情報量統計学的データ可視化ツール

スクリプト説明書

2016/04/28

（株）NTT データ数理システム 作成[1]

[1] 本ソフト及び付随するドキュメントは、大学共同利用機関法人情報・システム研究機構統計数理研究所が作成した仕様に基づいて、株式会社 NTT データ数理システムが作成したものである。

1. 概要

本ソフトは、スプレッドシート形式データ（カテゴリ変数・数値変数が混在しても構わない）が与えられた場合に、データの関連度及び構造を可視化する機能を提供する R スクリプトから構成されている。

他変数間の関係を見る道具として一般に広く使われているものとして相関係数行列がある。しかし、それには次の２つの大きな欠点がある。

（１） 相関係数行列はカテゴリ変数が扱えない

（２） 相関係数行列は非線形関数関係が扱えない

本ソフトは、上記相関係数行列が持つ欠点を、変数間に成り立つモデルの赤池情報量基準（ AIC ）を考えることにより解決した汎用ソフトである。

本ソフトは入力引数として、データファイル名、解析設定ファイル名、及び作業ディレクトリパスを指定する。実行結果として、作業ディレクトリ以下に２つの CSV ファイル（「変数間関連度ファイル.csv」 ファイルと「変数間関連度傾向ファイル.csv」 ファイル）及び１つの PNG 画像ファイル（「変数間関連度プロット.png」）を出力する。

2. 使用方法

visualize.infoStat(inputFILE,colTypesFILE,,workDIR="./",distribution=T,histogram=T)

3. 引数

- inputFILE

  可視化対象の入力データファイル名（CSV ファイル名）を指定する。このファイルは、workDIR で指定しているディレクトリの下に配置しなければならない。データのフォーマットは 1 行目に列名を指定し、列方向にデータの値が定義されているものとする。また、各列のデータ数（行数）は同じであるとする。

- colTypesFILE

  inputFILE で指定した入力データファイルの解析設定ファイル名（CSV ファイル名）を指定する。このファイルは、workDIR で指定しているディレクトリの下に配置しなければならない。
  データのフォーマットは、
  - 第一列に「列名」列
  - 第二列に「属性」列
  - 第三列に「離散化数」列
  - 第四列に「描画対象」列

  の 4 列からなるものとする。
  列名と列順はこの通りでなければならない。

各列の設定値の方法と様式を説明する。

1. 「列名」列
   解析対象の列を指定する。
   データファイルにある列名で、この箇所に名前がないものは、解析及び描画対象から除外される。しかし、この箇所に名前があるが、データファイルには列名がない場合はエラーとなり解析が停止する。
2. 「属性」列
   1 の列名を取り込む際のデータ型を指定する。「numeric」（実数データ）か「factor」(カテゴリデータ)のどちらかを指定する。
3. 「離散化数」列
   整数値を指定する。
   2 の属性で「numeric」を指定した列に対し、最小分割数を指定する。なお、2 の属性で「factor」を指定したものに対して設定値は無視される。
4. 「描画対象」列
   「T」か「F」の文字列を指定する。
   T に対応する列のみ解析・描画対象となる。F に設定された列は解析・描画対象とならない。また、「T」と「F」の値以外の文字列を指定した場合はエラーとなる。

- workDIR

  オプション指定。解析における作業ディレクトリを指定する。デフォルト値は、カレントディレクトリとなる。なお、inputFILE、colTypesFILE で指定したファイルがこのディレクトリの直下に配置されている必要がある。

- distribution

  オプション指定。T とした場合、説明変数の値と目的変数の分布の関連度を調べる。説明変数数値変数はすべてヒストグラムの AIC が最小になる分割数で離散化した後に、相関の計算を実施する。デフォルトで T であり、数値変数を離散化して離散変数として取り扱う。

- histogram

  オプション指定。T とした場合、対象となる目的変数の分布を混合ヒストグラムとして出力する。F とするとヒストグラムの

各区分におけるデータの混合割合を示す帯グラフとして出力する。T がデフォルト.。

4. 結果

関数の出力値はない。次の 4 章で記述するファイルを workDIR で指定したフォルダに出力する。

5. 動作

本関数の出力結果として以下の３つのファイル workDIR　引数で指定したフォルダに出力される。

- 変数間関連度ファイル(CSV ファイル)
  行方向に目的変数名、列方向に説明変数名、対応するセルに変数間関連度（モデルの AIC 差÷データ数）を入れたクロス表である。この変数間関連度の計算は次の 5 章の「動作詳細」を参照のこと。
- 変数間関連度傾向ファイル（CSV ファイル）
  行方向に目的変数名、列方向に説明変数名、対応するセルに変数間関連度から CATDAP の出力に応じた記号を入れたクロス表である。なお、各記号の対応は次の通りとする。
  ・-0.1 未満：*　（半角アスタリスク）
  ・-0.1～-0.05：#　（半角シャープ）
  ・-0.05～-.0.01：+　（半角プラス）
  ・-0.01～0：-　　（半角ハイフン）

- 変数間関連度プロット結果（png 画像ファイル）
  説明変数と目的変数の間の関係のマトリックス表示である。対角成分は、説明変数の分布の表示である。実数値のデータの場合ヒストグラムを表示し、カテゴリデータの場合、集計結果の棒グラフを表示する。非対角成分は distribution = F の場脚は、説明変数、目的変数の散布図を表示し、distribution = T の場合は、目的変数の分布の説明変数の値への依存を可視化したヒストグラム(histogram=T の場合)あるいは帯グラフ(hisutogram=F の場合)を表示する。表示されるのはカラム指定ファイルで描画対象とした変数である。

6. 変数間の関連度の算出方法

本章では、本スクリプトにおける AIC の定義式及び変数間関連度の計算方法について詳説を行う。

まず、出力ファイルである変数間関連度ファイルの目的変数 A と説明変数 B の間の変数間関連度は、次の式で与えられる。

$$\text{変数間関連度} = \frac{AIC_2 - AIC_1}{\text{データ数}}$$

ここで、AIC_2 は「目的変数に A をとり、目的変数に B をとった回帰モデルを考えた際の赤池情報量基準の値」である。 AIC_1 は「目的変数に A をとり、切片項だけの回帰モデルを考えた際の赤池情報量基準の値」である。この AIC_2 とAIC_1の定義式と求め方は目的変数と説明変数の型に応じて後に記述する。なお、データ数はデータの行数である。

6.1. 目的変数が実数・説明変数が実数

$(y_i)_{i=1}^N$を目的変数列の値、$(x_i)_{i=1}^N$を説明変数列の値とする。

この時、固定した n に対して以下の線型回帰を求める。

$$y_i = a_0 + a_1 \cdot x_i + a_2 \cdot x_i^2 + \cdots + a_n \cdot x_i^n + \varepsilon_i$$
$$\varepsilon_i \sim i.i.d.\ N(0, \sigma^2)$$

AIC_2^n は次式により求められる。

$$AIC_2^n = N \cdot \log\left(\sum_{i=1}^N \frac{(y_i - \hat{y}_i^n)^2}{N}\right) + 2(n+1)$$

ここで、

$$\hat{y}_i^{\mathrm{n}} = a_0 + a_1 \cdot x_i + a_2 \cdot x_i^2 + \cdots + a_n \cdot x_i^n$$

とする。

本スクリプトでは、

$$AIC_2 = \min_{n \le 5} AIC_2^n$$

としている。

また、

$$AIC_1 = N \cdot \log\left(\sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{N}\right) + 2$$

ここで、

$$\hat{y}_i = a_0$$

とする。これは、すなわち

$$y_i = a_0 + \varepsilon_i$$

とした場合の回帰モデルでの AIC である。

6.2. 目的変数がカテゴリ・説明変数がカテゴリ

目的変数のカテゴリ値の集合を $\{a_i \mid i = 1,2,\dots,I\}$ とし、説明変数のカテゴリ値の集合を $\{b_j \mid j = 1,2,\dots,J\}$ とする。また、全データのうち、目的変数値が a_i でかつ説明変数値が b_i となるものの数を $f(a_i, b_i)$ と表すこととする。このとき、

$$\mathrm{p}(a_i \mid b_j) \stackrel{\text{def}}{=} \frac{f(a_i, b_j)}{\sum_i f(a_i, b_j)}$$

とする。このとき、

$$AIC_2 = -2\sum_{i,j} f(a_i, b_i) \cdot \log p(a_i \mid b_j) + 2(I-1) \cdot J$$

となる。また、

$$p(a_i) \stackrel{\text{def}}{=} \frac{\sum_j f(a_i, b_j)}{\sum_{i,j} f(a_i, b_j)}$$

として、

$$AIC_1 = -2\sum_i \sum_j f(a_i, b_j) \cdot \log p(a_i) + 2(I-1)$$

とする。

6.3. 目的変数が実数・説明変数がカテゴリ

目的変数のカテゴリ値の集合を $\{b_i \mid i = 1,2,\dots,J\}$ とする。この時、以下の回帰モデルを考える。

$$y_i = \mu_{b_i} + \varepsilon_i$$
$$\varepsilon_i \sim i.i.d.\ N(0, \sigma^2)$$

この時、AIC_2 は

$$AIC_2 = N \cdot \log\left(\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}\right) + 2J$$

$$\hat{y}_i = \mu_{b_i}$$

となる。また、AIC_1 は

$$y_i = \mu + \varepsilon_i$$
$$\varepsilon_i \sim i.i.d.\ N(0, \sigma^2)$$

なるモデルに対し、

$$AIC_1 = N \cdot \log\left(\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}\right) + 2$$

$$\hat{y}_i = \mu$$

とした場合の値である。

6.4. 目的変数がカテゴリ・説明変数が実数

この場合には、次の方法でまず説明変数の離散化を行う。初めに、説明変数の数値が

同じ値を一つの水準と考えて（順序のある）カテゴリ変数とみなす。ここで、次の記号を用いる：

- 説明変数の水準の数を I 個として、これらを $X_1, X_2, ..., X_I$ とする。
  （添え字の順に順序のある順序尺度であるとする）
- 目的変数の水準の数を J 個として、これらを $Y_1, Y_2, ..., Y_J$ とする。

データの観測数（行数）の記号として $n(\cdot)$ を用いる：

- 説明変数が水準 X_i をとる行数： $n(X_i)$
- 目的変数が水準 Y_j をとる行数： $n(Y_j)$
- 説明変数が X_i でかつ目的変数が Y_j をとる行数： $n(X_i, Y_j)$

x, y がカテゴリである場合の確率モデル：

$$P(y \mid x) = \frac{n(x, y)}{n(x)}$$

に対し、説明変数の隣接する水準 X_i, X_{i+1} を合体して 1 つの水準 X_i' とすることを考える。ここで、この結合操作について、結合前と結合後の AIC を比較し、AIC の減少幅を計算する。その減少幅は、水準 X_i, X_{i+1} に関する対数尤度和と自由度の差のみが寄与し、それ以外の水準は計算に寄与しない。

・結合前の AIC（水準 X_i, X_{i+1} に関する部分のみ）

$$-2\sum_j \left[n\left(X_i, Y_j\right) \log \frac{n\left(X_i, Y_j\right)}{n\left(X_i\right)} + n\left(X_{i+1}, Y_j\right) \log \frac{n\left(X_{i+1}, Y_j\right)}{n\left(X_{i+1}\right)} \right] + 2(J-1)$$

・結合前の AIC（水準 X_i, X_{i+1} に関する部分のみ）

$$-2\sum_j \left[n\left(X_i', Y_j\right) \log \frac{n\left(X_i', Y_j\right)}{n\left(X_i'\right)} \right]$$

となるので、この結合に伴う確率モデルの AIC の減少は、

$$-2\sum_j \left[n_{ij} \log \frac{n_{ij}}{n_i} + n_{i+1,j} \log \frac{n_{i+1,j}}{n_{i+1}} - n_{ij}' \log \frac{n_{ij}'}{n_i'} \right] + 2\left(J-1\right)$$

となる。ここで、

$$n_{ij} = n(X_i, Y_j)$$
$$n_i = n(X_i)$$
$$n'_{ij} = n(X'_i, Y_j)$$
$$n'_i = n(X'_i)$$

である。

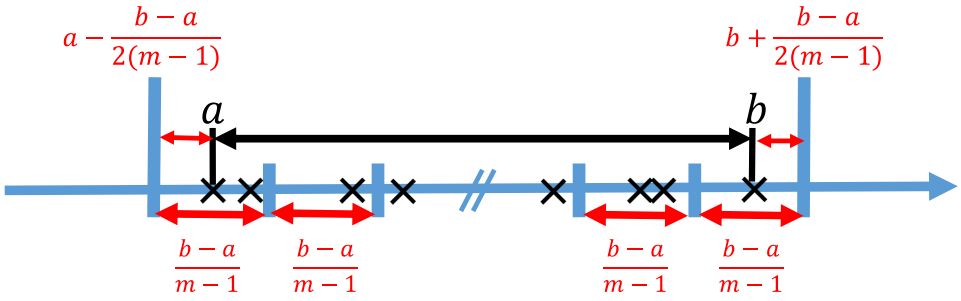
この式を用いて、次の手順で離散化を繰り返す：

① まずすべての数値を、値が同じ数値を同じ水準と見なして、カテゴリ化（順序尺度化）する。
② 隣接する水準を結合した場合の AIC 減少を求める。
③ AIC 減少幅が最大になる水準の組を探し、もしその値が正であれば、その水準同士を結合する。
④ この操作を AIC 減少幅の最大値が正になるものがなくなるか、指定した離散化数になるまで繰り返す。

このようにして数値変数を離散化した後で、「目的変数がカテゴリ、説明変数がカテゴリ」の方法と同じようにして計算を行う。

7. 数値変数のヒストグラムによる離散化

ヒストグラムの分割は、対象となる変数を m 分割した場合の AIC を算出し、その AIC が最小となる分割数を選ぶ。目的変数となる数値変数の最小値・最大値がそれぞれ a, b である場合に、m 分割ヒストグラムの AIC の算出は、図のように最小値・最大値を両端に $\frac{b-a}{2(m-1)}$だけ拡張し、最小値・最大値をそれぞれ $a - \frac{b-a}{2(m-1)}, b + \frac{b-a}{2(m-1)}$ とすると、拡張されたデータ範囲の長さは、$\frac{m}{m-1} \times (b-a)$ となるので、これを区間幅 $\frac{b-a}{m-1}$ の均等な m 個の区間を持つヒストグラムについて考える。



この場合の AIC は、全データ数を N, 分割した各区間 i の頻度を n_i, 各区間の推定確率を $\hat{p}_i = n_i/N$, データの最小粒度を d として、

$$-2\sum_i n_i \log \hat{p}_i + 2N \log \frac{b-a}{d(m-1)} + 2m$$

となる。第 2 項はヒストグラムを最小粒度の頻度表からヒストグラムの粒度の頻度表に変えたことによる補正項である。ここで分割数 m と関連のない定数項を除外すると、

$$-2\sum_i n_i \log \hat{p}_i - 2N \log(m-1) + 2m$$

となり、ヒストグラムの AIC 最小となる m を求める際には、変数ごとに分割数 m についてのこの値を求め、この値が最小となる m を用いる。

なお、m の探索範囲については、最適な階級数を求める簡便な式であるスタージェスの式

$$c' = 1 + \frac{\log_{10} N}{\log_{10} 2}$$

の階級数 c' を参考値として、2 から$2\lfloor c' \rfloor$までを探索範囲とする。

ヒストグラムを利用する場合で、かつ目的変数が数値である場合は、上記の基準で最小の AIC となる分割数で目的変数を離散化する。さらに説明変数との相関を計算する際、次の両方を計算して値の低い方を選ぶ。

- 目的変数を離散化しない場合の AIC 差
- 目的変数を離散化した場合の AIC 差