DT

KWONHYUNJIN

2018년 9월 19일

#tree: tree패키지

```
bc <- read.xlsx("New_version_breast_cancer.xlsx",1)
head(bc)
```

```
##   NA. age tumor.size inv.nodes deg.malig irradiat   Class lt40 ge40
## 1   1   4          4         1         3        0   recur    0    0
## 2   2   5          4         1         1        0 norecur    0    1
## 3   3   5          8         1         2        0   recur    0    1
## 4   4   4          8         1         3        1 norecur    0    0
## 5   5   4          7         2         2        0   recur    0    0
## 6   6   5          6         2         2        1 norecur    0    0
##   premeno node.capse no.node.capse breast.left breast.right quad.cen
## 1       1          1             0           0            1        0
## 2       0          0             1           0            1        1
## 3       0          0             1           1            0        0
## 4       1          1             0           0            1        0
## 5       1          1             0           1            0        0
## 6       1          0             1           0            1        0
##   quad.Ll quad.Lu quad.Rl quad.Ru
## 1       0       1       0       0
## 2       0       0       0       0
## 3       1       0       0       0
## 4       1       0       0       0
## 5       0       0       0       1
## 6       0       1       0       0
```

#set.seed란 랜덤한 값을 시작하기 전에 사용하면 이후에도 같은 값으로 랜덤값을 갖는다. #동일한 랜덤값을 계속해서 받기위해 #과연 필요할지에 대해 고찰

```
set.seed(123)
bc_shuffle <- bc[sample(nrow(bc)), ]
head(bc_shuffle)
```

```
##     NA. age tumor.size inv.nodes deg.malig irradiat   Class lt40 ge40
## 80   80   3          1         1         2        0 norecur    0    0
## 218 218   3          5         1         3        1   recur    0    0
## 113 113   4          8         4         2        1 norecur    0    0
## 242 242   5          6         1         1        0 norecur    0    0
## 257 257   5          1         1         2        0 norecur    0    1
## 13   13   5          7         1         1        0 norecur    0    1
##     premeno node.capse no.node.capse breast.left breast.right quad.cen
## 80        1          0             1           0            1        1
## 218       1          0             1           1            0        0
## 113       1          1             0           0            1        0
## 242       1          0             1           1            0        0
## 257       0          0             1           1            0        1
## 13        0          0             1           0            1        1
##     quad.Ll quad.Lu quad.Rl quad.Ru
## 80        0       0       0       0
## 218       0       1       0       0
## 113       0       1       0       0
## 242       1       0       0       0
## 257       0       0       0       0
## 13        0       0       0       0
```

#현재 bc 데이터 프레임에 id라는 컬럼이 필요없다. #bc_shuffle에서 1번컬럼은 제외하고 나머지 컬럼을 bc2에 할당한다.

```
bc2 <- bc_shuffle[-1]
head(bc2)
```

```
##     age tumor.size inv.nodes deg.malig irradiat   Class lt40 ge40 premeno
## 80    3          1         1         2        0 norecur    0    0       1
## 218   3          5         1         3        1   recur    0    0       1
## 113   4          8         4         2        1 norecur    0    0       1
## 242   5          6         1         1        0 norecur    0    0       1
## 257   5          1         1         2        0 norecur    0    1       0
## 13    5          7         1         1        0 norecur    0    1       0
##     node.capse no.node.capse breast.left breast.right quad.cen quad.Ll
## 80           0             1           0            1        1       0
## 218          0             1           1            0        0       0
## 113          1             0           0            1        0       0
## 242          0             1           1            0        0       1
## 257          0             1           1            0        1       0
## 13           0             1           0            1        1       0
##     quad.Lu quad.Rl quad.Ru
## 80        0       0       0
## 218       1       0       0
## 113       1       0       0
## 242       0       0       0
## 257       0       0       0
## 13        0       0       0
```

#데이터를 train 과 test 9:1로 나눈후 train 에 대한 tree 제작 .

```
train_num<-round(0.9*nrow(bc2),0)
bc_train<-bc2[1:train_num,]
bc_test<-bc2[(train_num+1):nrow(bc2),]
```

```
btree<-tree(Class~.,bc_train)
btree
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 249 292.000 norecur ( 0.72691 0.27309 )
##    2) deg.malig < 2.5 180 174.500 norecur ( 0.81111 0.18889 )
##      4) tumor.size < 3.5 36   9.139 norecur ( 0.97222 0.02778 )
##        8) age < 3.5 5   5.004 norecur ( 0.80000 0.20000 ) *
##        9) age > 3.5 31   0.000 norecur ( 1.00000 0.00000 ) *
##      5) tumor.size > 3.5 144 155.000 norecur ( 0.77083 0.22917 ) *
##    3) deg.malig > 2.5 69  95.640 norecur ( 0.50725 0.49275 )
##      6) inv.nodes < 1.5 40  50.450 norecur ( 0.67500 0.32500 )
##       12) premeno < 0.5 24  24.560 norecur ( 0.79167 0.20833 )
##         24) age < 5.5 14   0.000 norecur ( 1.00000 0.00000 ) *
##         25) age > 5.5 10  13.860 norecur ( 0.50000 0.50000 ) *
##       13) premeno > 0.5 16  22.180 norecur ( 0.50000 0.50000 ) *
##      7) inv.nodes > 1.5 29  34.160 recur ( 0.27586 0.72414 ) *
```

#분류를 얻을때는 type="class"를 지정해야하지만, 기본 값이 class이므로 생략 가능

```
pred <- predict(btree,bc_test,type="class")
table(pred,bc_test$Class)
```

```
##
## pred      norecur recur
##   norecur      14     6
##   recur         1     7
```

#confusionMatrix(예측값, 결과값) 함수를 이용하면 정확도 확인

```
confusionMatrix(table(pred, bc_test$Class))
```

```
## Confusion Matrix and Statistics
##
##
## pred      norecur recur
##   norecur      14     6
##   recur         1     7
##
##                Accuracy : 0.75
##                  95% CI : (0.5513, 0.8931)
##     No Information Rate : 0.5357
##     P-Value [Acc > NIR] : 0.01688
##
##                   Kappa : 0.4842
##  Mcnemar's Test P-Value : 0.13057
##
##             Sensitivity : 0.9333
##             Specificity : 0.5385
##          Pos Pred Value : 0.7000
##          Neg Pred Value : 0.8750
##              Prevalence : 0.5357
##          Detection Rate : 0.5000
##    Detection Prevalence : 0.7143
##       Balanced Accuracy : 0.7359
##
##        'Positive' Class : norecur
##
```

#가지치기(과적합화의 문제를 해결하기 위해 Pruning단계) #train셋을 여러번 쪼개서 테스트 한 다음 분산이 가장 낮은 가지의 수

#3