Solutions to Homework Module 13

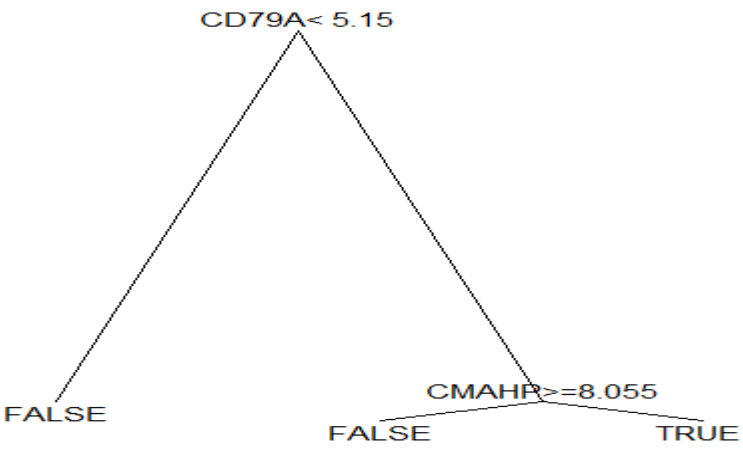
Solution to Problem 1

1a)

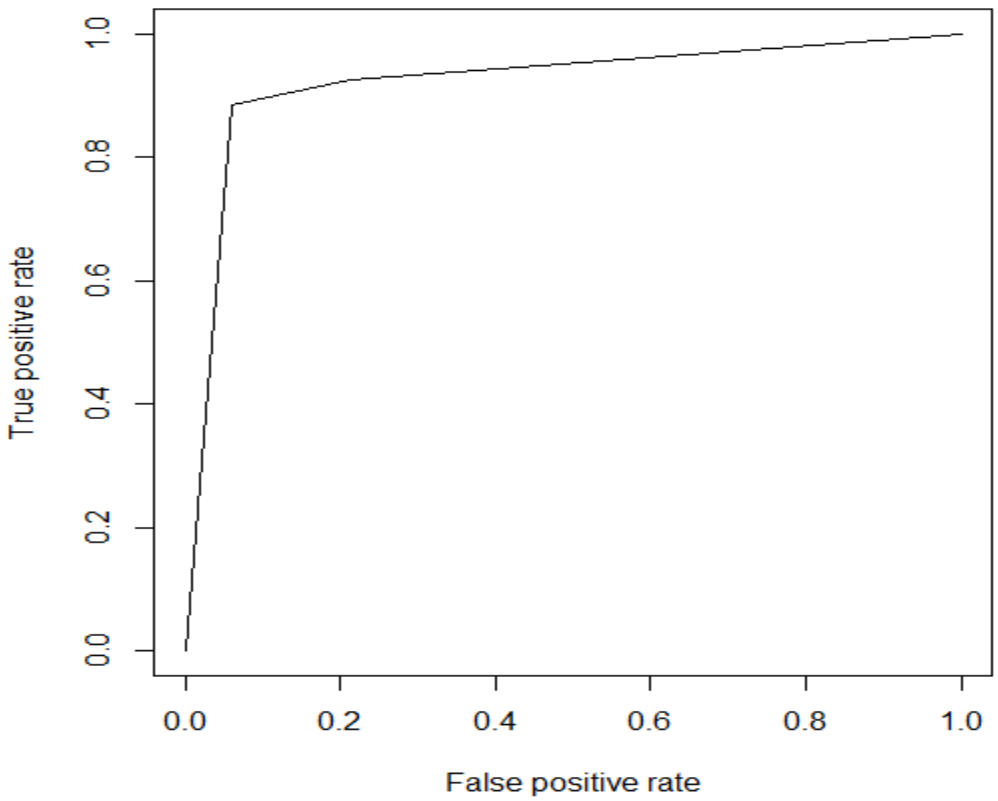
```
#defining indicator variable
> IsB <- factor(ALL$BT %in% c("B","B1","B2","B3","B4"))
```

1b)

```
> table(predicted.part, stages)
              stages
predicted.part FALSE TRUE
         FALSE    31   11
         TRUE      2   84
```



Solutions to Homework Module 13



1c)

```
> print("the empirical misclassification rate is")

[1] "the empirical misclassification rate is"
> (11+2)/128
[1] 0.1015625
> # fpr =
> print("the false positive rate is")
[1] "the false positive rate is"
> 2/(31+2)
[1] 0.06060606
> # fnr =
> print("the false negetive rate is")
[1] "the false negetive rate is"
> 11/(84+11)
[1] 0.1157895
> # Specificity = tnr =
> print("the specificity is")
[1] "the specificity is"
> 31/(2+31)
[1] 0.9393939
> print("the area under curve AUC is:")
[1] "the area under curve AUC is:"
> performance(pred,"auc")
Slot "y.values":
[[1]]
[1] 0.922807
```

Solutions to Homework Module 13

1d)

```
> fnr.true
[1] 0.08310606
```

The estimate fnr is 0.08310606

1e)

Logistic regression:

```
> table(pred.B1, IsB1)
       IsB1
pred.B1  B not_B
  B     90     6
  not_B  5    27
```

80% confidence interval for the coefficient of gene "39317_at"

```
> confint(fit.data, level=0.8)
Waiting for profiling to be done...
                    10 %       90 %
(Intercept) -13.767525 -2.8118382
X39317_at    -1.427390 -0.6047588
X38018_g_at   2.120174  3.9861802
```

CI is (-1.427, -0.6047)

1f)

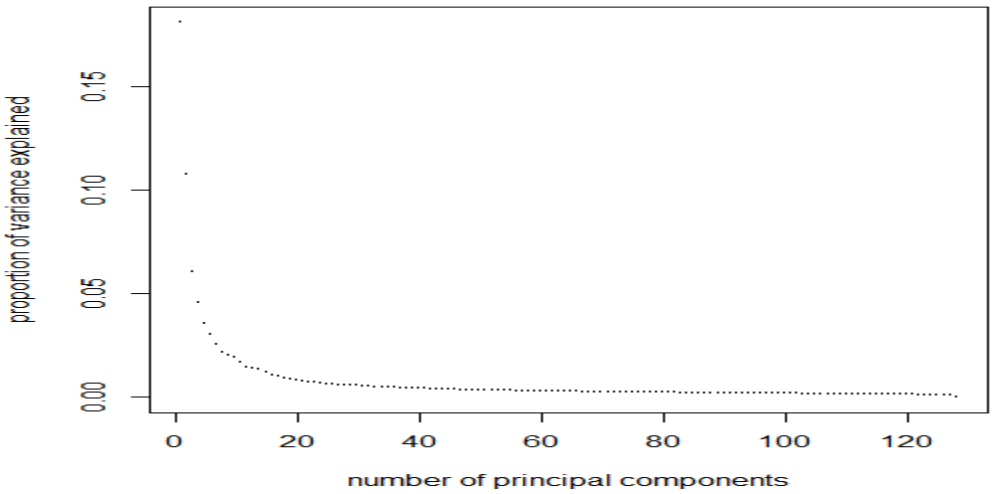
```
> mcr.cv
[1] 0.09375
```

The estimated mcr 0.09375

1g)

There is a rapid drop in the proportion in the beginning as seen until about 5 and then slows down near 15. Thus we can conclude that we should be using about 15PCs

Solutions to Homework Module 13




```
> summary(pca.ALL)
Importance of components:
                           PC1      PC2      PC3     PC4      PC5      PC6
PC7
Standard deviation     47.8103 36.9157 27.73208 24.0204 21.29449 19.64675 18.
00937
Proportion of Variance  0.1811  0.1079  0.06092  0.0457  0.03592  0.03057  0.
02569
Cumulative Proportion   0.1811  0.2890  0.34991  0.3956  0.43153  0.46211  0.
48780
                           PC8      PC9     PC10     PC11     PC12     PC13
PC14
Standard deviation     16.52815 16.08110 15.68492 14.73970 13.49120 13.46128
13.1528
Proportion of Variance  0.02164  0.02048  0.01949  0.01721  0.01442  0.01435
0.0137
Cumulative Proportion   0.50943  0.52992  0.54940  0.56661  0.58103  0.59538
0.6091
                          PC15     PC16     PC17     PC18     PC19     PC20
PC21
Standard deviation     12.50326 11.62651 11.33948 10.95969 10.56977 10.27269
9.98280
Proportion of Variance  0.01238  0.01071  0.01018  0.00951  0.00885  0.00836
0.00789
Cumulative Proportion   0.62147  0.63217  0.64236  0.65187  0.66072  0.66908
0.67698
                          PC22    PC23    PC24    PC25    PC26    PC27    PC2
8    PC29
Standard deviation     9.76071 9.69351 9.35307 9.07879 8.97473 8.85997 8.7242
1 8.59119
Proportion of Variance 0.00755 0.00744 0.00693 0.00653 0.00638 0.00622 0.0060
3 0.00585
Cumulative Proportion  0.68452 0.69196 0.69889 0.70542 0.71180 0.71802 0.7240
5 0.72989
                          PC30    PC31    PC32    PC33    PC34    PC35    PC3
6   PC37
```

Solutions to Homework Module 13

```
Standard deviation     8.53111 8.27069 8.23309 8.08897 8.07028 7.83775 7.8023
5 7.7043
Proportion of Variance 0.00576 0.00542 0.00537 0.00518 0.00516 0.00487 0.0048
2 0.0047
Cumulative Proportion  0.73566 0.74108 0.74645 0.75163 0.75679 0.76165 0.7664
8 0.7712
                           PC38    PC39    PC40    PC41    PC42    PC43    PC4
4    PC45
Standard deviation     7.56167 7.54920 7.47852 7.40003 7.33338 7.21207 7.1816
5 7.07957
Proportion of Variance 0.00453 0.00451 0.00443 0.00434 0.00426 0.00412 0.0040
9 0.00397
Cumulative Proportion  0.77571 0.78022 0.78465 0.78899 0.79325 0.79737 0.8014
5 0.80542
                           PC46    PC47    PC48    PC49    PC50    PC51    PC5
2    PC53
Standard deviation     7.00761 6.88692 6.86142 6.81489 6.79102 6.70541 6.6873
7 6.61719
Proportion of Variance 0.00389 0.00376 0.00373 0.00368 0.00365 0.00356 0.0035
4 0.00347
Cumulative Proportion  0.80931 0.81307 0.81680 0.82048 0.82413 0.82769 0.8312
3 0.83470
                           PC54    PC55   PC56    PC57    PC58    PC59   PC60
PC61
Standard deviation     6.58283 6.51855 6.4543 6.42488 6.40876 6.33493 6.2549
6.22671
Proportion of Variance 0.00343 0.00337 0.0033 0.00327 0.00325 0.00318 0.0031
0.00307
Cumulative Proportion  0.83813 0.84150 0.8448 0.84807 0.85132 0.85450 0.8576
0.86067
                           PC62    PC63    PC64    PC65   PC66    PC67    PC68
PC69
Standard deviation     6.19791 6.17728 6.13009 6.07863 6.0457 6.00987 5.98143
5.95744
Proportion of Variance 0.00304 0.00302 0.00298 0.00293 0.0029 0.00286 0.00283
0.00281
Cumulative Proportion  0.86371 0.86674 0.86971 0.87264 0.8755 0.87839 0.88123
0.88404
                           PC70    PC71    PC72    PC73    PC74    PC75    PC7
6    PC77
Standard deviation     5.87113 5.84515 5.82817 5.76546 5.74950 5.69443 5.6701
9 5.66516
Proportion of Variance 0.00273 0.00271 0.00269 0.00263 0.00262 0.00257 0.0025
5 0.00254
Cumulative Proportion  0.88677 0.88948 0.89217 0.89480 0.89742 0.89999 0.9025
3 0.90508
                           PC78    PC79    PC80    PC81   PC82    PC83    PC84
PC85
Standard deviation     5.64871 5.60202 5.56279 5.53503 5.5092 5.48405 5.44800
5.42787
Proportion of Variance 0.00253 0.00249 0.00245 0.00243 0.0024 0.00238 0.00235
0.00233
Cumulative Proportion  0.90760 0.91009 0.91254 0.91497 0.9174 0.91975 0.92210
0.92444
                           PC86    PC87    PC88    PC89    PC90    PC91    PC9
2    PC93
Standard deviation     5.36975 5.33498 5.29293 5.28715 5.25939 5.22066 5.1895
3 5.17319
Proportion of Variance 0.00228 0.00225 0.00222 0.00221 0.00219 0.00216 0.0021
3 0.00212
Cumulative Proportion  0.92672 0.92898 0.93119 0.93341 0.93560 0.93776 0.9398
9 0.94201
                          PC94    PC95    PC96    PC97    PC98   PC99   PC100
PC101
```

Solutions to Homework Module 13

```
Standard deviation     5.1529 5.10942 5.09202 5.07675 5.03399 5.0260 4.99505
4.96053
Proportion of Variance 0.0021 0.00207 0.00205 0.00204 0.00201 0.0020 0.00198
0.00195
Cumulative Proportion  0.9441 0.94618 0.94824 0.95028 0.95228 0.9543 0.95626
0.95821
                          PC102   PC103   PC104   PC105   PC106   PC107   PC108
PC109
Standard deviation     4.92004 4.8988 4.86395 4.85237 4.81879 4.80002 4.72967
4.69107
Proportion of Variance 0.00192 0.0019 0.00187 0.00186 0.00184 0.00182 0.00177
0.00174
Cumulative Proportion  0.96013 0.9620 0.96390 0.96577 0.96761 0.96943 0.97120
0.97295
                          PC110   PC111   PC112   PC113   PC114   PC115   PC11
6   PC117
Standard deviation     4.68160 4.64703 4.61168 4.59981 4.56873 4.54519 4.4547
6 4.42230
Proportion of Variance 0.00174 0.00171 0.00168 0.00168 0.00165 0.00164 0.0015
7 0.00155
Cumulative Proportion  0.97468 0.97639 0.97808 0.97975 0.98141 0.98304 0.9846
2 0.98616
                          PC118   PC119   PC120   PC121   PC122   PC123   PC12
4   PC125
Standard deviation     4.38206 4.36821 4.30457 4.27980 4.25439 4.18120 4.1683
7 4.14619
Proportion of Variance 0.00152 0.00151 0.00147 0.00145 0.00143 0.00138 0.0013
8 0.00136
Cumulative Proportion  0.98769 0.98920 0.99066 0.99212 0.99355 0.99493 0.9963
1 0.99767
                          PC126   PC127     PC128
Standard deviation     4.00554 3.65401 1.033e-13
Proportion of Variance 0.00127 0.00106 0.000e+00
Cumulative Proportion  0.99894 1.00000 1.000e+00
```

1h)

```
> print("tpr.svm")
[1] "tpr.svm"
> tpr.svm
[1] 0.9894737
```

Sensitivity of the classifier 0.9894737

1i)

```
> mcrcv<- mean(mcrcvraw)#average the mcr over all n rounds.
> mcrcv
[1] 0.0390625
```

Estimated MCR is 0.0390625

Solutions to Homework Module 13

1j)

When we compare classifiers (e) and (h), I would say (h) is better as it has better sensitivity [0.9894] than (e) [90/(90+5) = 0.9473]

Solution to problem 2

```
[1] "K = 1"
[1] "Classifier : MCR, leave-one-out MCR"
[1] "Logistic Regression: " "0.0733333333333333" "0.0733333333333333"
[1] "SVM: " "0.0733333333333333" "0.08"
[1] "Classification Tree: " "0.0666666666666667" "0.106666666666667"

[1] "K = " "2"
[1] "Classifier : MCR, leave-one-out MCR"
[1] "Logistic Regression: " "0.08" "0.08"
[1] "SVM: " "0.0866666666666667" "0.0866666666666667"
[1] "Classification Tree: " "0.0666666666666667" "0.106666666666667"

[1] "K = " "3"
[1] "Classifier : MCR, leave-one-out MCR"
[1] "Logistic Regression: " "0.0133333333333333" "0.0266666666666667"
[1] "SVM: " "0.0266666666666667" "0.0466666666666667"
[1] "Classification Tree: " "0.0666666666666667" "0.14"

[1] "K = " "4"
[1] "Classifier : MCR, leave-one-out MCR"
[1] "Logistic Regression: " "0.0133333333333333" "0.02"
[1] "SVM: " "0.02" "0.0266666666666667"
[1] "Classification Tree: " "0.0666666666666667" "0.14"
```

Based on the above data obtained, through analysis, we can see that the best logistic regression as well as smallest cross validation error was given by K=4 ,

thus we can say that K=4 gives best suitable principle component analysis.