 'q()' to quit R. > x <- read.delim("C:/Users/Ferryo/Desktop/predicti on.tutorial.data.txt", row.names=1) > View(x) > View(xt) > lowsxt[xt[_TERBB2.cxv"]<2,]</pre>	3:30	(Top Level) 0	R Script 0			
<pre>'help.start()' for an HTML browser interface to he Type 'q()' to quit R. > x < - read.delim("C:(JUSers/Ferryo/Desktop/predicti on.tutorial.dta.txt", row.names_l) > View(x) > view(x) > view(x) > lowsxt[xt], "ERB82.cxv"]<2,]</pre>	Console	Terminal ×	-0			
<pre>]p. 'Type 'q()' to quit R. > x <- read.delim("C://users/Fenyo/Desktop/predicti on.tutorial.data.txt", row.names=1) > xtatf(x) > view(x) > lowsxt[xt], "ERB2.cxv"]<2,]</pre>						
on.tutorial.data.txt", row.names=1) > View(X) > xtmt(X) > View(X) > lowaxt[xt]_["ERB82.CNV"]<2,]	lp.		er interface to he 🔨			
<pre>> View(xt) > low=xt[xt[,"ERBB2.CNV"]<2,]</pre>	on.tut > V	torial.data.txt", row.names iew(x)				
	> View > low:	w(xt) =xt[xt[,"ERBB2.CNV"]<2,]				

And then when you open the code you will see this and we are going to go through that and oh yes. So, there is one more thing that is important is, if you select a few lines and then click on here, I think I have seen some people have it is different ways that it says run may be, but that will then execute the line that you have marked and you can mark several lines.

So, now can anyone explain what this first line means?

And with a we are applying a function called t to x. So, x we can look at x it is where is it, here it looks like this. So, what is going to happen when we applied t the function t?

Student: we are extracting xt.

Yes yes.

Student: Input.

So, everyone knows what transpose is, you guys have transposed it already and I will do it now. So, like that now it appeared here. So, it we can look at it and it is now has the gene names as columns and the samples as rows, then what we want to do is what we are going to predict try to predict copy number that we will try to predict that is the samples that have a high copy number and so, we have this column now in the transposed one and we are going to say that any copy when the copy number is smaller than 2, then it is low and when it is larger than 2 it is high. And this is now the log version of it. So, we can run these 2 and see that we get in the high we have 16. So, if you remember my slides from the presentation there were a few now few dots and it turns out that they were 16 that had high ERBB 2 copy number and, but most of them 61 or low.

So, now let us try our first plot. So, you should see definitely look here that is there are 16 that have high copy number you should get that and 61 that had low. So, this taking a subset is very useful and we can that is something you will need very often and so, we just. So, what this notation means is that, we take xt and then buts we are only interested in the ones where the ERBB2 copy number is less than 2.

And those we assigned to a new variable that is called low. So, we can actually we can look at let us look at the high because it is a little bit smaller.

(Refer Slide Time: 07:31)

RSture	dio						>
-							
		Plots Session					-
		Gat Gat		11 65 -	Addin		Project: (Non
		orialR × 🗍 xt ×				Environment History Connections	-
	2 7 Filter		Q			😅 🔚 🖙 Import Dataset 🔹 🖉	List •
-	ERBB2.CNV	A CONTRACTOR AND AND	ERBB2	ERB84	GRB7	Global Environment -	9,
A2.A0CM	-0.271			-2.33622759	• •	Data O x 7 obs. of 77 v	ariables
A2.A0D2	-0.078			-5.65718510] -0.271 -0
A2.A0EQ	-0.003		-2.79560068				
A2.A0EV A2.A0EX	0.858		-3.25220905	-2.46804080 -0.23684699			
A2.A0EX	0.033		5.80097306			-	
A2.A0EY	-0.562			-2.20198022		Files Plots Packages Help Viewer	
A2.A0SX	-0.302			-2.20198022		De De Pacenger Help Helle	
AZ.A0T3	0.460			-3.93713370			
<					>		
Showing 1	to 10 of 77 entries						
Console	Terminal ×			-			
help, 'help. lp. Type '	or start()' fo q()' to qui read.delin orial.data. ew(x) (x)	some demos, " ir an HTML brow t R. (("C:/Users/Fen txt", row.name	ser inter	face to P	ie		

So, the; however, now we should only have 16 samples here, which probably we have switch this now is a subset is a view of the overall table that has 77 samples but that we have selected only the 16 that are high copy number.

(Refer Slide Time: 07:55)

B RStudio	- 🗆 X
ile Edit Code View Plots Session Build Debug Profile	Tools Help
- 😪 😅 - 🕞 🗐 👛 🔿 Go to file/function 📅 - Ad	
x × • prediction.tutorial.R ×	Environment History Connections
Source Q / · · · · · · · · · · · · · · · · · ·	
1 xt=t(x)	Global Environment - Q
<pre>2 low=xt[xt[,"ERBB2.CNV"]<2,] 3 high=xt[xt[,"ERBB2.CNV"]>=2,]</pre>	Data
4	🔿 x 7 obs. of 77 variables
<pre>5 gene1="ERBB4" 6 plot(xt[,gene1],xt[,"ERBB2.CNV"])</pre>	
7	
<pre>8 points(high[,gene1],high[,"ERBB2.CNV"],col=") 9</pre>	
<pre>9 10 plot(xt[,gene1],xt[,"ERBB2.CNV.binary"])</pre>	
11 points(high[,gene1],high[,"ERBB2.CNV.binary"	Files Plots Packages Help Viewer
12 13 threshold=-0.7	Zoom Zoport · O
14 high_above=high[high[,gene1]>=threshold,]	
<pre>15 high_below=high[high[,gene1]<threshold,] 16="" <="" pre=""></threshold,]></pre>	
1.0 < > 1.1 (Top Level) : R Script	
Console Terminal	
-/ ->	
, incacions.	
Type 'demo()' for some demos, 'help()' for on-line	
help, or 'help.start()' for an HTML browser interface to he	
lp.	
Type 'q()' to quit R.	
<pre>> x <- read.delim("C:/Users/Fenyo/Desktop/predicti</pre>	
on.tutorial.data.txt", row.names=1)	
View(x)	

So, now we just going to plot we defined gene one as ERBB 4 you just picked one out of the 5 proteomics tracks.

And we are going to plot on the x axis, we are going to plot ERBB 4 and on the y axis we are going to plot the copy number.

(Refer Slide Time: 08:23)

RStudio	X
File Edit Code View Plots Session Build Debug	Profile Tools Help
○ - 😪 💣 - 🔄 🗐 🌰 🖂 🚸 Go to file/function	💮 - Addins - 🚯 Project: (None) -
x × • prediction.tutorial.R × xt ×	Environment History Connections
Source on Save Q / · · · ·	🕂 Source • 🮯 📊 🖙 Import Dataset • 🥖 🛛 List • 🕥
1 xt=t(x)	A Global Environment - Q
<pre>2 low=xt[xt[,"ERBB2.CNV"]<2,] 3 high=xt[xt[,"ERBB2.CNV"]>=2,]</pre>	xt num [1:77, 1:7] -0.271 -0
4 angn=xt[xt[, "EKBB2.CNV"]>=2,]	Values
5 genel="ERBB4"	ene2 "TBCB"
<pre>6 plot(xt[,gene1],xt[,"ERBB2.CNV"]) 7</pre>	genel "ERBB4"
<pre>8 points(high[,gene1],high[,"ERBB2.CNV"]</pre>	,col=" gene1_cnv "TBCB.CNV.binary" gene2 "ERBB4"
9	genez EKBB4
<pre>10 plot(xt[,gene1],xt[,"ERBB2.CNV.binary" 11 points(high[.gene1],high[,"ERBB2.CNV.b</pre>])
12 points (high_, gener), high_, ERBB2.CNV.D	Files Plots Packages Help Viewer
13 threshold=-0.7	🦛 📫 🎤 Zoom 🦓 Export 🔹 🧕 💋 🔹 💮
14 high_above=high[high[,gene1]>=threshol 15 high_below=high[high[,gene1] <threshold< p=""></threshold<>	d,]
16 <	>
6:34 (Top Level) 0	R Script 0
Console Terminal x	- E 00 00 00 0 0
-1 00	EREB2.0W1
> threshold=-0.7	
<pre>> high_above=high[high[,gene1]>=threshold,]</pre>	
<pre>> high_below=high[high[,gene1]<threshold,]> low_above=low[low[,gene1]>=threshold,]</threshold,]></pre>	
<pre>> low_below=low[low[,gene1]<threshold,]< pre=""></threshold,]<></pre>	
<pre>> plot(xt[,gene1],xt[,"ERBB2.CNV.binary"]) > points(high[,gene1],high[,"ERBB2.CNV.binary"])</pre>	¥ - 80000000
<pre>> points(nign[,gene1],nigh[, "ERBB2.CNV.Dinar l="red")</pre>	
> lines(c(threshold,threshold),c(0,1))	-5 0 5 10
> genel="ERBB4"	
<pre>> plot(xt[,gene1],xt[,"ERBB2.CNV"]) ></pre>	xt[, gene1]

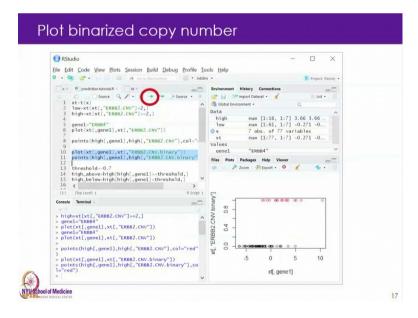
So, and that is plot is going to appear here. So, now, we see we have x axis is ERBB 4 and a copy number is for ERBB 2 on the y axis and we see that is there are these we know that there are 16 in our high group which are these here that is have a higher a copy number.

So, we have those 2 um. So, now, we can do one more thing here is to colour these in red. So, let me run it and then I will explain it. So, now, all the ones we are considering high or shown in red. So, what we did is we first plotted everything in the black here in the plot command, but then we overlaids in red but just for the ones that are in this the 16 in this high group.

So, we have now separated them. You can put search for maybe R points and then you will get some and then it will show you that that is what I did yesterday that to set define the colour, you write "col" short for colour equal to red and the red has to be in quotations and there are lots of other things you can do you can change the shape of these you can have them filled and all that you can look up on the internet how to how to do ok.

So, now we prop that copy number here, but we are really only interested if it is the copy number is low or high and so, there is one other column. So, here we plot the ERBB 2 dot CNV that is column but we have another one that is was ERBB 2 dot CNV dot binary. So, now, we are going to plot that. So, it is the same plot but we changing from the actual copy number to either 0 if the copy number is low or 1 if it is high.

(Refer Slide Time: 10:43)



So, then we run this and now the y axis will either be 0 or 1 and all the red ones will be up here and that is have high copy number they will have one and the other ones will be 0. So, you remember in my lecture I showed some slides where we had exactly these kinds of plots and then what is we want to do is to find a threshold here that where we have most of the red points above the thresholds and most of the black points below the threshold.

Now, as you see it is impossible to find a perfect thing for this, but let us try just one. So, now, there is a whole section here that is we are going to run, we are going to just define the threshold at 5 because all the black points are below 5. So, we are going to run this and get another plot where we put the line here at 5 and then we see we are going to get these define this low above.

(Refer Slide Time: 11:51)

B RStudio	0		- 🗆 🗙
le Edit	Code View Plots Session Build [Debug Profile To	pols Help
• O KI	💣 📲 🖶 💼 🖌 🔿 Go to file/functio	n 🗌 🔛 • Addin	ns • IB Project: (None) •
x× 0	prediction.tutorial.R* × 🗍 xt ×	-0	Environment History Connections
	- Source Q / - (-+)++	_→ Source • ≥	😅 🚽 🖙 Import Dataset + 🥑 📃 List • 🎯
	ot(xt[,gene1],xt[,"ERBB2V.b		Global Environment • Q
11 po 12	ints(high[,genel],high[,"ERBB2.	CNV.binary"	Data
13 th	reshold=5		high num [1:16, 1:7] 3.66 3.66 high_above num [1:3, 1:7] 3.66 3.66 3
	gh_above=high[high[,genel]>=th		high_below num [1:13, 1:7] 3.66 3.66
	gh_below=high[high[,gene1] <thre w_above=low[low[,gene1]>=thresh</thre 		low num [1:61, 1:7] -0.271 -0
17 low	w_below=low[low[,gene1] <thresho< td=""><td>old,]</td><td>low_above num[0 , 1:7]</td></thresho<>	old,]	low_above num[0 , 1:7]
	ot(xt[,gene1],xt[,"ERBB2.CNV.b ints(high[,gene1],high[,"ERBB2.		low_below num [1:61, 1:7] -0.271 -0 ~
20 lin	nes(c(threshold, threshold), c(0,		Files Plots Packages Help Viewer
21 22 ger	ne1="ERBB2"		🖕 🔰 🏓 Zoom 🎾 Export - 🧿 🕖 🤹 - 💿
	nel= ERBB2		
	nt(vtf nene11 vtf nene711	~	
25 <	oplevel :	R Script :	
	erminal ×	-0	8
-/ 10	(high[,gene1],high[,"ERBB2.CNV	hipary"] co	4 N 10
l="red")			8 1
> thresh			0.4
	bove=high[high[,gene1]>=thresh elow=high[high[,gene1] <thresho< td=""><td></td><td>8 -</td></thresho<>		8 -
> low_ab	ove=low[low[,gene1]>=threshold	,1	
	low=low[low[,gene1] <threshold, t[,gene1],xt[,"ERBB2.CNV.binar</threshold, 		¥ o tottalite to o o o
	(high[,gene1],high[,"ERBB2.CNV		-5 0 5 10
="red")			
lines(c(threshold, threshold), c(0,1))		xt[, gene1]

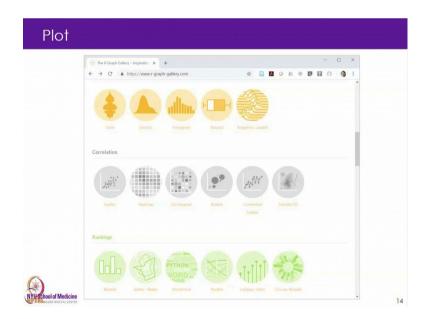
So, the num that is the number of black points above the line which is 0 and low below which is 61 which is all the black, then we have from the high group above we have only 3 which is a bit disappointing 3 out of 16 and that were classified right and then 13 is below. So, this is not a good choice of a thresholds but. So, what I want you to do is now find a good threshold oh yeah.

(Refer Slide Time: 12:43)

G r plot example	s - Google Search x +	×
$\leftrightarrow \rightarrow C$	a https://www.google.co.in/search?q=r+plot+examples&coq= 🖈 📄 🗾 🛛 💩 💿 🔯 🚱	1
Google	r plot examples	^
	All Images News Videos Maps More Settings Tools	
	About 44,90,000 results (0.39 seconds)	
	Producing Simple Graphs with R https://www.iharding.edu/infaccoventre C+ Jul 1, 2010 - First well produce a very simple graph using the values in the car vector In this next example, well save the file ta PP and choop of exits	
	5. Basic Plots — R Tutorial - cyclismo.org https://www.cyclismo.org/butorialR/plotting htm	
	People also search for X	
	how to plot imported data in r plot discrete function in r	
	strip chart example r par tutorial strip chart in excel plotting in r studio	
	R plot() Function (Add Titles, Labels, Change Colors and Overlaying https://www.atlamentor.ion-programming-plot-function/ ● ◆ In the article, you'l learn to use plot() function in R altich is used to make For example, the command plot(12, 12, 13), would refer points (13, 13).	
	The R Graph Gallery – Inspiration and Help with R Graphics	

So, this is another Google search I did is to how to plot with R and you get a lot of different answers.

(Refer Slide Time: 12:53)



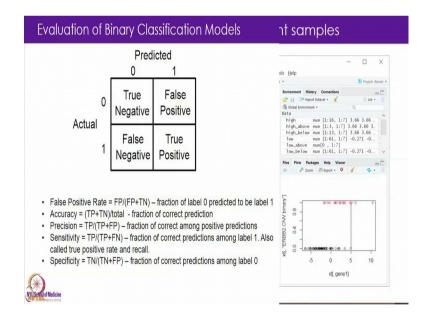
So, and one very nice thing is that they have a graph gallery that shows different types of graphs. So, you can look at the graphs and see I want to do that is kind of graph, then they have the code there you can just copy and paste and modify it a little bit.

(Refer Slide Time: 13:11)

RStudio File Edit Code View Plots Session Build Debug Profile Io	
<pre>>> @predecementAr = r</pre>	Environment Histry Connections Control Datase → Connections Control Datase → Connections Data Data Data Control Datase → Connections Data Data Control Datase → Connections Data Control Datase → Connections Control Datase → Connections Conne
Censole Terminal #	

And so, these will just now show the exactly what we have done.

(Refer Slide Time: 13:21)



And then if you remember one of the slides what we have done now is actually may a table like this. So, we had our actual groups which are the 0 would correspond to the lower group, the copy num ERBB2 copy numbers low and the one is when the copy number is high and then the predicted group is when we set or threshold now at 5. So, we get this table.

(Refer Slide Time: 13:57)

Test different thresholds	
What is the optimal threshold?	
 Try the different genes in the table: ERBB4, ERBB2, GRB7, TBCB, TRIM65 	
 Explore plot the data in different ways (see e.g. https://www.r-graph-gallery.com/) 	
NYUREsould Medicine	
NPuschool of Medicine	20

So, what I want you to do now is to change the threshold and just find what you think is a good threshold to choose for this example and this is for ERBB 4 then I want you to redo it for ERBB 2 protein values which of course, are going to be better that is what we expect, but also maybe try TRIM5 TRIM 65 and see find that for each of these the best threshold and also if you have more time you can explore the to plot the data in different ways by looking here at gallery yes.

So,. So, then every time you try and new thresholds you get a table like this. So, maybe fill out the table if it on paper and pen and forever and try a few and then move on to from ERBB 4 to ERBB 2 which is going to give a better result this because there is going to be more separation not surprisingly because the copy number effects the same gene more but as we saw it is before I show that ERBB 4 does not have a copy number change, but it is still affected by the copy number change in indirect way and so, on yeah.

So, that is what I want you to do and also TRIM 65 is negatively correlated, but not that strongly. So, that will be then the red and the black will switch it will be the other way around, but yeah. So, that is just explore and think about definitely think about what does it mean that the threshold is optimal. So, what we want to we want to get the values of these 4 values for every threshold.

And so and we have calculated these for a copy number high the ones that are above the threshold the for a copy number high advance that are below the threshold. So, can anyone

tell me where. So, there are 3 that are above the threshold these 3 here of the copy number high, 3 of the red points. So, where does that3 go in the in the in the confusion matrix which we have 4 different positions where would we put the number 3 here?

Student: may be in the true positives

Yes true positives. So, we put the 3 here and then we have the copy number high below which are the 13 and what are those.

Student: false negatives

Yes false negatives. So, we put 3 it in through positives and thirteen in false negatives and then we have the copy number low which then are the other two. So, we have the low above which are 0 the where does that end up.

Student: false positives.

True negative yes no I am sorry false false positive sorry yes. So, those are the false. So, we do not have any false positives and then the 61 the low below or the 61 true negatives and always with these we want as many as possible on the diagonal and as few as possible off the diagonal. Let us say that we have a very in additional test that we do for all the positive on that is rather quick and easy then we do not worry so, much about false positives then maybe we can allow more false positive, but we are worried about the false negatives.

So, it is not always, but the taking the sum is definitely one option. So, the other example is let us say the consequence of having a false positive is that the PhD student will spend 5 years investigating this protein and then we of course, want to have very few false positives. So, it is it always depends on the situation.

So, we had the high the copy number high that are above the threshold that is our true positives and then we have the copy number high that are below the threshold those was are the false negatives; then we have the copy number low above those are the false positives.

So, in this case with this initial threshold we had no false positives and then the copy number low that are below or the true negatives you should try to find is to minimize the sum of the false positives and the false negatives, that I think for this case we could call as the optimal case, but remember that is that is not the general statement that is it is not vary from case to case.

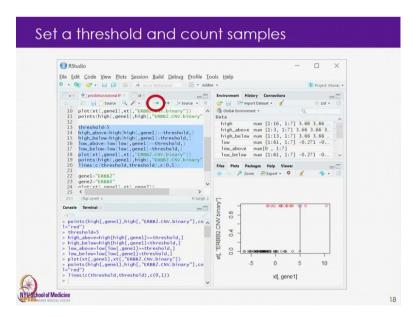
So, what you have done now is actually optimized a very simple machine learning algorithm. It is probably the simplest machine learning algorithm one can imagine, but it is still an algorithm that is where we separate try to predict what is positive and negative. And so, the what I want you to people talk a lot about machine learning and its, but and there is a lot of hype but there is no magic to it this is; it is I mean this is really when if you understand this the rest is just tricks to do it better.

Student: Sir. So, when you are calculating the threshold locking a threshold now based on that table if you take that precision value and I want to get the most precise values can that be used as the point for calculating on optimization?

Yeah, yeah that is a yeah if you want to do that definitely yeah.

So, I think this will kind of show this will kind of show the power of R. So, that is in. So, what David and I will show you all kind of showing of the power of R. So, David went through how you can manually.

(Refer Slide Time: 21:27)



Fill out that that confusion matrix or the table for all programming languages like R that comes with the fact that you can fill that the table with single line of code and so, now, if you have to optimize this we have to know the table let us say 10 lines if you do it by hands, some

may take you a month or maybe you are faster but still it will take you a lot time but you can very quickly do it with 10 lines of code, and if you learn this thing for a for loop you can do it with 2 lines of code and it all let you go through everything if you want to calculate precision or you want to calculate accuracy or sensitivity specificity there are formulas based on that table.

And R will automatically do those for you there are functions that calculate all those things. And we can do pretty much everything and you can even plot all the values as you are threshold changes and then see where the maximum or minimum occurs. So, that is the power of programming. So, I think David and I are basically introducing to you to what it can do but it provides power of those wave beyond brought excel or people can do and like David said it is all do by hand, but if you do it by code it is much faster one thing. The more important thing is the next and few each later you are trying to figure out what exactly you did it would be there in code if you did it in excel what ever you say is what we did before.

So, re reproducibility in terms of or remembering what you did in showing others what you did and in the publication making it available to others is made much easier if you can code I think that is where the power of programming comes. So, I will just show you how to create that table automatically and then I let you and Google deal with the best. To do this to do the table you need to have 2 vectors one that gives you the high and low based on copy number and another one gives you the high and low based on this threshold that we have chosen and so, when you have these 2 vectors you can call a function them create the table for you

(Refer Slide Time: 24:03)

Statement of the Area and the second s		- 11 ×
Q 13xala		
File Edit Code View Plots Session Build Debug Profile Tools Help		X Project (None) *
0 . 0		and the second second
T-	Environment History Connections	
i D prediction tutonal R high Rt	C H P import Dataset *	List *
Source on Save 4 /	Global Environment •	
11 threshold-S	Data	
14 high_above_high_highgene1 threshold.	high num [1:16, 1:7] 3.66 3.66 3.66	2.9 2.66.
15 high_below_high_high_genel_threshold, 16 low_above_low_low_genel_threshold,	high_above num [1:3, 1:7] 3.66 3.66 3.66 1	1
17 low below low low genel threshold.	high_below num [1:13, 1:7] 3.66 3.66 2.9 2	0.003.00
14 platist genel st ["FEREZ, ON, binary"]]	low num [1:61, 1:7] -0.271 -0.078 -	0.003 0.2
<pre>19 points high ,genel, high ,"rene2.cvv.binary"),col="red") 20 lines c threshold,threshold,c(0,1)</pre>	low_above num[0 , 1:7] low_below num [1:61, 1:7] -0.271 -0.078 -	0.003 0
20 Thes C threshold, threshold, the characteristic	Tow_Delow num (1.01. 1.7) - other of the	
22 gene1-"ERBS4"	Files Plots Packages Help Viewer	40
<pre>23 gene2 "TRIM65" 24 plot(xt(.gene1),xt(.gene2))</pre>	a P Zoom A Export • O	S Publish •
25 points high [.gene1].high [.gene2],col="red")		
26		
27 Algent t	F	
		0
Console Terminal		
n under wergeneeundent too dauf uitkeurent ereis ereise ereis	207	
<pre>> low_above=low[low[.gene1]>=threshold.]</pre>	6 1	
. low_below=low[low[.gene1]cthreshold.]	0.4	
<pre>plot(xt[.genel].xt[."ERBB2.CNV.binary"]) points(high[.genel].high[."ERBB2.CNV.binary"].col="red")</pre>		
lines(c(threshold, threshold), c(0, 1))		
and the second se	" o leader the t	-
and the second se	-5 0 5	10
6		
	xt[, gene1]	
NETTER		

So, the first vector threshold was 5 and then create another vector that says when it is when that gene is high and when that gene is low and once you have these 2 vectors that is the single line of code for the data plotting.

(Refer Slide Time: 24:31)

and the state of the		- 0
0 Okudo		
File Edit Code View Plots Session Build Debug Profile Tools Help		B bregart (Nara
0 · 0 · Addrs ·	and the second sec	-1
a = 0 prediction tutorial R · high = xt =	Environment History Connections	Unt +
Source on Save Q / - PRun ** Source *	🖹 💣 🛃 🧖 Import Dataset • 🦸	
1 st-t(x)	Global Environment *	
2 low-xt[xt[,"ERB82.CNV"]-2.]] high xt[xt[,"ERB82.CNV"]-2.]	Data high num [1:16, 1:7] 3.66 3.	66 3.66 2.9 2.66.
4	high_above num [1:3, 1:7] 3.66 3.6	6 3.66 1 1
<pre>5 genel="ERB84" 6 plot(xt[,genel],xt[,"ERB82.CNv"])</pre>	high below num [1:13, 1:7] 3.66 3.	66 2.9 2.66 3.66_
7	low num [1:61, 1:7] -0.271	-0.078 -0.003 0
<pre>8 points(high[,gene1 ,high(,"ER8B2.CNV"),col="red")</pre>	low_above num[0 , 1:7]	
9 10 plot(xt[.gene1].xt[."ERBB2.CNV.binary"])	low_below num [1:61, 1:7] -0.271	-0.078 -0.003 0
<pre>10 plot(xt .genel .xt ,"ERSB2.CNV.Dinary") 11 points/high(.genel].high[,"ERSB2.CNV.binary"].col="red")</pre>	Files Plots Packages Help Viewer	-
	a P Zoom - Export - O	Gr Publish •
13 threshold=5 14 high_above_high[high],genel]==threshold,]		
15 high below high high .genel threshold.		
16 low_above_low low [gene1] - threshold,] R Sene		
	>	00000
onsole Terminal		
er flor "en" rement i flor funder i florand vit e name rame rame i er fil	201	10 400 TO 10
low_above=low[low[.gene1]>=threshold,]		
low_below=low[low[,genel] <threshold,]< td=""><td>32 (</td><td>A CONTRACTOR</td></threshold,]<>	32 (A CONTRACTOR
<pre>plot(xt[,genel],xt[,"ERBB2.CNv.binary"]) points(high[,genel],high[,"ERBB2.CNv.binary"],col="red")</pre>	89 -	
lines(c(threshold, threshold), c(0, 1))		0
	8	
	5 0	5 10
(*)	xt[.gene1]	
NPTE,	¥.	

So, basically what you need to do is you have to say high is 1. So, you just stick that and just copy it over ok.

(Refer Slide Time: 24:43)

A REAL PROPERTY AND A REAL		- 0 ×
() Agudo		
File Edit Code View Plots Session Build Debug Profile Tools Help		E Propert (Name) *
0 . On . Addins *		
······································	Environment History Connections	
Source on Save Q / . PRun ** * Source ·	🔐 🔛 💷 Import Dataset *	
12	Global Environment *	
11 threshold S	Data	
14 high above high high genel threshold.	erbb2.cnv num [1:16, 1:7] 3.66 3.6	6 3.66 2.9 2.66
15 high_below high high .genel threshold. 16 low_above low low .genel threshold.	high num [1:16, 1:7] 3.66 3.6	6 3.66 2.9 2.66
16 low_above_low low of the threshold,	high_above num [1:3, 1:7] 3.66 3.66	3.66 1 1
18 olot(vt/ ownel xt, "FRREZ, CNV, binary"))	high_below num [1:13, 1:7] 3.66 3.6	6 2.9 2.66 3.66
19 points high , genel , high , "ERBB2 CNV, binary", col- red)	low num [1:61, 1:7] -0.271 -	0.078 -0.003 0
<pre>20 lines(c(threshold, threshold), c(0,1))</pre>	low_above num[0 , 1:7]	
21 22 gene1="(R884"	Files Plots Packages Help Viewer	
23 gene2 "TRIM65"	Zoom Tabport . O 2	S Publish ·
24 plot(xt[.gene1],xt[.gene2])	a the the	
<pre>25 points(high].gene1[.high[.gene2].col="red")</pre>		
26		
also (the Level) t R Screet t	2	
	CON binary	0 0
Console Terminal		
and a state of the	207	
1: In doTryCatch(return(expr), name, parentenv, handler) :	4 0	10000
display list redraw incomplete 11 In doTryCatch(return(expr), name, parentenv, handler) :	10 33	
<pre>invalid graphics state</pre>		
1: In doTryCatch(return(expr), name, parentenv, handler) :		
invalid graphics state		
	× 5 0 5	10
repart awapected input in "erbb2.cnv <- xt[xt[,"ERBB2.cNv"]>=2,]\"	xt[.gene1]	
1 1		
NPTEL -		

So, I now have a variable called erbb2 cnv

I just.

So, I want to mark things that are more than 2 as 1 and. So, so this statement is saying there is this matrix called xt and.

(Refer Slide Time: 25:35)

File Edd Code View Plots Session Build Debug Profile Tools Help	and the second		- 0 ×
• ABBS • • P publics house in the A + . • Source to leve A + . • P publics house it level and the high is and the source is and the so	() Ebudia		
P productional evolution with interval i			E Popet (Nene)
0 package magnetic magnetic <thmagnetic< th=""> magnetic magnetic</thmagnetic<>	0 . Of 12"		
Source on Ser Q. # - 1 **** **** **** ***** ************************************	9 emptotembatorial embblors high above high it -		
12 12 threshold.5 13 threshold.5 14 threshold.5 15 threshold.5 15 threshold.6 15 threshold.6 16 threshold.6 17 threshold.6 19 threshold.6 19 threshold.6 19 threshold.7 19 threshold.7 19 threshold.7 10 thresh		🐨 🗐 🤎 Import Dataset * 🦸	- CH -
<pre>13 Unterstold = // fight below high high (penel) = threshold.] 15 trigh_below high high (penel) = threshold.] 16 tog_box how low low (interstold, interstold, interstold,</pre>		Global Environment *	-U.UUS U
<pre>it imp</pre>	13 threshold-5	a to a standahlar	
16 Tog_blow_low low igenel_threshold, 17 Tog_blow_low low igenel_threshold, 18 Poly:X(:genel_it,":EBB2,CON:(insr/")) 19 points:(high).genel.h(h); 10 points:(high).genel.h(h); 10 points:(high).genel.h(h); 10 points:(high).genel.h(h); 10 points:(high).genel.h(h); 10 points:(high).genel.h(h); 10 points:(h); 11 # points:(h); 12 genel.***********************************	14 high_above_high_high_genel_threshold,	num [1:77, 1:7] -0.271 -0.078	-0.003 0
17 Totage bolow-low low (genel). threshold.] 18 plot:th:(penel).th:(pe	16 low above low low .gene1 = threshold.		
18 plot xt (_gene1, xt, _tema2, th, _tema2, th, _tema7, _teo1, "red") 19 plot xt (_gene1, _tema3, th, _tema2, th, _tema7, _teo1, "red") 20 threas (t threshold, threshold, xt(0,1)) 21 gene1, _tema3, " 23 gene1, _tema3, " 24 plot xt (_gene1, th, th, _tema2); 25 points (th, th, gene1), th, th, gene2); 26 plot xt (_gene1, th, th, th, gene2); 27 th, th, th, th, gene2); 26 th, th, th, th, gene2); 27 th, th, th, th, gene2); 28 th, th, th, th, gene2); 29 th, th, th, th, th, th, gene2); 20 th,	17 low below low low , genel =threshold,	erbb2.cny Named logi [1:77] FALSE FALSE	FALSE FALSE
20 Times (c threshold, threshold, e(0,1)) 21 genel. "EREA" 22 genel. "EREA" 31 genel. "TIMES" 32 genel.	18 plot(xt, genel, xt, "CR882, CNV. Dinary"), col="red")		FALSE FALSE
22 22 22 23 25 25 26 26 26 26 26 26 26 26 26 26	20 lines(c(threshold, threshold), c(0,1))	genel "ERB84"	
23 gene2-market. 24 plotx(1,gene1),*t(.gene2); 25 point(x1,gene1),*t(.gene2); 26 plotx(1,gene1),*t(.gene2); 27 to the field of the second			-5
<pre>34 plot xt (genel), +t(1, gene2), ext(1, gene2), col^{-red}) 35 points (high), gene2), ext(-red⁻) 36 37 38 (Buildent 1</pre>			A DALL .
<pre>55 points(high).genel).htgh.genel).col-"red") 52 53 (buind 1</pre>	24 plot(xt[.gene1],xt[.gene2])	D Zoom Ve Esport · · ·	D. Laborat
27 28 (bound lemma erbb).cmv ~ xt[xt[,"E882.CMV"]>=2,])." erbb.cm. c xt[xt[,"E882.CMV"]>=2,])." erbb.cm. c xt[xt[,"E882.CMV"]>=2,])." viewespected input timerbol.cmv ~ xt[xt[,"E8882.CMV"]==2,])." viewespected input timerbol.cmv ~ xt[xt[,"E8882.CMV"]==2,])."	<pre>25 points(high(,gene1),high(,gene2),col="red")</pre>		
10 Clauser 1 Causer 1 Clauser 1			
Canada Terminal ertbbl. chr.v = xt[xt[, "ERB2.CW"]>=2,])." ertbbl. chr.v = xt[xt[, "ERB2.CW"]>=2,])." visie(erbbl. chr.v = xt[xt[, "ERB2.CW"]>=2,])		2	
erbb2.cmv < xt[xt[.*EB82.CMV"]>=2.]\ erbb2.cmv < xt[xt[.*EB82.CMV"]>=2.]\ visia(erbb2.cmv) = xt[xt[.*EB82.CMV]>=2.]\ visia(erbb2.cmv) = xt[xt[.*EB82.CMV]>=2.]\\visia(erbb2.cmv) = xt[xt[.*EB82.CMV]>=2.]\\visia(erbb2.cmv) = xt[xt[.*EB82.CMV]>=2.]\\visia(erbb2.cmv) = xt[xt[.*EB82.CMV]>=2.]\\visia(erbb	and the second se	a oo ao a ao oo	0
erb02.cmv <- xt[xt[.*CR802.CNV"]>>2.]\ Erb02.cmv <- xt[xt[.*CR802.CNV"]>>2.]\ erb02.cmv <- xt[xt[.*CR802.CNV"]>>2.]\ Erb02.cmv <- xt[xt[.*CR802.CNV"]>>2.]\ view(erb02.cmv)	Console Terminal	La m	
erbb2.cmv <- xt[xt["CR882.CNV"]>=2.]\ rror: unexpected input in "erbb2.cmv <- xt[xt["R882.ON"]>=2.]\" view(erb2.cmv <- xt[xt["R882.CNV"]>=2.] view(erb2.cmv) <- xt["xt["R882.CNV"]>=2.]\" view(erb2.cmv) <- xt["xt["R882.CNV"]>=2.]\" view(erb2.cmv <- xt["xt["xt["R882.CNV"]>=2.]\" view(erb2.cmv <- xt["xt["xt["xt["xt["xt["xt["xt["xt["xt["	CONTRACTOR DESIGNATION ADDRESS .	30	
error: unsequent continue to the contin the continue to the continue to the cont		ō "]	
view(erb2,cm) view(erb2,cm) view(high.above) view(high.above) -5 0 5 10	arthD cov or vt[vt]. "FRB2. CNV"]>=2.]	6 01	
view(erb2,cm) view(erb2,cm) view(high.above) view(high.above) -5 0 5 10	<pre>irror: unexpected input in "erbb2.cnv <- xt[xt[, "ERBB2.CNV"]>=2,]\"</pre>	82 -	1000
vine(etta).ctv/[/38882.ctv/]>=2 Vine(big(1,40ve) Vine(big(1,40ve))			
View(high, above)		4	
		5 0 5	10
A former it (. reset 1) >> 0	(A)	at any th	
MPTTL I	teres pene t [, 'ERBEA'] >= th	xt[.gene1]	
	NPTEL 1		

I am picking the erbb 2 cnv column and if that column is greater than or equal to 2, then the Boolean value is true in other words whenever that is true for any sample, then that will have

a 1 otherwise it will have a 0. So, it is just a different way of representing what David had before. So, if you look here. So, now, you can see erbb 2 cnv is logical and it has false false false true and so, far.

So, for every sample it will say whether it was greater than 2 or less than 2. So, it now you have taken values and discretized it. So, it has become true or false I will do the same for the other I am going to set my threshold to 5 I will call it th. So, one of the tricks that R programmers use is to how names that are really small. So, you do not have to type too much when you are writing code.

But that is not a good idea when you are writing code to keep this is like just to try things out when you are writing code to publish if you have the then you do not know whether it was a what it meant. So, then you need a longer name, but when you are just trying things out it helps to how smaller names and so, now, I am going to say erbb 4.

Student: What does that arrow means?

No, the error happened because I had an extra xt here.

Student: no the arrow arrow.

Or the arrow means get the result and put it into some name. So, what I am saying here is. So, this line is basically saying I have the number 5, I want to call it th. So, that way when I write code using th I can go on change it to 3 and then next time I can run the same code and it will do it for 3 instead of 5.

Student: So, I will use the equal to sign.

Equal to sign is also the same.

Actually equal to is a more newer thing they would not allow equal to in R before.

I think, but the later versions allow equal to, but I am used to doing the less than minus is the same as equal to.

Student: Yeah.

So, I will used the equal to because I do not really know R.

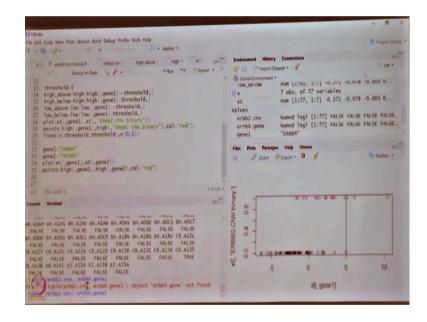
In many languages it is equal to.

Yeah.

And all I think realize that people are getting confused. So, only recently they have allowed the equal to sign till than it was the less than minus. So, that shows that I am I am a little old.

So, for the erbb 4 gene I am going to say I am going to take the array xt and I want the gene ERBB 4 and I want this to be greater than or equal to threshold in order for it to be called high. So, now, for ERBB 4 gene again you see here there are 77 things and they are all either true false or some true or false. So, you can even look at it by just typing the name here. So, now, we save the result in ERBB 4 dot gene when.

(Refer Slide Time: 28:41)



So, you can see for which sample it is true for which sample it is false it has the whole thing encoded in that one name.

So, now all we need to do to get your table that you did manually before is to say I want a table that compares erbb 2 copy number and erbb 4 the gene what did I called.

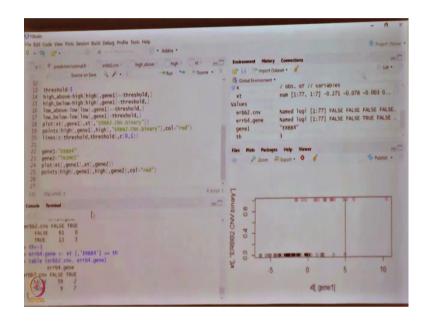
Student: So, actually I think that erbb.

Er yeah I spelt it wrong the first time. So, these are things that you have to keep in mind. So, many of your had issues when you load at the table for example, if you did not remove the

column that said genes and make it the row name then when you do the transpose of the matrix, the transpose loads only when you have a real numbered matrix.

So, there are all these perks and debugging that you need to learn along the way, it is not like as straightforward as one would think because the error that comes out does not really tell you what is happening or what is wrong. So, you need to kind of logically think through and do it, but you do it a couple of times you will you will get it. So, there is your table.

(Refer Slide Time: 29:49)



So, you can see the true false here is the predicted and the true false on the on the rows or the actual. So, in our case the this is for the copy number and this is for the gene and so, you can this is the table that you manually created by counting dots before. So, now, let us say I want to change the threshold to 3 I said threshold to 3 and then I just repeat the same command I had before.

And then I just plot the table again. So, now, if you set it to 3 this is the result you will get and so, on and there is a construct called a for loop where you can say start with my threshold of 1 and go up to 10 and it will calculate this table automatically and print it for values from 1 to 10.

But I thought one should never use for loops in R.

That is true. That is more complicatedSo, there is a way to do it without doing for loops in a single line of code and I think that is when you come for the advanced R course next time.

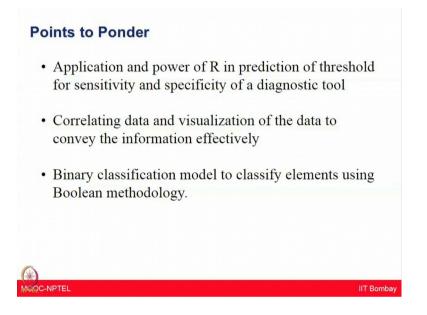
Yes yeah.

That is for yeah. So, you are allowed to use for loops until next year.

Yeah.

But after that is no more.

(Refer Slide Time: 30:57)



I hope this session was informative for you all, where you learned some basics of R followed by the prediction analysis. Doctor Fenyo showed you how to set an optimal threshold and count the sample we have also learned how to colour code the samples that is coming with high copy number. Further he showed you how binary classification model can be used to classify the elements of a given set into 2 groups.

There are many matrices that can be used to measure the performance of a classifier or predictor. Different fields have provided different preferences for a specific matrices due to different goals. For example, for the clinical applications the sensitivity and a specificity are often used. I will recommend you all to learn some basics of R that will help you a lot for doing the statistical analysis and prediction analysis very easily.

There are many publically available resources where various software various codes are already available, but you at least need some basic ability to run those codes in r. In the next

supplementary lecture, we will have TA for this course Deeptarup Biswas who will discuss about pathway enrichment and network analysis.

Thank you.