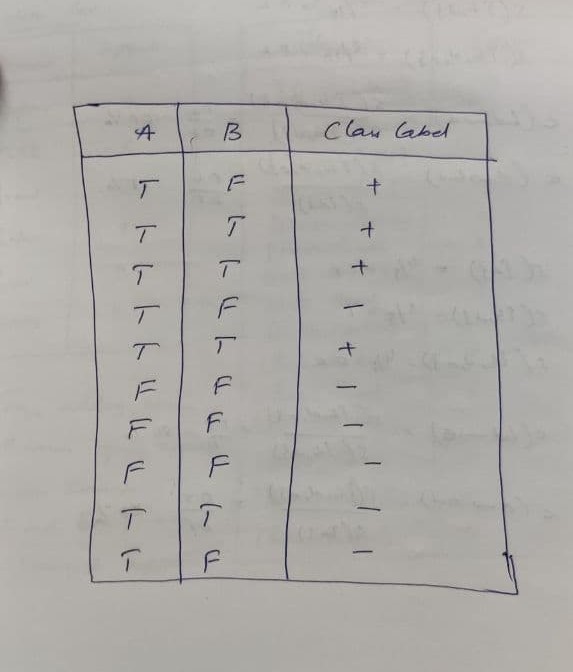
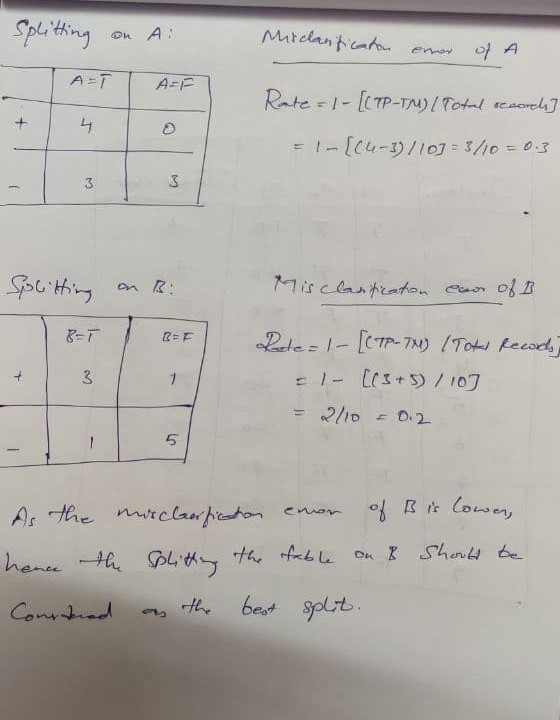
Data Mining Assignment 4

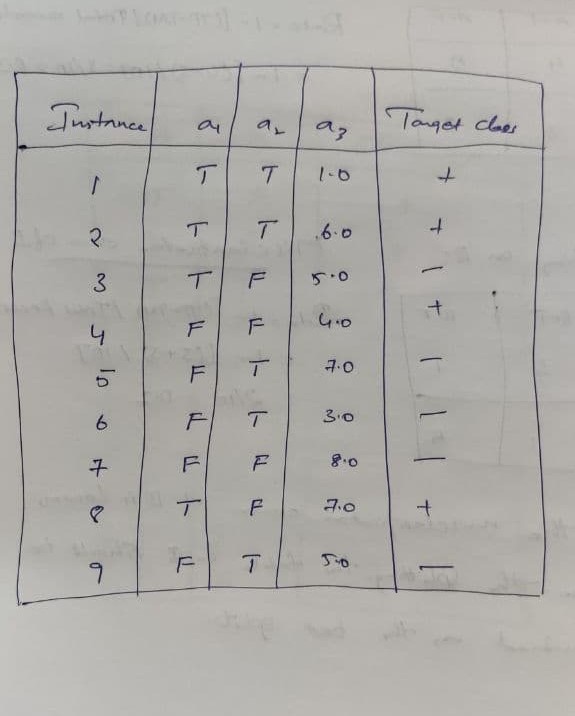
1. Read Chapter 4 (all sections) and Chapter 5 (Sections 5.2, 5.5, 5.6 and 5.7).
2. Consider the following data set for a binary class problem.



Calculate the misclassification error rate when splitting on A and B to determine the best split. Which of these splits considered is the best according to misclassification error rate?

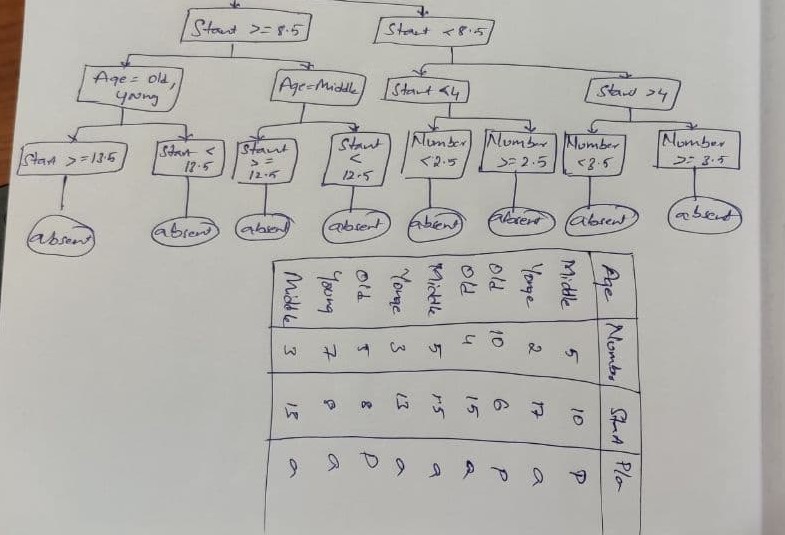


1. Consider the training examples shown below for a binary classification problem.



For a3, which is a continuous attribute compute misclassification error rate for every possible split to determine the best split. Which of these splits considered is the best according to misclassification error rate?

1. The file <http://www-stat.wharton.upenn.edu/~dmease/rpart_text_example.txt>gives an example of text output for a tree fit using the rpart() function in R from the library rpart. Use this tree to predict the class labels for the 10 observations in the test data [http://www- stat.wharton.upenn.edu/~dmease/test\_data.csv](http://www-stat.wharton.upenn.edu/~dmease/test_data.csv) linked here. Do this manually - do not use R or any software.



1. I split the popular sonar data set into a training set ([http://www- stat.wharton.upenn.edu/~dmease/sonar\_train.csv](http://www-stat.wharton.upenn.edu/~dmease/sonar_train.csv)) and a test set ([http://www- stat.wharton.upenn.edu/~dmease/sonar\_test.csv](http://www-stat.wharton.upenn.edu/~dmease/sonar_test.csv)). Use R to compute the misclassification error rate on the test set when training on the training set for a tree of depth 5 using all the default values except control=rpart.control(minsplit=0,minbucket=0,cp=-1, maxcompete=0, maxsurrogate=0, usesurrogate=0, xval=0,maxdepth=5). Remember that the 61st column is the response and the other 60 columns are the predictors.

* sonar\_f it <- rpart {y - . , x , contr of - rpart. control {ml ns pJ Its , md nbucket-0, cp--1, tea xcolnpeze=0, maxsurrogaze=0, usesur rogaze=0, xval =o, maxdepzh=S))
* pJ at {sonar\_f1t)
* text {sonar\_f1t)
* pr 1rrr {sonar\_fiz) n= 130

node) , split , n, 1ass , yval , {yprob)

^ denotes termi na1 node

1 rool 130 64 -1 {0. 50Z69231 0. 492 30769)

2) v11>=0.17095 79 21 -1 (0.F34177Z2 0.26582278)

4) vZ7>=0.8191 3Z 2 -1 (0.94594898 0.084O84O5)

8) v9>=0.0889 34 0 -1 (1.00000000 0.OOOOOO00) \*

9) v9< 0.0889 3 1 1 (0.3333333J 0.66666667)

I8) V2< 0.OL9S 1 0 -1 (1.00000000 D.00000000)

19) V2>=0.0195 2 0 1 (0.00000000 1.OOOOOO00) \*

5) vZ7< 0.8191 42 19 -1 (0.34761903 0.48238099)

10) V54>=0.020ZS 12 0 -1 (1.00000000 O.OOOO0000) \*

11) v54< 0.O2OZS 30 11 1 (0.36666667 D.63333333)

ZZ) V8< 0.17045 17 7 -l (O.588ZJ5Z9 0.41176471)

44) v36‹ o.4491 12 2 —1 (o.gJJJJJ33 o.iee6666F) °

43) V36>=0.4491 S 0 1 (0. OOOOOOOD l.OOO00000) \*

2 3) VB>=O. 17045 13 1 1 {0 . 07692 3D8 0. 92 307692)

46) V1>=0.0925 1 0 -1 (1.0000000D O.OOO00000) \*

47) V1< 0.0925 12 0 1 (0.0000000D l.OOO00000) \*

3) v11< 0.17095 51 8 1 (O.l5686Z75 0.84Jl3725)

6) v82>=0.0209 6 1 —1 (0.8333333J 0.16666667)

IZ) v1>=0.02225 S 0 -1 (1.00000000 0.00000000) \*

13) v1< 0.02225 1 0 1 (O.OOOO0000 1.DOOOOOOO)

7) v82< 0.0209 4S 3 1 (0.06666667 0.9J333333)

14) v19>=0.8351 S 2 -1 (0.60000000 0.4OOOOO00)

Z g) V26< 0. 61S3 3 0 —1 (1. 0 0000000 0. 00000000) •

Z9) V26>=0.6153 2 0 1 (0.00000000 1.00000000) \*

15) v19< 0.8351 40 0 1 (0.00000000 1.00000000) \*

* + pr eds etc one <- as . numer i c{pr edict {sonar\_f1t , sonar\_test , type = " c1ass ") )
  + pr eds etc ons <- repJ ace{pr edi ct:i ons , predlct1ons == 1,
  + pr eds etc one <- repJ ace{pr ed1cti ons , predl ct1ons == 2, 1)
  + pr eds etc one

[lj —1 -I —l 1 -1 -1 1 1 -1 —1 —1 I 1 1 -l -L —1 —1 1 1 -1 —I 1 -1 L 1 —1

28j -1 -1 -1 -1 1 -1 1 1 1 1 1 -1 -1 -1 -1 -l -1 1 -1 1 1 1 -1 -1 -l 1 1

SS] -1 1 1 -1 1 l 1 -l -l 1 -1 -1 1 -1 1 -l 1 l -1 -1 l 1 1 1

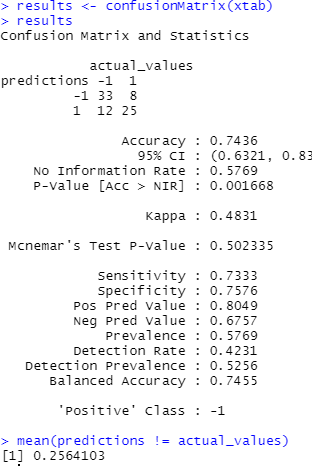
* acIual\_values sonar\_zesr[, 61a
* actual\_values

[1] -1 -1 -l 1 1 l 1 l -l -1 -1 1 -1 1 -l -l -1 l 1 1 -l -1 1 -1 -l 1 -1

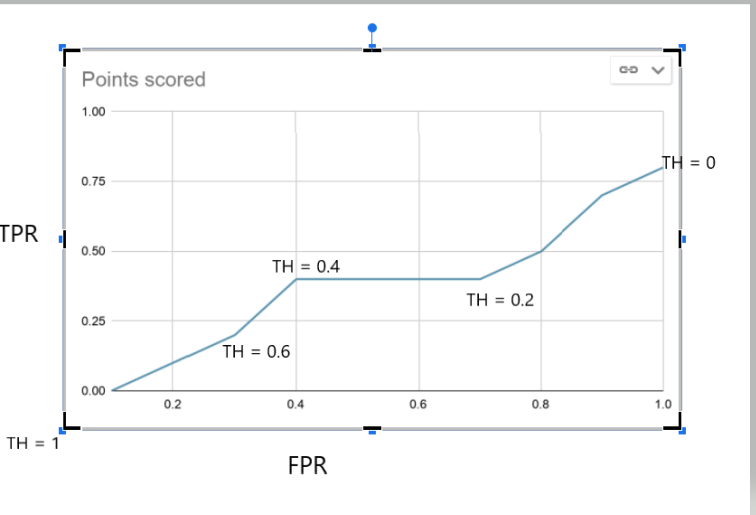
28a -1 I 1 1 1 l 1 l l -1 1 -I -1 -1 -l l -1 l -1 1 l I -1 -1 -l 1 -1

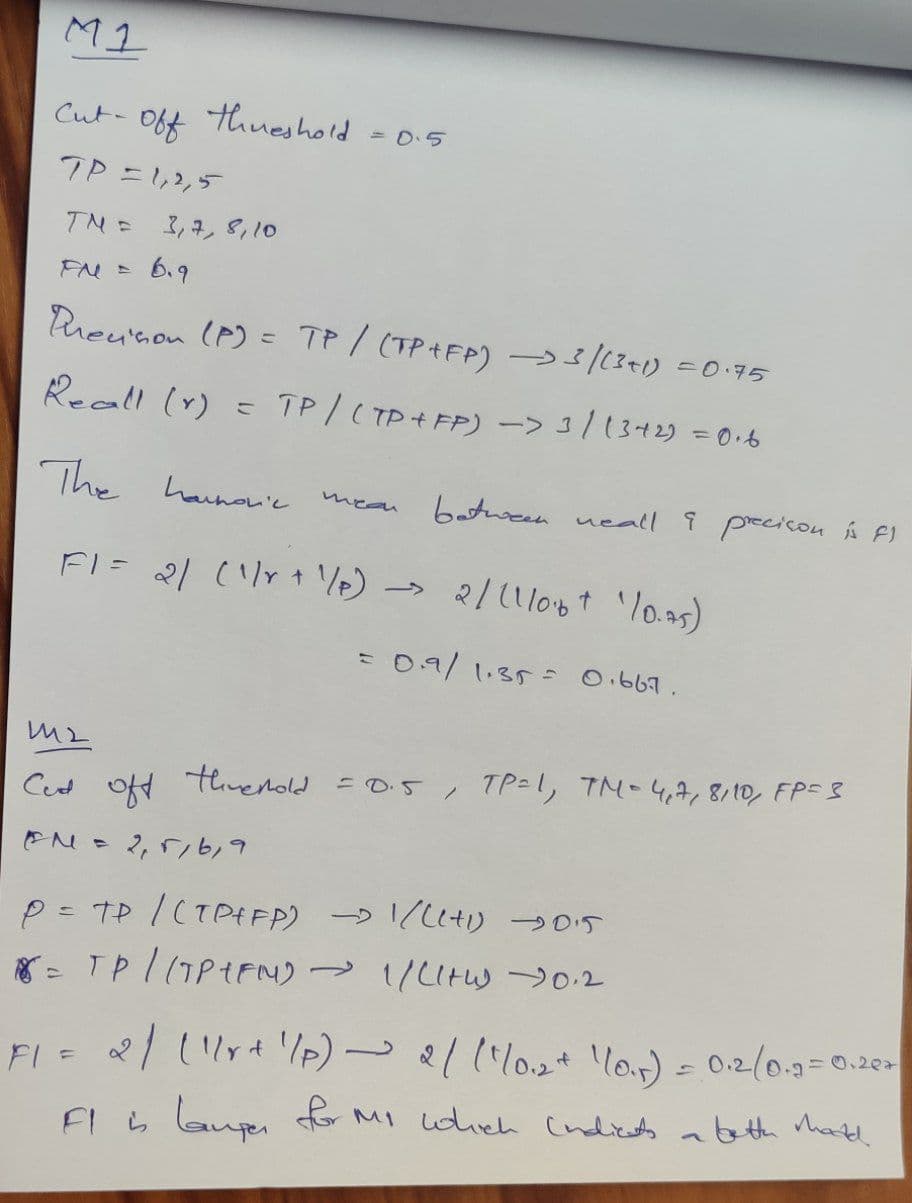
33j -1 -1 -1 -1 -1 -1 1 -1 -1 -l -1 -1 1 -1 -1 -1 -1 1 -l -l 1 1 -1 1

* xzab <— z ableCPredlcci ons , acz ua1\_va1ues )I

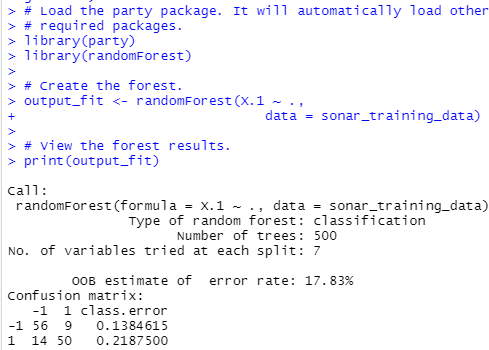


1. Do Chapter 5 textbook problem #17 (parts a and c only) on pages 322-323. Note that there is a typo in part c - it should read "Repeat the analysis for part (b)". We will do part b in class.





1. Compute the misclassification error on the training data for the Random Forest classifier to the last column of the sonar training data. Show your R code for doing this.



Misclassification error rate = 1 – ((True Positives + True Negatives) / Total observations)

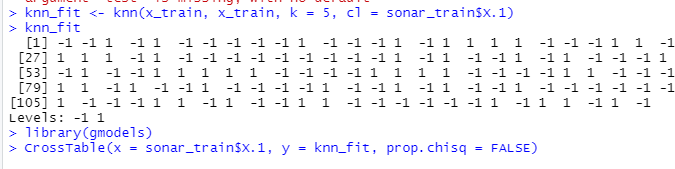
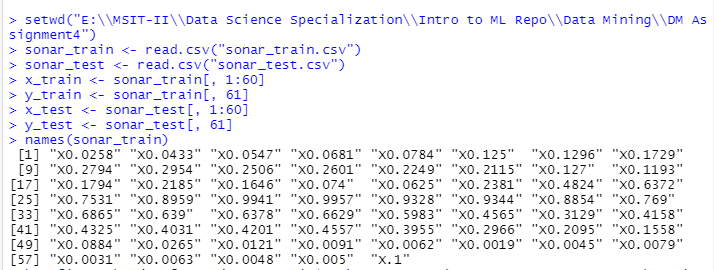
= 1 – ((56 + 50) / (56 + 9 + 14 + 50))

= 1 – (106 / 129)

= 1 – 0.8217054264

# = 0.1782945736

1. This question deals with sonar data
2. Use knn() for the k-nearest neighbor classifier for k=5 and k=6 to the last column of the sonar training data. Compute the misclassification error on the training data and also on the test data.



ceJJ coments



N I

N / ROw TOt I

N / cot mot et I

N / Table Tot al I



rotas observations in rable: 129



s onar\_rr a5 n$x. 1 l —+ l ROw Tot a3



—\* 1 1 6 | g5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| I | | 0.908 | I | 0.092 | | | 0.504 |
| I | | 0.738 | I | 0.122 |  |  |
| I | | 0.457 | I | 0.047 |  |  |
| \* | I \*\* | | 1 | 43 | | | 64 |
|  | I 0.328 | | I | 0.672 | | | O.49g |
|  | I 0.263 | | I | 0.878 |  |  |
|  | I 0.163 | | I | 0.333 |  |  |
| Coluon Top 1 | 1 | 80 | 1 | 49 | | | 129 |
|  | I | O. 620 | 1 | 0. 380 |  |  |



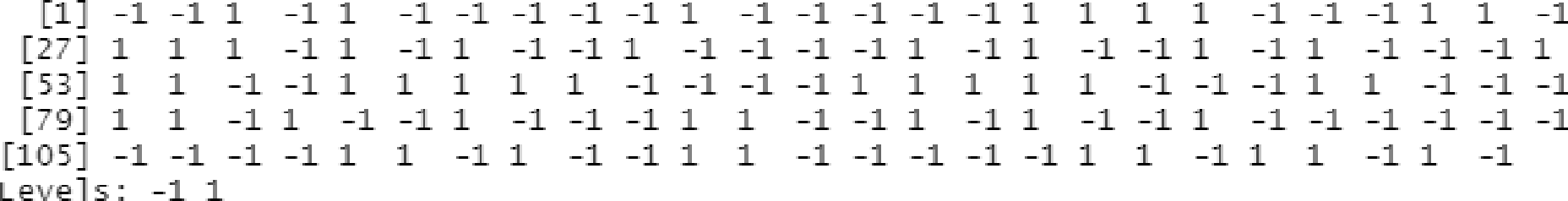
* summary(knmfit)

80 49

* n1 *s\_error\_Yr a\ n* = 1— s um {knn \_f1 z == y\_zrañ n) /1ength {y\_t r ai n)
* n1 s\_error\_tr a n

§1] 0. 2093023

* knri\_f t: <— knn (x\_t r zti n , x\_7r añ n , k 6, cl sonar\_t r zti n$x. I)



* 11 brary(gmodels)
* Cr os slabl e(x = s onzr\_t r zi nsx. 1 , y knn\_fJ t: , prop. chi 3q FALSE}

cel l cont ents

N

X / RO ' Tot al

\ / col Total

N / Tnb l e Tom al

yot6l absei- Vnt i ons n Tabl e : 12 9

I k nn\_f1t

s on6i-\_ti- ai n$x. 1 I —1 1 | ROW' Tot nl

I I I

-1 I \*\* I \*\* I

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | I | 0.877 | 0.123 | | | 0. 304 |
| I | 0.750 | 0.151 |  |  |
| I | 0.442 | 0.062 |  |  |
| 1 | 1 | 19 | 4 5 | | | C4 |
|  | I | 0.297 | 0.703 | | | 0.49C |
|  | I | 0.250 | 0.849 |  |  |
|  | I | 0.14? | 0.349 |  |  |
| Col u 1n TDt n1 | 1 | 7 6 | 5 3 | | | 12 9 |
|  | 1 | 0. 5 89 | 0. 411 |  |  |

* n1s\_ei- i- oi-\_t r ni n\_k C = 1— s u u(knn\_f ñ t == y\_ti- nJ n) /l e ngt h {y \_t r &ñ n}
* n1s\_ei- i- oi-\_t r ni n\_k C

§1§ 0. 2093023

knri\_f i t ti- ni n\_k 3 knn {x\_t r BJ n , x\_ti- ai n, k knn\_f i t\_ti- ni n\_k ñ

|  |  |  |  |
| --- | --- | --- | --- |
| [1] -1 -1 1 -1 1 | -1 -1 -1 -L-1 1 -1 -1 -1 1 | -1 1 | 1 L 1 -1 -1 -1 1 1 -1 |
| 2 ] 1 1 1 -1 1 | -1 -1 -1 -1 -1 -1 -1 -1 -1 1 | -1 1 | -1 -1 1 -1 1 -I -1 -1 1 |
| 53] -1 1 -1 -1 1 | 1 1 1 1 -1 -1 -1 -1 1 1 | 1 1 | -1 -1 -1 -1 1 1 -1 -1 -1 |
| ?9] 1 1 -1 1 -1 | -1 1 -1 -1 -1 -1 1 -1 -1 1 | -1 1 | -1 -1 1 -1 -1 -1 -1 -1 -1 |
| [105] 1 -1 -1 -1 1 | 1 -1 1 -1 -1 1 1 -1 -1 -1 | -1 -1 | -1 1 -1 1 1 -1 1 -1 |
| Levels: -1 1 |  |  |  |

*5 ,* c l 's onBr\_t: r aJ n $X. 1}

* + library(9models)
* cr oss Tab e (x = s onar\_z r at n Sx . 1 , y = knri\_f i f\_t r ni n\_k 5 , pr op. ch1s q = r LTE}

cell contents

N N / RoM TOt 61 N / col for&l

N / Table Tot ml

Tonal obser\law J ons 1n vzbl e : LZ9

knn\_fi t\_t r at n\_k 3

s onar\_tra1n$X. 1 | -1

-1 | S9

0. 908

0. Z38

0.4SZ

1 | 21

0. 32 8

0. 2 83

0. 163

1 | Row Total 6 | 65

0. 092 I 0.504 I

0. 122

0.047

43 | 64

O. 672 | 0. 496

0. B7B

0. 333

Co umn Total | 80 | 49

I o. ego I o. zso I

12 9



* ml ecu as sJ f y\_error\_t r of n\_k S 1 {sum{ knn\_f *t\_tra!* n\_k S y\_tr at n) ' ength (y\_tr at n}}
* misc1ass4fy\_error\_trRin\_ks

Ij 0.2093023

knri\_fJ t\_traJri\_kd <- knn {x\_t.r at n , x\_t:rat n, k d, ct s anar\_1zr at n $x. 1)

* knri\_f t\_traJri\_k6

|  |  |  |  |
| --- | --- | --- | --- |
| [1j -1 -1 1 -1 1 -1 -1 -1 -1 -1 1 -1 -1 -1 1 | -1 1 | 1 1 1 -1 | -1 -1 -1 1 -1 |
| *[27j* 1 1 1 -1 1 -1 -1 -1 -1 1 -1 -1 -1 -1 1 | -1 1 | -1 -1 1 -1 | 1 1 1 -1 1 |
| [53j -1 1 -1 -1 1 1 1 1 1 —1 -1 -1 -1 1 1 | 1 1 | 1 —1 —1 —1 | 1 1 -1 -1 -1 |
| [79j 1 1 -1 1 -1 -1 1 -1 -1 -1 -1 1 -1 -1 1 | -1 1 | -1 -1 1 -1 | -1 -1 -1 -1 -1 |
| [105j 1 -1 -1 -1 1 -1 -1 -1 -1 -1 1 1 -1 -1 -1 | -1 -1 | 1 1 -1 1 | 1 1 1 -1 |
| Levels: -1 1 |  |  |  |
| * library(gmodels) |  |  |  |

› crossiable(x = sonar rr&inSx.1, y = knifii rraiik6, prop.chisq = FALSE)

cev conrenrs

N

N / ROvY TOt61

N / cot +oz EU

N / Table Total

Total Observations in TRble: 129

knn\_fi t\_t r at n\_k6

sonar\_rrain$x.1 | -1 I | Row Tot a4

-1 | 58 | 7 | 68

l 0.892 l 0.108 | 0.504

0.F53 | 0.135

0.450 | 0.054

1 | 19

0. 297

0. 247

0.147

Col unn Tota1 | 77

O . 597

45 I

0.Z03

0.865

0.349

32

0 . 40 3



O. 496

129

* + nJ scl as st *fy\_error\_z* r z1n\_k6 1 {sum(knn\_f JI:\_t:raJ n\_k6 y\_zr aln) , l engt:h(y\_t:raln})
  + nJ scl as sJ *fy\_error\_t* r of n\_k6

Ij 0.2015504

knri\_fJ t:\_vest\_k5 <- knn(x\_t end, x\_ces7 , k 5, ct 3onzr\_1z es $x. L) knri\_fJ t:\_t:est\_k'i

§L] -1 -1 1 -1 1

-1 1 1 1 -1

[53] 1 -1 -1 -1 -1

Level s: -1 1

› Jibrary(gmodeJsj

-1 1 -1 -1 -1 1 -1 1 -1 1 -1 -1 1 -1 -1 -1 -1 -1 -1 1 -1

1 -1 -1 1 -1 -1 1 -1 1 -1 -1 -1 1 -1 1 1 1 -1 -1 -1 1

-1 -1 1 -1 -1 -1 -1 -1 1 -1 -1 -1 -1 -1 -1 -1 1 1 -1 -1

* Cr o3 3Tabl e (x = 3 onzr\_t e3t $x.1 , y = knrl\_f ñ 7\_t es t\_k 5 , prop. ch1s q = FALSE)

cell contents

N

N / RO\N TOt61

N / cot +avztJ

N / Table +avztJ

Total Observations in Table: 77

knn\_fi z\_1esz\_k 3

s onar\_z est $x. 1 | -1 1 | ROW TOIdl

-1 | 40 | 4 | 44

0. 909 | 0. D91 | 0. 571

0.F69 | 0.160

0.519 | 0.032

1 | 12

0.364

0.231

0.1F6

column Total | 52

*0. 67* S

21

0. 6 36 I

O. 840

O. Z7’ 3 I

25

0. 325

3 3

0. 4 29



77



i misclassify\_error\_test\_ks I (sum(knn\_fit\_test\_k5 == y\_test) / length(y\_test))

* misclassify\_error\_test\_ks

[Ij 0.2077922

* knn\_fJ t:\_t:est\_k6 knn(x\_t est , x\_cest , k 6, ct 3anztr\_t.est.Sx. 1)
* knn\_fJ t:\_t:est\_k6

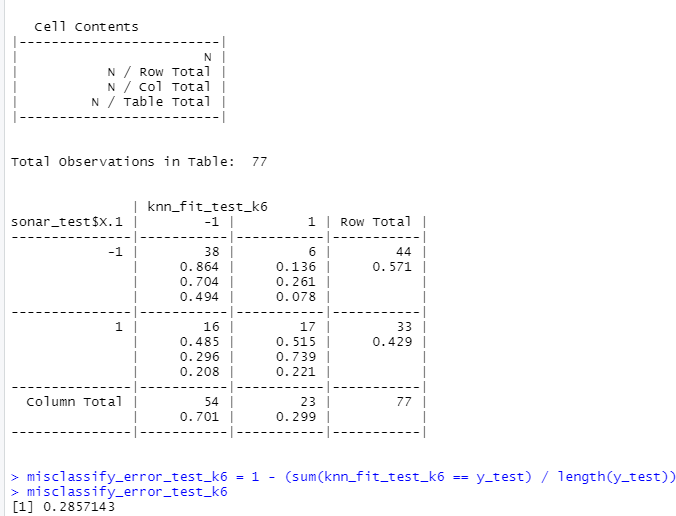
[1j -1 -1 1 -1 1 -1 1 -1 -1 -1 1 -1 1 -1 -1 -1 -1 1 -1 -1 -1 -1 -1 -1 1 -1

[Z7j 1 -1 1 1 -1 1 -1 1 -1 1 -1 1 1 1 -1 -1 -1 1 -1 1 1 1 -1 -1 -1 -1

[53j 1 -1 -1 -1 -1 -1 -1 1 —1 —1 —1 -1 -1 -1 -1 -1 -1 -1 —1 —1 —1 1 -1 -1 -1

Level s: -1 1

* Library(gmodeJs)
* crossTabJe(x = sonar\_test$x.1, y = knn\_f4i\_test\_k6, prop.chisq = PALSE)



1. Repeat part a using the exact same R code a few times. Explain why both the training errors and the test errors often change for k=6 but not for k=5. Hint: Read the help on the knn function if you do not know.

If there exists any tie in the values of Kth nearest vector then all the candidates are included in vote.

For the value, k = 6, there are ties for 6th nearest vector.

Hence there are changes in the misclassification error rate as all the candidates will be included in the vote.