**Classification Decision Tree Example Using CART for the Kyphosis Data Set**

- There are several packages which support classification and regression tree analysis.
- This discussion integrates the task output for the classification tree example for the kyphosis data set with the CART package in R.

**Kyphosis Data Frame Description**
The kyphosis data frame has 81 rows and 4 columns representing data on children who have had corrective spinal surgery.

- **Kyphosis**: A factor with two levels (absent/present) indicating if the kyphosis condition was present after the operation
- **Age**: In months
- **Number**: The number of vertebrae involved
- **Start**: The number of the first (topmost) vertebra operated on

#Build Classification Tree With RPART for kyphosis data frame

```r
> library(rpart)

> kyphosis[1:3,]

Kyphosis Age Number Start
1 absent 71 3 5
2 absent 158 3 14
3 present 128 4 5
```

```r
> pairs(kyphosis,panel=panel.smooth)
```
# Fit a Model [Use defaults cp=.01, minsplit=20, minbucket=round(minsplit/3),xval=10]
> fit <- rpart(Kyphosis~Age+Number+Start, method="class", data=kyphosis)
> fit

n= 81

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 81 17 absent (0.7901235 0.2098765)
   2) Start>=8.5 62  6 absent (0.9032258 0.0967742)
      4) Start>=14.5 29  0 absent (1.0000000 0.0000000) *
      5) Start< 14.5 33  6 absent (0.8181818 0.1818182)
         10) Age< 55 12  0 absent (1.0000000 0.0000000) *
         11) Age>=55 21  6 absent (0.7142857 0.2857143)
            22) Age>=111 14  2 absent (0.8571429 0.1428571) *
            23) Age< 111 7  3 present (0.4285714 0.5714286) *
   3) Start< 8.5 19  8 present (0.4210526 0.5789474) *

# Display the Complexity Parameter table for the fitted rpart object
> printcp(fit)

Classification tree:
rpart(formula = Kyphosis ~ Age + Number + Start, data = kyphosis,
       method = "class")

Variables actually used in tree construction:
[1] Age   Start

Root node error: 17/81 = 0.20988

n= 81

<table>
<thead>
<tr>
<th>CP</th>
<th>nsplit</th>
<th>rel error</th>
<th>xerror</th>
<th>xstd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0.176471</td>
<td>0</td>
<td>1.000000</td>
<td>1.0000</td>
<td>0.21559</td>
</tr>
<tr>
<td>2 0.019608</td>
<td>1</td>
<td>0.82353</td>
<td>1.0588</td>
<td>0.22010</td>
</tr>
<tr>
<td>3 0.010000</td>
<td>4</td>
<td>0.76471</td>
<td>1.1176</td>
<td>0.22433</td>
</tr>
</tbody>
</table>

# Automatically select the complexity parameter associated with the smallest cross-validation error. [HSAUR]
> fit$cptable[which.min(fit$cptable,"xerror"),"CP"]
[1] 0.1764706
# Visualize the Cross-Validation Results

> plotcp(fit)

![Cross-Validation Results Plot](image)

# Plot Original Tree to a postscript file

> post(fit,file="tree.ps",title="Classification Tree for Kyphosis")

# Plot Original Tree

> plot(fit,uniform=TRUE,compress=TRUE,lty=3,branch=.7,main="Classification Tree for Kyphosis")
> text(fit,use.n=TRUE,all=TRUE,cex=.8,xpd=TRUE)

![Original Tree](image)
# Plot Pruned Tree

```r
> pfit <- prune(fit, cp=fit$cptable[which.min(fit$cptable[,"xerror"]),"CP")
> plot(pfit,uniform=TRUE,compress=TRUE,lty=3,branch=.7,main="Pruned Classification Tree for Kyphosis")
> text(pfit,use.n=TRUE,all=TRUE,cex=.8,xpd=TRUE)
```

![Pruned Classification Tree for Kyphosis](image)

# Compute the Misclassification Error for Pruned Tree

```r
> (tt <- table(actual=kyphosis$Kyphosis,predicted=predict(pfit,type="class")))
predicted
          actual     absent present
absent     56       8
present     6      11
```

# The Pruned Tree has a 17% error rate.

```r
> 1-sum(diag(tt))/sum(tt)
[1] 0.1728395
```

# Create a Conditional Inference Tree

```r
> library(party)
> fit <- ctree(Kyphosis~Age+Number+Start,data=kyphosis)
> plot(fit,main="Conditional Inference Tree for Kyphosis")
```
# Perform Random Forest Prediction of Kyphosis Data

```r
> library(randomForest)

   randomForest 4.6-2
     Type rfNews() to see new features/changes/bug fixes.

> fit <- randomForest(Kyphosis~Age+Number+Start, data=kyphosis)
> print(fit)

Call:
  randomForest(formula = Kyphosis ~ Age + Number + Start, data = kyphosis)
Type of random forest: classification
  Number of trees: 500
No. of variables tried at each split: 1

  OOB estimate of error rate: 18.52%
Confusion matrix:

             absent present class.error
absent        60     4       0.0625000
present       11     6       0.6470588

> importance(fit)

          MeanDecreaseGini
Age         8.557202
Number      5.435533
Start      9.975925
```

F. References
Decision Tree Topic Notes
