第3章　人工知能エンジンと決定木

3.4 偽札データ再考
3.5 不動産の鑑定
3.6 Rによる決定木

07t4072f　岡崎　駿

3.4　偽札データ再考(1)

- タイタニックのデータでは予測変数も基準変数も質的変数であった。
  - 質的変数
    - 2値変数（ダミー変数）
    - 多数変数（カテゴリカル変数）

- 偽札データでは、予測変数が「対角線」、「下部マージン」の連続変数である。
  - 予測変数が連続変数になった場合はどうなるのだろうか？

3.4　偽札データ再考(2)

- 「真札」と「偽札」を見分ける基準を探すためにCARTで分析

対角線＜140.45cm(103枚中99%)→偽札

対角線≧ 140.45cm

　　下部マージン≧9.45cm(16枚中81%)→偽札

　　下部マージン＜9.45cm(96枚中100%)→真札

もし対角線＜140.45cm　ならば偽札

もし対角線≧140.45cm　かつ下部マージン≧ 9.45cmなら偽札

もし対角線≧140.45cm　かつ下部マージン＜9.45cmなら真札

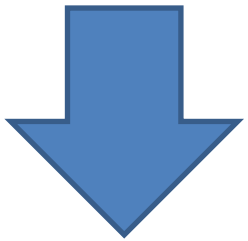
3.4　偽札データ再考(3)

- 予測変数が連続変数の場合
  - 当該の親ノード内のN個のオブザベーションをその変数に対してソートする。
  - 重複しない測定値が何種類あるか数え、それをM個とする。(連続変数による測定値の数値は大方、M=N)
  - M-1個の分岐基準を計算して、最大になる点が、分岐点とする。

  ⇒「1つの連続変数の代わりに、あたかもM-1個のダミー変数があるかのように扱う」

3.5　不動産の鑑定(1)

- 分類木･･･基準変数が質的変数である木
- 回帰木･･･基準変数が連続変数である木



　分類木でも回帰木でも予測変数の扱いは同等である。

　質的変数はその水準によって枝の分岐を行わない

3.5　不動産の鑑定(2)

変数名	性質	内容
予測変数		
犯罪率	連続	犯罪発生率
邸宅率	連続	2万5千sq.ft以上の広さの住宅百分率
非小売	連続	非小売業者の百分率
川沿い	2値	チャールズ川沿いは1それ以外は0
NX濃度	連続	窒素酸化物濃度
部屋数	連続	平均部屋数
古さ率	連続	1940年以前建築の建物の百分率
安定所	連続	5つの職業センターまでの重み付き距離
高速道	多値	高速道路の接続・利用のしやすさ
資産税	連続	1万ドルに対する資産税
生徒数	連続	先生1人あたりの生徒の人数
黒人率	連続	1000×（黒人比率－0.63）^2
下層率	連続	下流階層の百分率
基準変数		
家価格	連続	住宅価格の中央値(単位1000ドル)

3.5　不動産の鑑定(3)

部屋数＜6.941、平均：19.934、効果：-2.599
　下層率≧14.400、平均：14.956、効果：-4.978
　　犯罪率≧6.992、平均：11.978、効果：-2.978
　　犯罪率＜6.992、平均：17.138、効果：+2.182
　下層率＜14.400、平均：23.350、効果：+3.416
　　安定所≧1.385、平均：22.905、効果：-0.445
　　部屋数＜6.543、平均：21.630、効果：-1.275
　　部屋数≧6.543、平均：27.427、効果：+4.522
　　安定所＜1.385、平均：45.580、効果：+22.230
部屋数≧6.941、平均：37.238、効果：+14.705
　部屋数＜7.437、平均数：32.113、効果：-5.125
　　犯罪率≧7.393、平均：14.400、効果：-17.713
　　犯罪率＜7.393、平均：33.349、効果：+1.236
　部屋数≧7.437、平均：45.097、効果：-11.797
　　生徒数≧18.300、平均：33.300、効果：-11.797
　　生徒数＜18.300、平均：46.407、効果：+1.310

3.5　不動産の鑑定(4)

- 部屋数の多い少ないに関わらず、「部屋数」と「犯罪率」は分岐基準として、複数回用いられている
- 「部屋数」の少ない地域の値段「安定所」が効いている
- 「部屋数」の多い地域の値段には「生徒数」が効いている

3.5 不動産の鑑定(5)

- 基準変数が連続変数である課題にはジニ係数を用いた分岐基準は使用できない
- 「平方和の分解」を使用
- 親ノードにおける基準変数の偏差平方和を求める

$$SS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

N : 親ノード内のオブザベーションの数

y_i: 基準変数の測定値

$\bar{y}$: 親ノード内の基準の平均値

3.5 不動産の鑑定(5)

予測変数Tのよって2分岐すると、

右にN_R個、左にN_L個のオブザベーションに分かれる。

子ノードの偏差平方和は

$$SS_{TR} = \sum_{i=1}^{N_R} (y_i - \bar{y}_R)^2$$

$$SS_{TL} = \sum_{i=1}^{N_L} (y_i - \bar{y}_L)^2$$

$\bar{y}_R, \bar{y}_L$: 子ノード内の基準の平均値

子ノード内の偏差平方和の和

$$SS_{TW} = SS_{TR} + SS_{TL}$$

3.5 不動産の鑑定(5)

親ノードの平方和と子ノード内の平方和の差

$$SS_{TB} = SS - SS_{TW}$$

候補となる予測変数ごとにSS_{TB}を計算し、その値が
最大になる予測変数で分岐を行ない、回帰木を成長させる。

予測変数に外れ値が多い場合など、平均値の基準で
回帰木を成長させることが望ましくない場合は

$$SS^{*}_{TB} = \sum_{i=1}^{N} | y_i - \tilde{y} | - (\sum_{i=1}^{N_R} | y_i - \tilde{y}_R | + \sum_{i=1}^{N_L} | y_i - \tilde{y}_L |)$$

を用いる

3.5　不動産の鑑定(6)

- 実際に現場で利用する場合には成長しすぎた決定木には成績が悪いことが多い

プルーニング(枝刈り)を行う

- 推定用のデータだけを使う方法
- 交差妥当化用のデータや検証用のデータを併用する方法

3.5　不動産の鑑定(7)

1.推定用のデータを用い、見かけ上の成績が頭打ちになるまで、十分に木を繁らせる

2.ターミナルノードを含む枝の中で推定用のデータに関して成績の良くない部分に着目し、その枝があった場合とない場合の両方の成績を交差確認用のデータで計算

3.交差確認用データに関して成績の落ちる枝はプルーニングする

4.プルーニングした枝の部分はターミナルノードになるので、2.3の過程を繰り返す

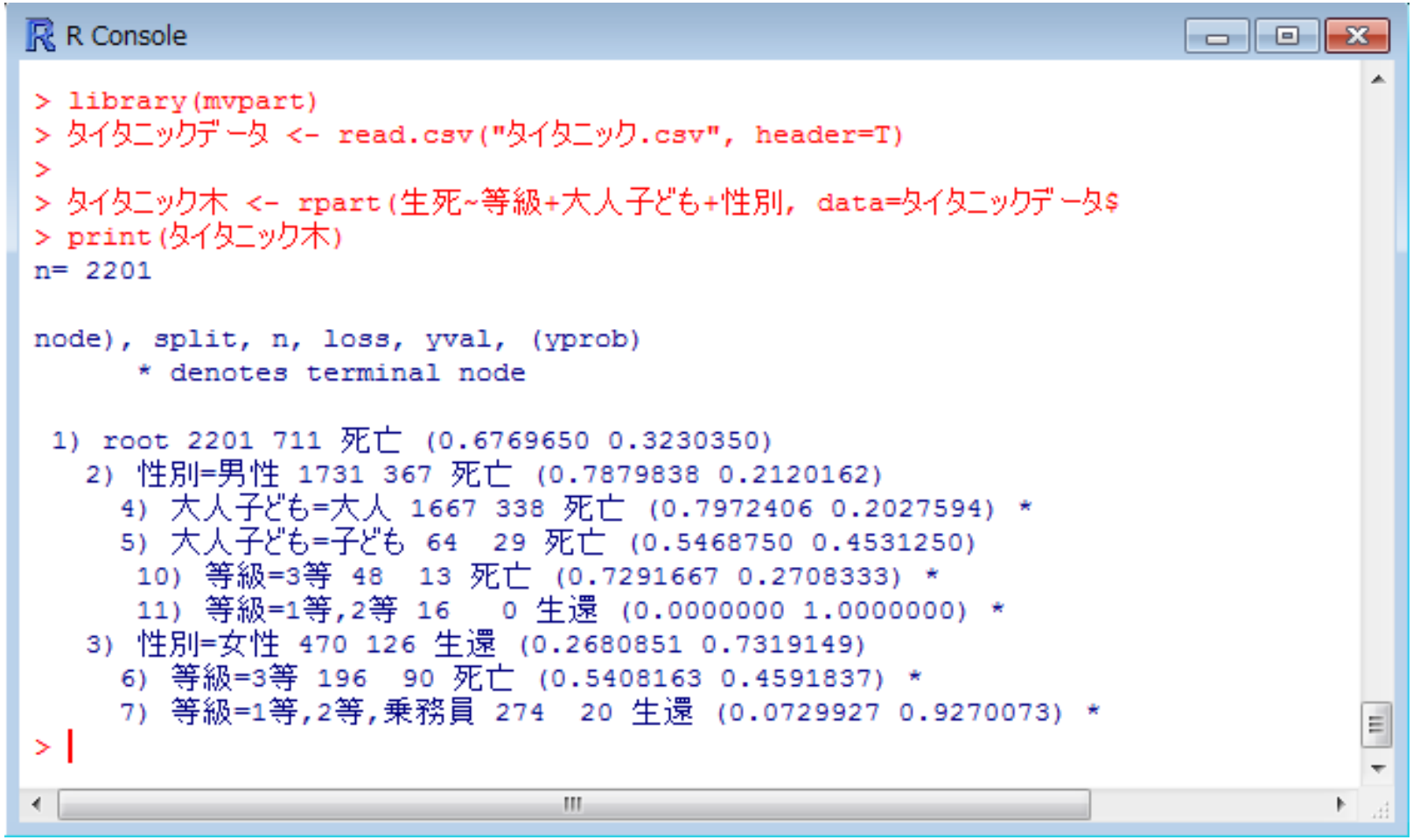
5.どのターミナルノードを含む枝をプルーニングしても、交差確認用データによる成績が下がるようであればプルーニング終了

6.最終的な決定木と交差確認用データは独立でなくなっているため、3番目のデータを検証用に使用

7.検証用データを用いて、計算した最終的な決定木の成績を選出

Rによる決定木(1)

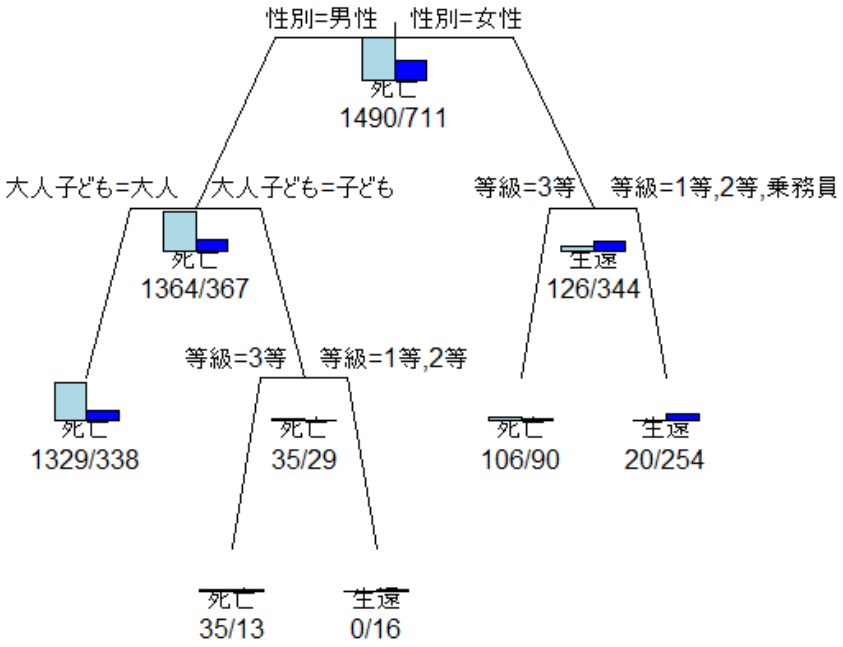
- タイタニックデータの分析



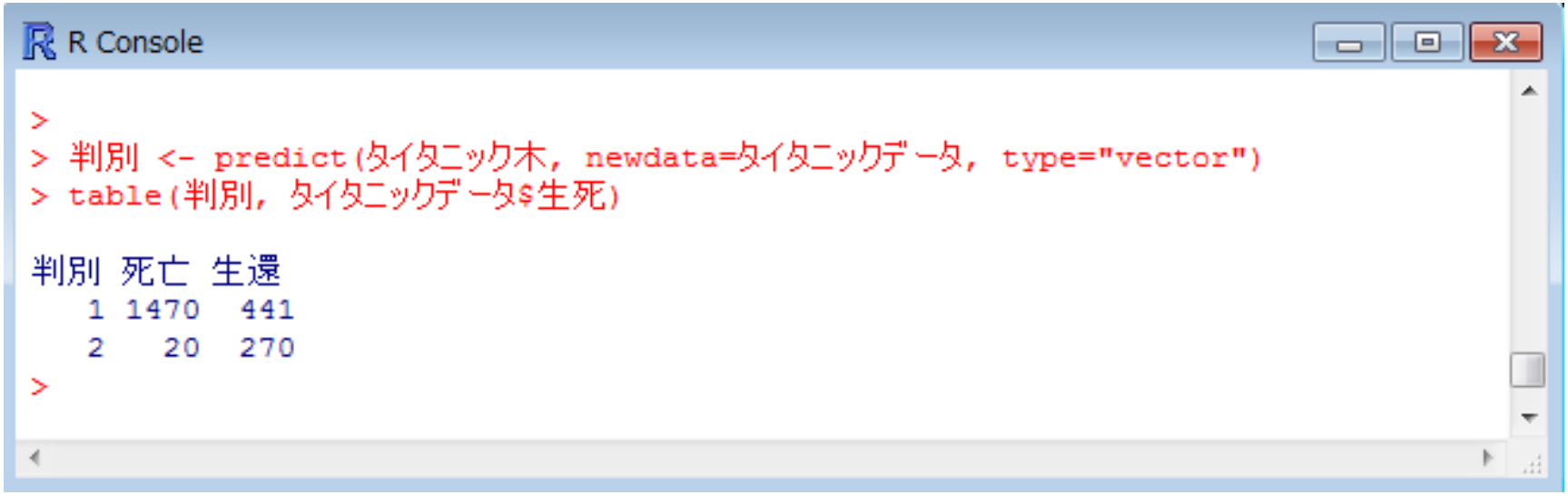
Rによる決定木(2)

R Console

```
> plot(タイタニック木, uniform=T, branch=0.6, margin=0.15)
> text(タイタニック木, all=T, use.n=T, pretty=0)
>
```

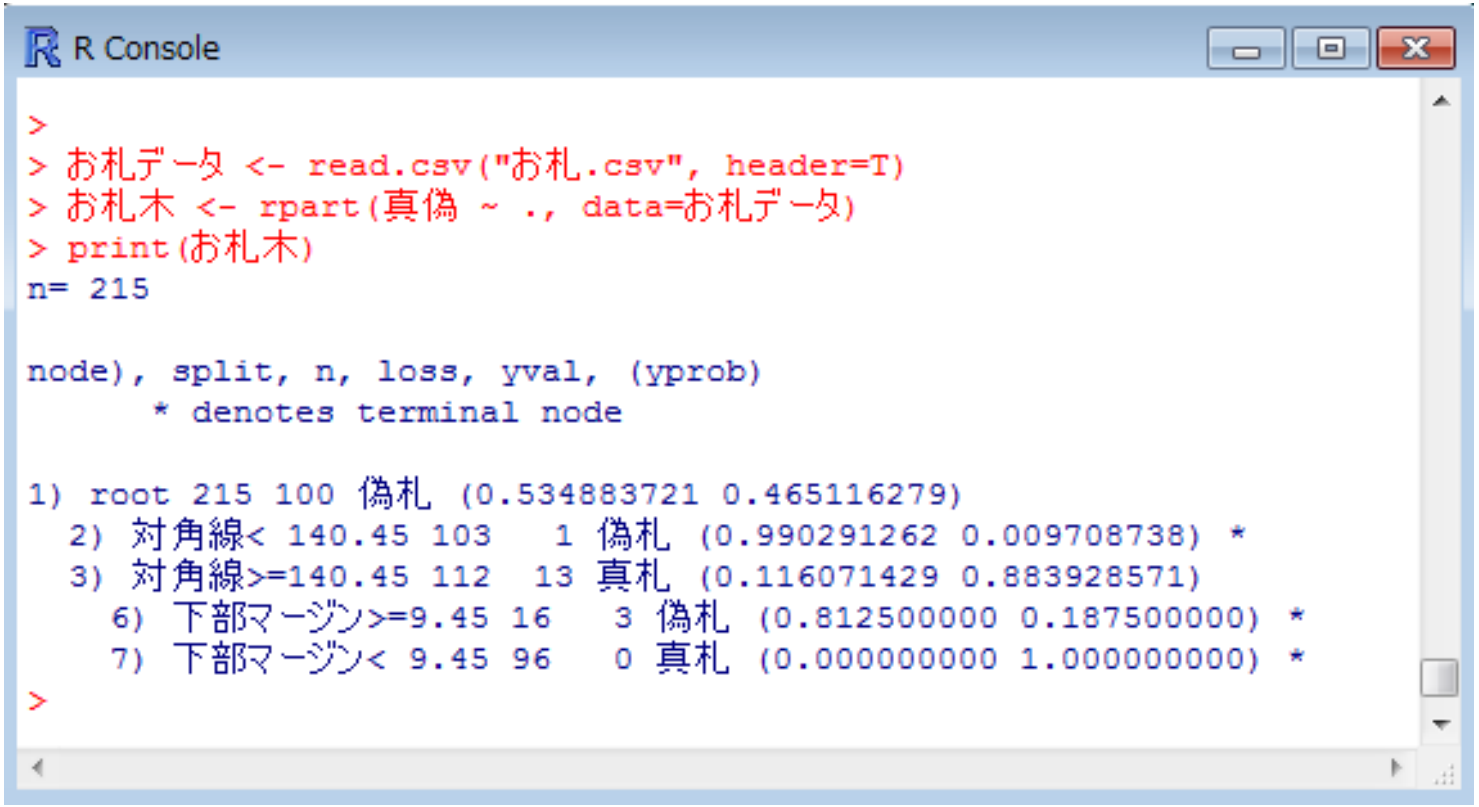


Rによる決定木(3)

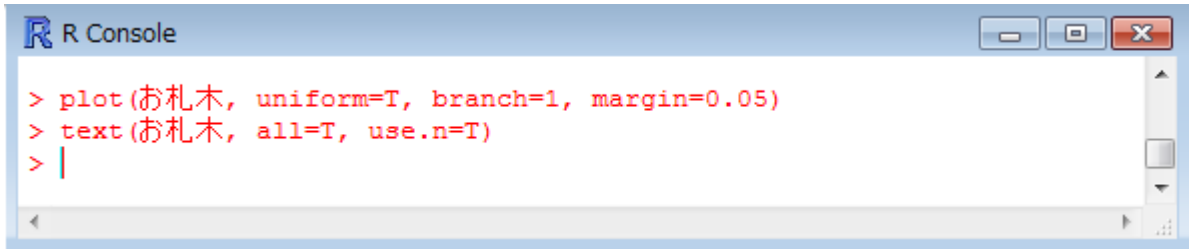

R Console
>
> 判別 <- predict(タイタニック木, newdata=タイタニックデータ, type="vector")
> table(判別, タイタニックデータ$生死)
判別 死亡 生還
1 1470 441
2 20 270
>

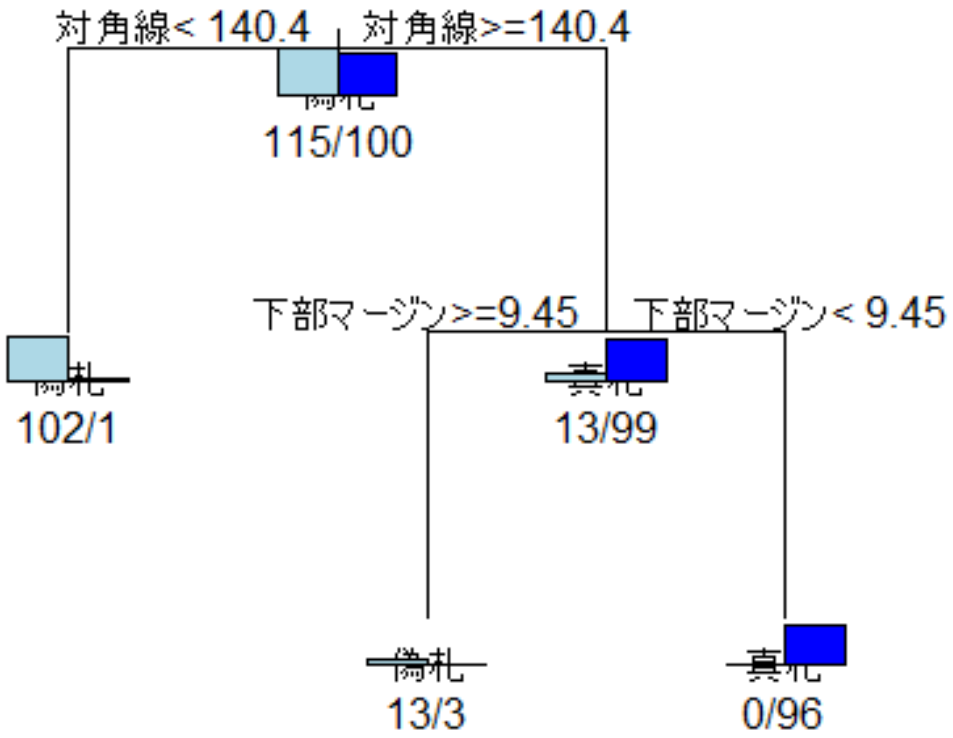
Rによる決定木(4)

- 偽札データの分析



Rによる決定木(5)


R Console
> plot(お札木, uniform=T, branch=1, margin=0.05)
> text(お札木, all=T, use.n=T)
>


対角線< 140.4
対角線>=140.4
115/100
下部マージン>=9.45
下部マージン< 9.45
102/1
13/99
13/3
0/96

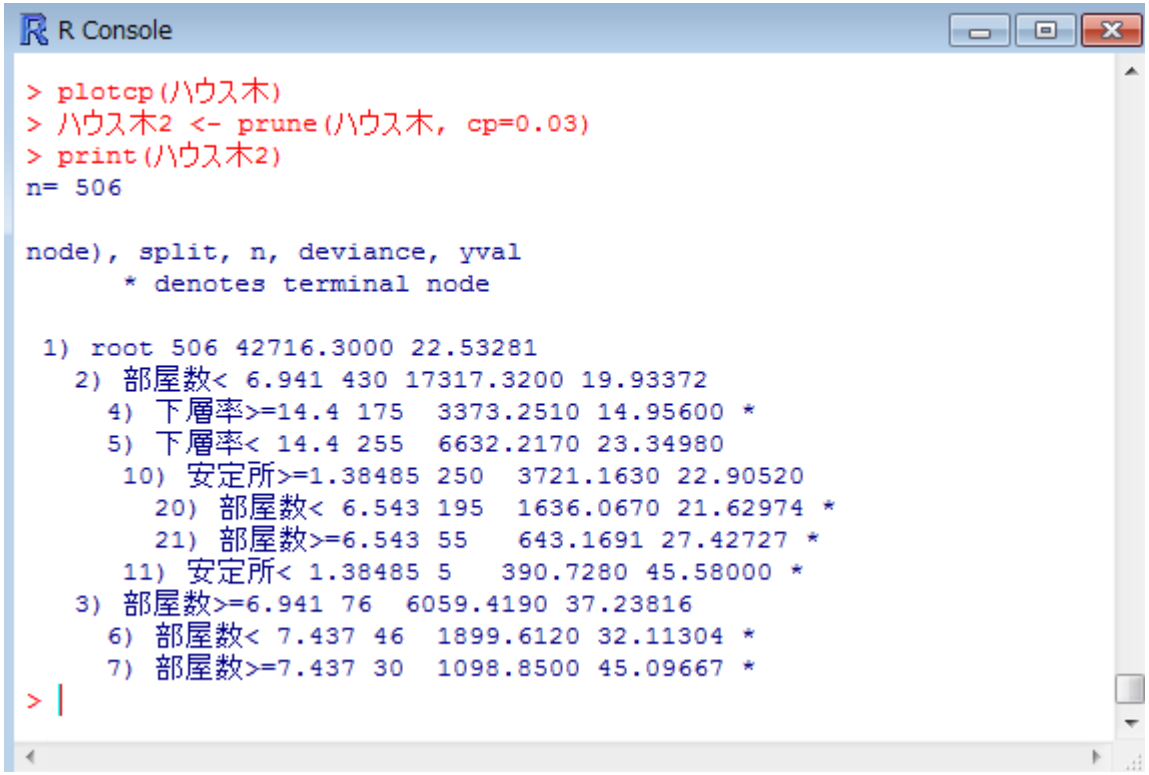
Rによる決定木(6)

- 不動産データの分析

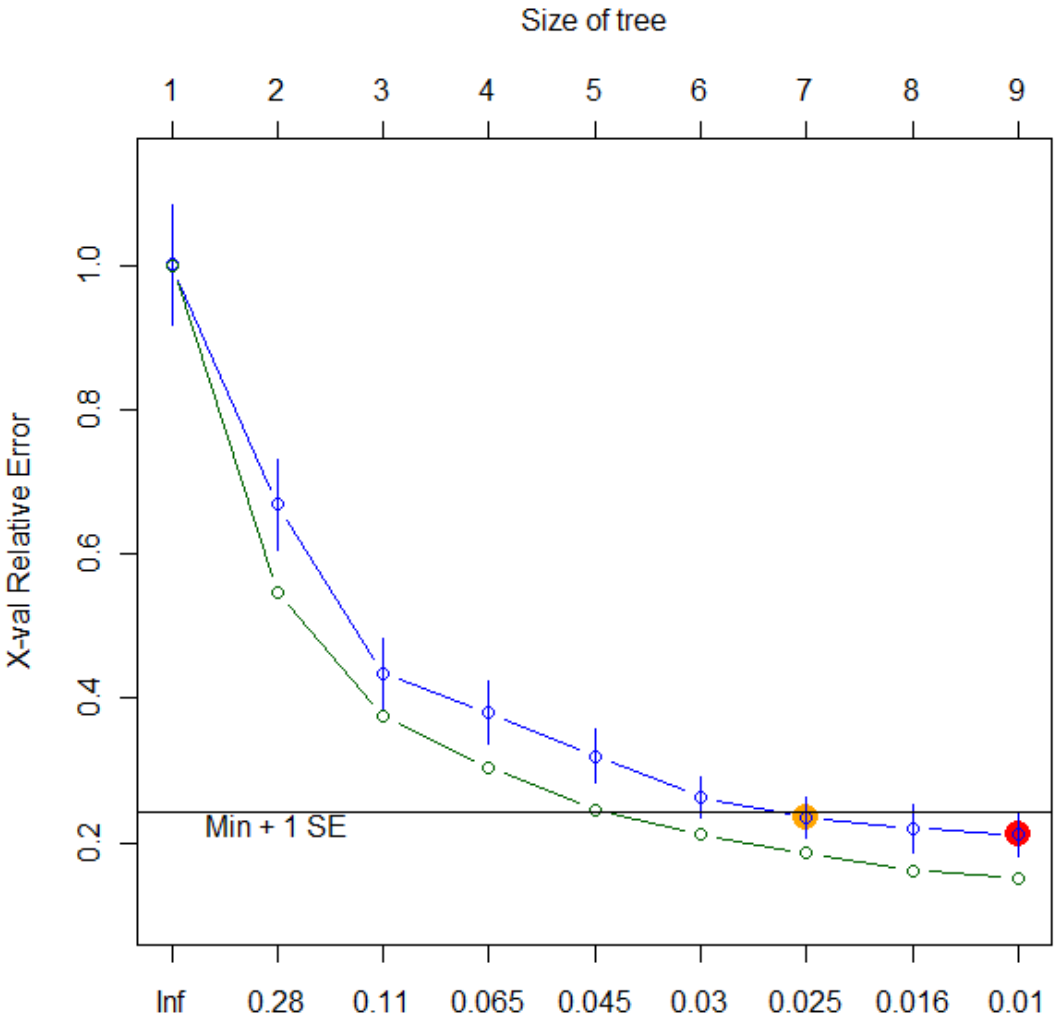
R Console

```
> library(mvpart)
>
> ハウスデータ <- read.csv("ハウス.csv", header=T)
> ハウス木 <- rpart(家価格 ~ ., data=ハウスデータ, method="anova")
> print(ハウス木)
n= 506

node), split, n, deviance, yval
      * denotes terminal node

 1) root 506 42716.3000 22.53281
   2) 部屋数< 6.941 430 17317.3200 19.93372
     4) 下層率>=14.4 175  3373.2510 14.95600
       8) 犯罪率>=6.99237 74  1085.9050 11.97838 *
       9) 犯罪率< 6.99237 101  1150.5370 17.13762 *
     5) 下層率< 14.4 255  6632.2170 23.34980
      10) 安定所>=1.38485 250  3721.1630 22.90520
        20) 部屋数< 6.543 195  1636.0670 21.62974 *
        21) 部屋数>=6.543 55   643.1691 27.42727 *
      11) 安定所< 1.38485 5   390.7280 45.58000 *
   3) 部屋数>=6.941 76  6059.4190 37.23816
     6) 部屋数< 7.437 46  1899.6120 32.11304
      12) 犯罪率>=7.393425 3    27.9200 14.40000 *
      13) 犯罪率< 7.393425 43   864.7674 33.34884 *
     7) 部屋数>=7.437 30  1098.8500 45.09667
      14) 生徒数>=18.3 3   223.8200 33.30000 *
      15) 生徒数< 18.3 27   411.1585 46.40741 *
> |
```

Rによる決定木(7)


R Console
> plotcp(ハウス木)
> ハウス木2 <- prune(ハウス木, cp=0.03)
> print(ハウス木2)
n= 506
node), split, n, deviance, yval
* denotes terminal node
1) root 506 42716.3000 22.53281
2) 部屋数< 6.941 430 17317.3200 19.93372
4) 下層率>=14.4 175 3373.2510 14.95600 *
5) 下層率< 14.4 255 6632.2170 23.34980
10) 安定所>=1.38485 250 3721.1630 22.90520
20) 部屋数< 6.543 195 1636.0670 21.62974 *
21) 部屋数>=6.543 55 643.1691 27.42727 *
11) 安定所< 1.38485 5 390.7280 45.58000 *
3) 部屋数>=6.941 76 6059.4190 37.23816
6) 部屋数< 7.437 46 1899.6120 32.11304 *
7) 部屋数>=7.437 30 1098.8500 45.09667 *
>

Rによる決定木(8)


Size of tree
1
2
3
4
5
6
7
8
9
X-val Relative Error
0.2
0.4
0.6
0.8
1.0
Min + 1 SE
Inf
0.28
0.11
0.065
0.045
0.03
0.025
0.016
0.01

Rによる決定木(9)

R Console

```
> plot(ハウス木2, uniform=T, branch=0.6, margin=0.1)
> text(ハウス木2, all=T, use.n=T)
>
```