先端メディアゼミナールⅡ

厚見　隆之介

第五回:タイタニックの運命

Rによる決定木

決定木とは予測モデルであり、ある事項に対する観察結果から、その事項の目標値に関する結論を導くものである。(wiki引用)

決定木には２つの種類があり

回帰木：実数値を取る関数の近似に用いられる(例　住宅価格の見積もり　患者の入院期間の見積もり)

分類木：分類に用いられる(例　性別　『男女』など)

分類木

まず決定木のパッケージは'tree','rpart','mvpart'の３つがありますが、今回は一番新しい'mvpart'を使います。

＊３つの違いを調べましたが'mvpart'は'rpart'の拡張版って事しかわかりませんでしたorz

まずはパッケージとデータを読み込みます。

まずパスを通します。

library(mvpart)

タイタニックデータ<-read.csv(“タイタニック.csv,header=T”)

エラーが出た人用

Error in library("mvpart") : there is no package called 'mvpart'

とエラーが出た人がいると思います。

そしたら上のメニューバーの

「パッケージ」ー＞

「パッケージのインストール」ー＞

「CRANのミラーサイトと選択」ー＞

でパッケージの「mvpart」を指定しインストールしてください。

決定木を生成します。

決定木を生成します。

タイタニック木 = rpart(生死~等級+大人子ども+性別, data=タイタニックデータ, method="class")

print(タイタニック木)

Node):分岐のノードの番号

split:分岐の条件

n:そのノードに含まれている個数

loss:誤分類の個体数

yval:そのノードの基準変数

yprob:各ノードの適合率

樹木の図を描く

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)

ニューラルネットでの分析

- 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)

回帰木

```
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)
```

プルーニング

逆伝播する誤差や重みの性質を調べ、

最終層への影響の少ないユニットを訓練中に破棄するオプション（67p）

```
plotcp(ハウス木)

ハウス木2 <- prune(ハウス木, cp=0.03)

print(ハウス木2)
```

宿題

- 年収.csvを使って決定木を作成してください。

- 上のデータを使って実測値と予測値の散布図を作成してください。

まとめと感想

- 決定木を使うと分ける事が難しい物に関しても判別し判断することができる。
- 少しの分散がまったく異なった木を生成するから結果が不安定。

- 決定木は判別したい物を木で表し目で理解しやすいのはいいと思った。
- しかし、cpの値によって木が細かく分散し結果が安定しないのはなと思った。