モデルの系譜と発展

2年4組85番山縣謙蔵

決定木の主要モデル

- CHAID
- SIMS
- C&RT
- QUEST
- CART
- QUEST

などが挙げられる

（今回はCARTというモデルを利用）

ジニ係数

- 親ノードAに属するオブザベーションから任意に一つ選んでそれがCjである確率をpAjとする
- Cjの分散はpAj(1-pAj)で表せることが知られている

これらから式を定義する

親ノードAにおける基準変数Cの総分散の数式

$$I(A)=i(pA)=\sum_{j=1}^{j} pAj(1-pAj)=1-\sum_{j=1}^{j} p^{2}Aj$$

親ノードAが子ノードALとARになる場合⊿Iを最大化する分岐基準を選択

$$\Delta I = P(A)I(A) - \{P(AL)I(AL) + P(AR)I(AR)\}$$

子ノードの不純度平均と親ノードの不純度の差を計算

偽札データ再考

表を作る

偽札データ

予測変数が対角線と下部マージン（連続変数）

基準変数が真札と偽札（質的変数）

決定木から抽出されるルールに基づき表を作成（P93　図3.2）

予測変数が連続変数の場合

- 当該親ノード内のN個のオブザベーションをその変数に関して分類する
- 重複のない測定値が何種類あるか数えそれをM個とする。このとき必ずしもM=Nとはならない（p93参考）
- M-1個の分岐点を作成（連続変数では分割点が無限にできるためダミー変数で代用）

演習

お札.csvを使って図式化した決定木を作成してみよう！

解説（1）

```
library(mvpart)
money = read.csv(“お札.csv”,header=T)
moneywood=rpart(真偽~.,data=money)
print(moneywood)
//ここで決定木が表示される
```

解説（２）

plot(moneywood,uniform=T,branch=1,marigin=0.05)

Text(moneywood,all=T,use.n=T)

//決定木が図式化される

まとめと感想

分類パターンが多数ある連続変数を使った決定木の作成も図式化可能であることが分かった。

今回のように大別していく考え方は今後の学習でも重要になっていくだろうと思った

宿題

ハウス.csvのデータを用いて決定木を作成しよう!