不動産の鑑定

濱永　千佳

決定木の種類(p94)

- 分類木
  - 基準変数が質的変数
    （例 「生死」、「真偽」、「男女」など）
- 回帰木
  - 基準変数が連続関数
    （例 「身長」、「犯罪率」など）

- 枝の分岐を行うときの、予測変数の扱いは同じ
  - 質的変数：その水準による
  - 連続変数：ソートし、適当な分岐点を決めてから

分岐点の決定(p96)

基準変数が連続変数

⇒平方和の分解を利用

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

（ジニ係数は、質的変数のどちらであるかを利用して計算していた。連続変数であると、基準の数値から求めなければならないため、平方和の分解を使う。）

プルーニングと交差妥当化

- プルーニング（枝刈り、剪定とも呼ばれる）
  - 繁りすぎた決定木の枝を、（結果が大きく変わらないように検証しながら）刈ること
  - ２種類の方法で実施
    - 推定用のデータだけ
    - 交差妥当化用データ、検証用データを併用する

- 交差妥当化(p１６)
  - 心理測定学における妥当性研究の分野で生まれた考え方

機械学習

- 機械学習のアルゴリズム
  - 最適な決定木に到達する保証がない
  - 分岐回数を定めて簡単なプログラムを書いても、組合せが多くなると計算しきれない

    ⇒計算の困難さを「組合せ爆発」と呼んでいる

演習

回帰木を作成してみよう

演習1

- Shareにある、「ハウス.csv」を使って、回帰木を作成してみよう。
  - 回帰木を指定する引数はmethod = ”anova”

解答1 NO.1

```
library(mvpart)
house = read.csv(“ハウス.csv")
header = TRUE
attach(house)

housetree = rpart(家価格~., data=house, method="anova“, cp=0.01)
print(housetree)
```

ここで、method=“anova”で回帰木であることを指定する

演習1 NO.2

- cpとは？
  - 複雑さを表すパラメータ(complexity parameter)
    値が小さいほど、木は複雑となる
    デフォルトでは、cp=0.01と設定されている

- プルーニングしてみよう

plotcp(housetree) ⇐プルーニングの目安を探る

※ランダムにデータを分割して交差確認を行った結果のため、いつも同じ値ではないことに注意

housetree2 = prune(housetree, cp=0.03)

prune関数を使い、cp値を指定して返す

print(housetree2)

演習1 NO.3

- 決定木を表示する

plot(housetree2, uniform=T, branch=0.6, margin=0.1)

text(housetree2, uniform=T, all=T, use.n=T)

~~時間の余った人は、~~プルーニングする前の決定木と比較してみよう

plot(housetree, uniform=T, branch=0.6, margin=0.1)

text(housetree, uniform=T, all=T, use.n=T)

感想

- プルーニングは何回か前の時間で学習していたが、データを比較してみると枝の数が減り、決定木の全体を見やすくなったと感じた。
- 途中で触れた計算は、組み合わせが多くなると計算量も増加していくものであり、最適な決定木に到達する保証がないということはよくわかった。

宿題

- 組み込みデータ「iris」を用いて、回帰木を作成してみよう。
  - 基準変数は「Species」
  - プルーニングを行ってください。