R

決定木分析

概要

・決定木
  ・ 分類木：目的変数がカテゴリー
  ・ 回帰木：目的変数が数値（連続）

■分析のアルゴリズム

・ C4.5：カテゴリー
・ CART
・ CHAID

基本

■モデルの作成

```
モデル <- rpart( 目的変数 ~ 説明変数 + 説明変数 , data= データ )
```

・ method="class" とオプションの指定で、目的変数をカテゴリーとして扱う
  ・ 指定がなければ、自動判別
・ method="anova" で、数値として扱う

■可視化：グラフにする

```
plot(as.party( モデル ))
```

■分析結果の三つの見方

1.as.party(モデル)
2. モデルそのものを見る
3.summary(モデル)

Complex Parameter (cp)

・複雑性パラメタ
・木が複雑になりすぎた場合（分岐が多すぎる場合)、木が伸びすぎないように「剪定」する

■ printcp(**モデル**)

・エラーの収束を見て、cp をどこまでにするかを判断する

■ plotcp(**モデル**)

標準を下回るところを見つける。カットオフポイント。

サンプルデータとスクリプト

サンプルデータ（各指標）から、
１）欠損値を含むデータ 2 つを除き、
２）ファイル名のカラムを削除したものを、
jpn.4c というデータフレームに入れているものとします。

■ rpart で Decision Tree Analysis

パッケージ rpart をインストール。

Score を目的変数（分類カテゴリー）として、残りの言語指標（説明変数）のどれがどのくらい分類に寄与するかを観察します。

式は：　rpart(目的変数 ~ ., data = データフレーム)
結果を、モデルとして保存するようにします。

具体的には、まず、attach して使うデータフレーム名を決めておく。

■ 結果のグラフ

```
install.packages("rpart.plot", dependencies = T)
library(rpart.plot)
rpart.plot(jpn.DTmodel)
```

・上から順に、重要な指標と位置づけられる。

1. 一番重要な指標が Type　131 以上あるか、
2. 次に Token が 398 以上あるか
3. そのうえ、ASL（平均文長）が 13 以上あれば、評価は 4.9 になる。

・一番下のレベルで、AWL（平均語長）が基準として選ばれている。
・まったく図に表れない指標（各種語彙多様性指標や文の数）はスコアの決定には寄与しないといえる。

・どんな指標を入れるかにより、判断は変わってくるが、逆に、どんな指標を入れようが、重要でない指標は選ばれないし、それほど重要でない指標は下のほうに位置づけられる。常に上のほうに位置づけられる指標は、常に重要であると判断される。

・どんな指標の組み合わせで分析を行うかは、分析の目的しだい。

スコアをカテゴリー変数とみなして分析しなおしてみる

■ summary(モデル) で、詳しい結果が表示される

partykit を使った別のグラフ表示

```
library(rpart)
library(partykit)
plot(as.party(jpn.DTmodel2))
```

■結果の解釈（Criterion で高得点を取るには）

・総語数が１６０語くらい書けないと評価３はもらえない。
・総語数が２６０語くらい書けると評価４がもらえる可能性が高くなる。
・総語数が４００語以下の場合、語彙力が高ければ４がもらえる可能性が高くなる。
　・多様な語彙（語彙タイプ数が１３０以上）もしくは
　・単語長の長い（平均単語長５文字以上）「難しい」語彙の使用
・総語数が４００語以上で、平均的に長い文（１３単語以上）を書くと評価５がもらえる可能性が高い。

■オプション

・ Node の出力をシンプルにする
　・ plot で、オプションをつける

```
, type="simple"
```

complex parameter (cp) によって剪定を行うことができる

https://toukeier.hatenablog.com/entry/2018/09/03/080713

■ printcp(モデル)

cp
　複雑さ
nsplit
　分岐の数
rel error
　エラー率
xerror
　交差確認のエラー率
xstd
　交差確認のエラー率の標準偏差

■ plotcp(モデル)

rpart.control で詳細設定

minsplit
　分岐に必要な最低データポイント数
minbucket
　端末の「葉」に必要な最低データポイント数
cp
　「complexity parameter」
maxcompete
　価値のない分岐を「剪定」する値。（価値がない＝説明率が上がらない）
maxsurrogate

対立項「competitor 」の数

usesurrogate

xval

交差検証の数

surrogatestyle

maxdepth

最大の深さ、30 まで

分岐

・デフォルトは gini 係数

・information を指定することもできる

・gini 係数が小さくなるように分岐する

分類精度を確認する confusionMatrix()

References

https://qiita.com/LTtoose/items/8d3ac23791ee6e1cb4ce