Pengantar Bioinformatika

Fadhillah Putri Taha
H071171301
Program Studi Ilmu Komputer

Chapter 8
Classification Methods

Exercise 8.8

1. Classification tree of Golub data. Use recursive partitioning in rpart

   a) Find a manner to identify an optimal gene with respect the Golub data to prediction of the ALL AML patients.

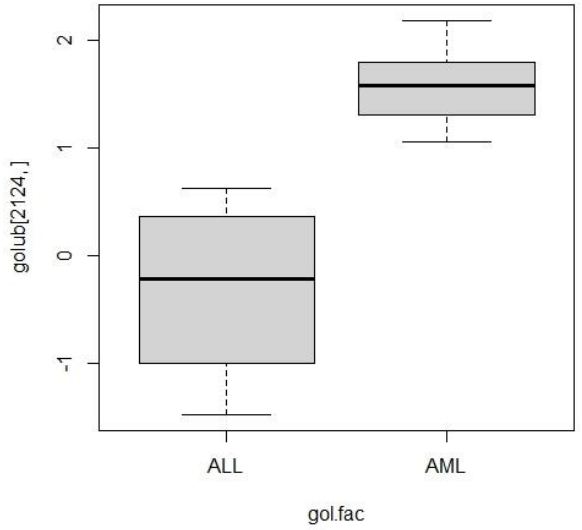
      Variable predictor dari nilai ekspresi gen adalah gen CCND3 Cyclin D3. Sintaks program dengan menggunakan bahasa pemrograman R, sebagai berikut:

```
> #Ex 3
> library(rpart);data(golub); library(multtest)
> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
> gol.rp <- rpart(gol.fac ~ golub[1042,] , method="class")
> predictedclass <- predict(gol.rp, type="class")
> table(predictedclass, gol.fac)
              gol.fac
predictedclass ALL AML
           ALL  25   1
           AML   2  10
> predict(gol.rp,type="class")
  1   2   3   4   5   6
   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25
ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL AML ALL ALL ALL AML ALL ALL ALL ALL
 26  27  28  29  30  31  32  33  34  35  36  37  38
ALL ALL AML ALL AML AML AML AML AML AML AML AML AML
Levels: ALL AML
> boxplot(golub[2124,] ~gol.fac)
> summary(gol.rp)
Call:
rpart(formula = gol.fac ~ golub[1042, ], method = "class")
  n= 38
         CP nsplit rel error    xerror      xstd
1 0.7272727      0 1.0000000 1.0000000 0.2541521
2 0.0100000      1 0.2727273 0.3636364 0.1719828

Variable importance
golub[1042, ]
          100
```

Fadhillah Putri Taha
H071171301
Program Studi Ilmu Komputer

```
Node number 1: 38 observations,    complexity param=0.7272727
  predicted class=ALL  expected loss=0.2894737  P(node) =1
    class counts:    27    11
   probabilities: 0.711 0.289
  left son=2 (26 obs) right son=3 (12 obs)
  Primary splits:
      golub[1042, ] < 1.198515 to the right, improve=10.37517, (0 missing)
Node number 2: 26 observations
  predicted class=ALL  expected loss=0.03846154  P(node) =0.6842105
    class counts:    25     1
   probabilities: 0.962 0.038
Node number 3: 12 observations
  predicted class=AML  expected loss=0.1666667  P(node) =0.3157895
    class counts:     2    10
   probabilities: 0.167 0.833
> golub.gnames[2124,]
[1] "4847"      "Zyxin"     "X95735_at"
> predict(gol.rp, type="prob")
          ALL         AML
1  0.9615385 0.03846154       20 0.9615385 0.03846154
2  0.9615385 0.03846154       21 0.1666667 0.83333333
3  0.9615385 0.03846154       22 0.9615385 0.03846154
4  0.9615385 0.03846154       23 0.9615385 0.03846154
5  0.9615385 0.03846154       24 0.9615385 0.03846154
6  0.9615385 0.03846154       25 0.9615385 0.03846154
7  0.9615385 0.03846154       26 0.9615385 0.03846154
8  0.9615385 0.03846154       27 0.9615385 0.03846154
9  0.9615385 0.03846154       28 0.1666667 0.83333333
10 0.9615385 0.03846154       29 0.9615385 0.03846154
11 0.9615385 0.03846154       30 0.1666667 0.83333333
12 0.9615385 0.03846154       31 0.1666667 0.83333333
13 0.9615385 0.03846154       32 0.1666667 0.83333333
14 0.9615385 0.03846154       33 0.1666667 0.83333333
15 0.9615385 0.03846154       34 0.1666667 0.83333333
16 0.9615385 0.03846154       35 0.1666667 0.83333333
17 0.1666667 0.83333333       36 0.1666667 0.83333333
18 0.9615385 0.03846154       37 0.1666667 0.83333333
19 0.9615385 0.03846154       38 0.1666667 0.83333333
```



Fadhillah Putri Taha
H071171301
Program Studi Ilmu Komputer

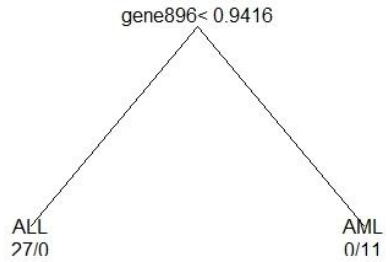
b) Explain what the code does.

Berdasarkan data Golub dkk (1999), nilai ekspresi gen dapat dibentuk suatu *decision tree*, yang memberikan prediktor, jika terdapat banyak prediktor maka fungsi **rpart** secara otomatis memilih gen yang berpengaruh dalam pengklasifikasian. Berdasarkan boxplot yang dihasilkan, dapat diketahui bahwa gen A merupakan prediktor ideal untuk membagi-bagi pasien ke dalam beberapa kelas. Data pasien ALL dan AML diklasifikasikan berdasarkan gen CCND3 Cyclin D3.

c) Use rpart to construct the classification tree with the genes that you found. Does it have perfect predictions?

Iya, itu merupakan prediksi yang sempurna dengan menentukan ekspresi gen sebagai variabel, kemudian mengubah operator **t**. Sehingga predictor dari dua kelas pasien dapat diprediksi dengan sempurna. Sintaks program dengan menggunakan bahasa pemrograman R, sebagai berikut:

```
> #Ex 4
> library(rpart);data(golub); library(multtest)
> row.names(golub)<- paste("gene", 1:3051, sep = "")
> goldata <- data.frame(t(golub[1:3051,]))
> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
> gol.rp <- rpart(gol.fac~., data=goldata, method="class", cp=0.001)
> plot(gol.rp, branch=0,margin=0.1); text(gol.rp, digits=3, use.n=TRUE)
> golub.gnames[896,]
[1] "2020"                    "FAH Fumarylacetoacetate" "M55150_at"
```

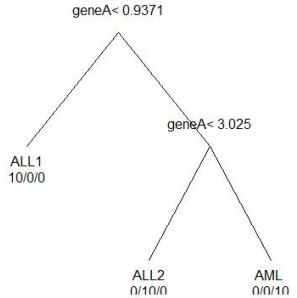


d) Find the row number of gene Gdf5, which is supposed not to have any relationship with leukemia. Estimate a classification tree and report the probability of misclassification. Give explanations of the results.

```
> grep("Gdf5",golub.gnames[,2])
[1] 2058
```

Fadhillah Putri Taha
H071171301
Program Studi Ilmu Komputer

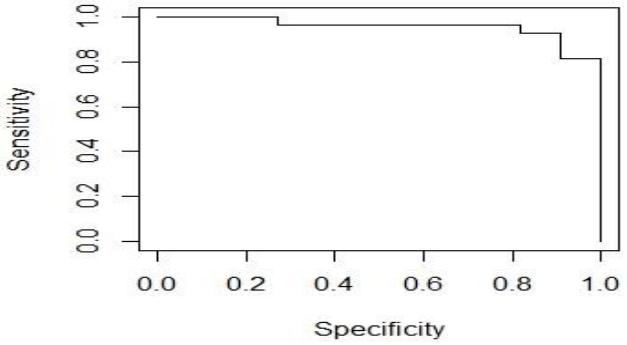
Ekspresi gen yang optimal. Misalkan data ekspresi microarray tersedia sehubungan dengan pasien yang menderita dari tiga jenis leukemia disingkat ALL1, ALL2, dan AML. Gen A memiliki nilai ekspresi dari populasi (kelompok pasien) N (0,0,52) untuk ALL1, N (2,0,52) untuk ALL2, dan N (4,0,52) untuk AML.



2. Sensitivity versus specificity.

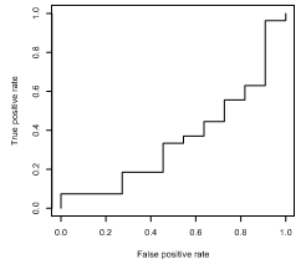
a) Produce a sensitivity versus specificity plot for the gene expression values of CCND3 Cyclin D3.

```
> library(multtest);library(ROCR);data(golub)
> golub.clchanged <- -golub.cl +1
> pred <- prediction(golub[1042,], golub.clchanged)
> perf <- performance(pred, "sens", "spec")
> plot(perf)
```



b) In what sense does it resemble Figure 8.2.

Titik potong ditentukan pada data terkecil, yaitu -0.74, yang kemudian pada titik ini semua pasien diuji positif, sehingga tingkat positif palsu adalah 11/11 dan tingkat positif sejati adalah 27/27. Ini ditunjukkan oleh titik akhir (1,1) pada plot.Dapat diamati bahwa tingkat positif sejati jauh lebih rendah ketika seseorang bergerak pada sumbu horizontal dari kiri ke kanan. Ini sesuai dengan area di bawah kurva 0,35, yang kecil. Ini menggambarkan bahwa gen dapat mengekspresikan perbedaan besar sehubungan dengan prediksi status penyakit pasien.

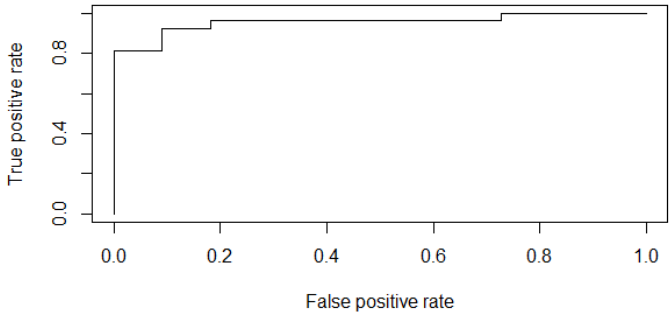


Fadhillah Putri Taha
H071171301
Program Studi Ilmu Komputer

c) Compute the area under the curve for sensitivity versus specificity curve.

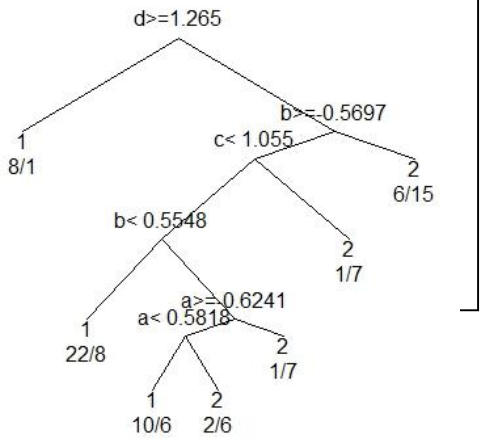
```
> #Ex 2

> library(ROCR)
> gol.true <- factor(golub.cl,levels=0:1,labels= c("TRUE","FALSE"))
> pred <- prediction(golub[1042,], gol.true)
> perf <- performance(pred, "tpr", "fpr" )
> performance(pred,"auc")
A performance instance
  'Area under the ROC curve'
> plot(perf)
```



a) Construct a factor with 100 values one and two and a matrix with predictor variables of 500 by 4 with values from the normal distribution. Use the first four letters of the alphabet for the column names.

```
> #3a
> library(rpart)
> predictors <- matrix(rnorm(100*4,0,1),100,4)
> colnames(predictors) <- letters[1:4]
> groups <- gl(2,50)
> simdata <- data.frame(groups,predictors)
> rp<-rpart(groups ~ a + b + c + d,method="class",data=simdata)
> predicted <- predict(rp,type="class")
> table(predicted,groups)
         groups
predicted  1  2
        1 38 10
        2 12 40
```



Fadhillah Putri Taha
H071171301
Program Studi Ilmu Komputer

b) Use rpart to construct a recursive tree and report the misclassification rate. Comment on the results.

```
> plot(rp, branch=0,margin=0.1); text(rp, digits=3, use.n=TRUE)
```

c) Do the same for support vector machines.

```
> library(e1071)
> svmest <- svm(predictors, groups, data=df, type = "C-classification", kernel = "linear")
> svmpred <- predict(svmest, predictors, probability=TRUE)
> table(svmpred, groups)
       groups
svmpred  1  2
      1 50 50
      2  0  0
```

d) Do the same for neural networks.

```
> library(nnet)
> nnest <- nnet(groups ~ ., data = simdata, size = 5,maxit = 500, decay = 0.01, MaxNWts = 5000)
# weights:  31
initial  value 72.875957
iter  10 value 61.281306
iter  20 value 48.491034
iter  30 value 44.000429
iter  40 value 41.593852
iter  50 value 41.253020
iter  60 value 41.238766
iter  70 value 41.238492
iter  80 value 41.238437
final  value 41.238433
converged
```

Fadhillah Putri Taha
H071171301
Program Studi Ilmu Komputer

e) Think through your results and comment on these.

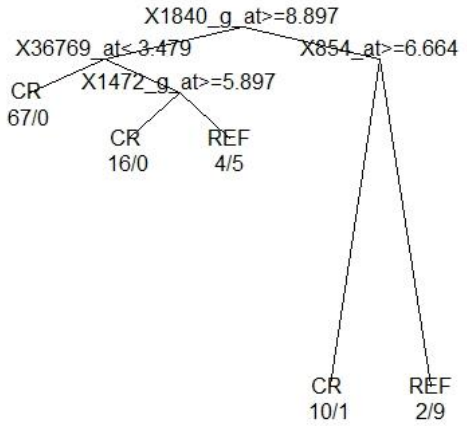
Tingkat kesalahan klasifikasi akan menurun berbanding terbalik dengan predictor yang diberikan. Semakin banyak predictor, semakin kecil tingkat kesalahan pada klasifikasi.

4. Prediction of achieved remission. For the ALL data from its ALL library the patients are checked for achieving remission. The variable ALL$CR has values CR (became healthy) and REF (did not respond to therapy; remain ill).

a) Construct an expression set containing the patients with values on the phenotypical variable remission and the gene expressions with a significant p-value on the t-test with the patient groups CR or REF.

b) Use recursive partitioning to predict the remission. Report the misclassification rate and the names of the genes that play a role in the tree.

```
> ALLrem <- ALL[,which(pData(ALL)$remission %in% c("CR","REF"))]
> remfac <-factor(pData(ALLrem)$remission)
> pano <- apply(exprs(ALLrem),1,function(x) t.test(x ~ remfac)$p.value)
> names <- featureNames(ALLrem)[pano<.001]
> ALLremsel<- ALLrem[names,]
> data <- data.frame(t(exprs(ALLremsel)))
> all.rp <- rpart(remfac ~., data, method="class", cp=0.001)
> plot(all.rp, branch=0,margin=0.1); text(all.rp, digits=3, use.n=TRUE)
> rpart.pred <- predict(all.rp, type="class")
> table(rpart.pred,remfac)
          remfac
rpart.pred CR REF
      CR  93   1
      REF  6  14
> 7/(93+1+6+14)
[1] 0.06140351
> mget(c("1840_g_at","36769_at","1472_g_at","854_at"), env = hgu95av2GENENAME)
$`1840_g_at`
[1] "RAN, member RAS oncogene family"
$`36769_at`
[1] "RB binding protein 5, histone lysine methyltransferase complex subunit"
$`1472_g_at`
[1] "MYB proto-oncogene, transcription factor"
$`854_at`
[1] "BLK proto-oncogene, Src family tyrosine kinase"
```



Fadhillah Putri Taha
H071171301
Program Studi Ilmu Komputer

5. Gene selection by area under the curve. A strategy of selecting genes is to compute the auc for each gene and to use the best 10 for further investigation. Compute the auc for each row with gene expressions of the Golub at al. (1999) data. Collect these in a vector and select the ten best. Is ”CCND3 Cyclin D3” among these?

```
> #5

> library(ROCR); data(golub, package = "multtest")
> gol.true <- factor(golub.cl,levels=0:1,labels= c("TRUE","FALSE"))
> auc.values <- apply(golub,1,
+     function(x) performance(prediction(x, gol.true),"auc")@y.values[[1]])
> o <- order(auc.values,decreasing=TRUE)
> golub.gnames[o[1:25],2]
 [1] "TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)"
 [2] "Macmarcks"
 [3] "VIL2 Villin 2 (ezrin)"
 [4] "TOP2B Topoisomerase (DNA) II beta (180kD)"
 [5] "C-myb gene extracted from Human (c-myb) gene, complete primary cds,
         and five complete alternatively spliced cds"
 [6] "RETINOBLASTOMA BINDING PROTEIN P48"
 [7] "RB1 Retinoblastoma 1 (including osteosarcoma)"
 [8] "CCND3 Cyclin D3"
 [9] "ALDR1 Aldehyde reductase 1 (low Km aldose reductase)"
[10] "T-COMPLEX PROTEIN 1, GAMMA SUBUNIT"
[11] "SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)"
[12] "Inducible protein mRNA"
[13] "Translational initiation factor 2 beta subunit (eIF-2-beta) mRNA"
[14] "NUCLEOLYSIN TIA-1"
[15] "Putative enterocyte differentiation promoting factor mRNA, partial cds"
[16] "IEF SSP 9502 mRNA"
[17] "ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain"
[18] "PROTEASOME IOTA CHAIN"
[19] "X-LINKED HELICASE II"
[20] "Stimulator of TAR RNA binding (SRB) mRNA"
[21] "MYL1 Myosin light chain (alkali)"
[22] "Transcriptional activator hSNF2b"
[23] "HKR-T1"
[24] "ADA Adenosine deaminase"
[25] "Transcriptional activator hSNF2b"
```