第八回ベイズ統計・機械学習研究会

Kyohei ITO

2020/11/1

教師あり学習・回帰分析

機械学習は「教師あり学習」と「教師なし学習」に大別されます。

「教師あり学習」は入力値xと出力値yが対となって観測されるとき、そのxとyの間の関係を導こうとする機械学習手法です。

「教師なし学習」は入力値xのみが観測されるとき、そのxの背後にある何らかの傾向を抽出しようとする機械学習手法です。

米国ボストン市郊外における地域別の住宅価格のデータセットである BostonHousing を扱います。
BostonHousing データセットには以下のカラムがあります。

- crim: 人口1人当たりの犯罪発生率
- zn: 25,000 平方フィート以上の住居区画の占める割合
- indus: 小売業以外のビジネスが占める面積の割合
- chas: チャールズ川の周辺かを表すダミー変数 (1:yes 0: no)
- nox: 窒素酸化物濃度 (0.1ppm単位)
- rm: 住居あたりの平均部屋数
- age: 1940年以前から所有されている物件の割合
- dis: ボストン市内5箇所の雇用施設からの重み付けされた距離
- rad: 高速道路へのアクセス性を表す指標
- tax: $10,000あたりの不動産税率
- ptratio: 小学校教員一人あたりに対する児童の人数
- b: Bを街ごとの黒人の比率として 1000(B-0.63)^2
- lstat: 階級が低い層の人口に占める割合
- medv: 持家価格の中央値 ($1,000単位)

今回使うデータの読み込み

```
install.packages(mlbench)
data(BostonHousing, package = "mlbench")
Data <- BostonHousing
```

回帰分析の概要

回帰分析は複数の説明変数を用いて目的変数を表すモデルを作成する分析方法です。

- 何らかの関係が想像される変数間の関係を調べる(相関関係を示す)
- 説明変数から被説明変数の値を予測する（予測）

などの用途で使用されます。

回帰分析の前提条件

- 独立性（説明変数と残差は独立）

- 等分散性（説明変数にかかわらず残差の分散が一定）
- 正規性（残差が正規分布に従う）
- 線形性（説明変数と目的変数の関係は直線で近似できる）
    - 参考　http://www.aoni.waseda.jp/abek/document/regression-3.html (http://www.aoni.waseda.jp/abek/document/regression-3.html)

回帰分析の詳細は

- 豊田（2012）『回帰分析入門　—Rで学ぶ最新データ解析—』
- 管（2016）『例題とExcelで学ぶ多変量解析ー回帰分析・判別分析・コンジョイント分析編ー』
- 【大学数学】最小二乗法(回帰分析)【確率統計】
    - https://www.youtube.com/watch?v=Zz1sgYxrA-k (https://www.youtube.com/watch?v=Zz1sgYxrA-k)
- 【機械学習】線形回帰（前編）| 線形回帰の理論
    - https://www.youtube.com/watch?v=zo8BmIGSO2Y (https://www.youtube.com/watch?v=zo8BmIGSO2Y)
- 【機械学習】線形回帰（後編）| 重回帰と正則化
    - https://www.youtube.com/watch?v=xh1OtbZyxqw (https://www.youtube.com/watch?v=xh1OtbZyxqw)

などを参考になさってください。

回帰分析をする際のRコード

```
lm(formula, data)
```

- formula は回帰式に用いる被説明変数と説明変数などのモデルの形式を指定する部分です。
- data は回帰分析に用いたデータセットの名前です。

ほかにも引数はありますがここでは省きます。

引数 formula と回帰のモデル式の対応表を挙げる．

formula	回帰におけるモデル式
y ~ x	モデル式 $y = a + bx + \varepsilon$（ε は誤差項）について，目的変数 y と説明変数 x をベクトルで指定する．
y ~ x1 + x2	モデル式 $y = a + b_1x_1 + b_2x_2 + \varepsilon$（$\varepsilon$ は誤差項）について，目的変数 y と説明変数 x_1，x_2 をベクトルで指定する．
y ~ x1 * x2	交互作用項を含んだモデル式（**x1:x2** でもよい） $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2 + \varepsilon$（$\varepsilon$ は誤差項）について，目的変数 y と説明変数 x_1，x_2 をベクトルで指定する．
y ~ x1 + x2 + x1*x2	上と同じ交互作用モデル．
y ~ (x1 + x2)^2	上と同じ交互作用モデル．
y ~ x - 1	切片(定数)項を除外する（＋0 でも可）．
y ~ 1 + x + I(x^2)	多項式回帰 ： $y = b_0 + b_1x_1 + b_2x^2 + \varepsilon$（$\varepsilon$ は誤差項）．I(x^2) は poly(x,2) でも可．
y ~ x | z	z で条件付けしたときの y の x への単回帰．
y ~ ., data = データ名	あるデータに目的変数 y と説明変数 x_1，・・・ が入っている場合で，モデル式が$y = a + b_1x_1 + \cdots + \varepsilon$（$\varepsilon$ は誤差項）である場合は，まず，目的変数 y をベクトルで指定し，右辺は「y 以外」という意味で．（ピリオド）を指定することも出来る．

http://cse.naro.affrc.go.jp/takezawa/r-tips/r/71.html (http://cse.naro.affrc.go.jp/takezawa/r-tips/r/71.html)　より

データの分割

```
library(tidymodels)

df_split = initial_split(Data,  p = 0.8) #データをpの割合で分割する
df_train = training(df_split)
df_test  = testing(df_split)
```

散布図、相関係数行列

```
library(GGally)
ggpairs(df_train)
```

回帰分析の要約

```
summary(model)
```

モデルを用いた予測

予測にはpredict関数を使います。データはnewdataに指定します。

```
predict(model,newdata = DATA)
```

逐次選択法（ステップワイズ法

） 変数の数を変えながら最適なモデルを探す。

```
step(lm(y~.,data = ))
```

VIF統計量を算出

carパッケージをインストールして使用します。

回帰分析のモデルを引数にとります。

```
vif(model)
```

散布図、相関係数行列

```
library(GGally)
ggpairs(df_train)
```

正規性の確認

残差の正規性の確認のためにQQplotを行います。

```
qqnorm(x)
```

モデルの評価（AIC）

```
AIC(lm)
```

モデルの評価（MSE,RMSE）

評価指標とは、学習させたモデルの性能やその予測値の良し悪しを測る指標 です。 https://aizine.ai/rmse-rmsle1114/ (https://aizine.ai/rmse-rmsle1114/)

```
library(tidymodels)
metrics(data,y1,y2)
```