線形判別分析

青木繁伸

2020 年 3 月 17 日

1　目的

線形判別分析を行う。

2　使用法

```
import sys
sys.path.append("statlib")
from multi import disc
disc(data, make_dummy=False, verbose=True)
```

2.1　引数

data	分類に必要な説明変数と，群を表す変数が最終列（最も右側の列）になるように用意されたデータフレーム
make_dummy	ダミー変数への変換が必要な場合には True を指定する。デフォルトは False
verbose	必要最小限のプリント出力をする

2.2　戻り値の名前

"ncase"	サンプルサイズ
"p"	説明変数の個数
"vname"	説明変数の名前
"gVname"	群変数の名前
"ng"	群の数
"gName"	群の名前
"num"	各群のサンプルサイズ
"t"	全体の変動・共変動行列
"w"	群内変動・共変動
"names2"	2 群判別の対象とする群の名前
"cFunction"	分類関数の係数
"dFunction"	判別関数の係数

"partialF"	分類関数の係数の検定統計量
"partialP"	分類関数の係数の検定統計量の p 値
"df1"	第 1 自由度
"df2"	第 2 自由度
"wilksLambda"	ウィルクスの Λ
"wilksLambdaF"	ウィルクスの Λ の検定統計量
"wilksLambdaP"	ウィルクスの Λ の検定統計量の p 値
"wilksLambdaDf1"	第 1 自由度
"wilksLambdaDf2"	第 2 自由度
"distance"	各群の重心からの距離（χ^2 分布にしたがう）
"df"	自由度（説明変数の個数に等しい）
"Pvalue"	各群の所属する確率 p
"prediction"	どの群に所属するか判別
"correct"	正しい判別がされたデータ数
"correctTable"	判別表
"correctRate"	正判別率
"discriminantValue"	判別値（2 群の場合）

3 使用例

3.1 3 群判別の場合

```
import pandas as pd

data = pd.read_csv("data/iris.csv")

import sys
sys.path.append("statlib")
from multi import disc

a = disc(data)
```

```
coefficients of discriminant functions
          setosa:versicolor  setosa:virginica  versicolor:virginica  \
sl                -7.845958        -11.098318             -3.252360
sw               -16.515361        -19.902591             -3.387230
pl                21.642090         29.197184              7.555094
pw                23.832640         38.477524             14.644884
constant          13.455863        -18.059850            -31.515713

                  F       p value
sl         4.721152  1.032884e-02
```

```
sw        21.935928  4.831201e-09
pl        35.590175  2.756205e-13
pw        24.904333  5.143154e-10
constant        NaN           NaN

coefficients of classification functions
              setosa  versicolor   virginica
sl        -47.088333  -31.396418  -24.891698
sw        -47.175741  -14.145020   -7.370559
pl         32.861278  -10.422902  -25.533090
pw         34.796822  -12.868458  -42.158226
constant  170.419715  143.507990  206.539415

result of discrimination
            setosa  versicolor  virginica
setosa          50           0          0
versicolor       0          48          2
virginica        0           1         49

correct rate = 98.0 %
```

3.2 2 群判別の場合

```
data = data.iloc[0:100, :]

a = disc(data)
```

```
coefficients of discriminant functions
          setosa:versicolor          F   p value
sl                -3.052770   0.715527  0.399741
sw               -18.022959  25.700452  0.000002
pl                21.766195  24.597600  0.000003
pw                30.844165  10.705452  0.001490
constant         -13.961740        NaN       NaN

coefficients of classification functions
              setosa  versicolor
sl        -50.078491  -43.972951
sw        -36.233259   -0.187341
pl          7.889082  -35.643307
pw         70.340670    8.652340
```

```
constant  173.031449  200.954928

result of discrimination
            setosa  versicolor
setosa          50           0
versicolor       0          50

correct rate = 100.0 %
```

3.3 ダミー変数への変換が必要な場合

ファイル "data/disc-2.csv" はデータの値は数値ではなく文字列で入力されている。このような場合，disc() は make_dummy=True によりダミー変数への変換を行う。

なお，文字列とはいっても，単に数字を文字列（"1", "2" など）にしただけではダミー変数に変換してくれないので注意が必要である。

```
dat = pd.read_csv("data/disc-2.csv")
print(dat.iloc[:10, :])
```

```
     X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11   X12  \
0   two  zero   one   one  zero   one  zero   one   one   two   two  zero
1  zero  zero  zero   one   one  zero  zero  zero  zero   two   two  zero
2   one   one  zero  zero   two  zero  zero  zero  zero  zero   one  zero
3   one  zero   one   two   two   one   one  zero  zero   two   two   one
4  zero  zero  zero   one   two  zero  zero   one   two   one  zero  zero
5   two   one  zero  zero  zero   one   two  zero   two  zero  zero   one
6   two  zero   one   two  zero   one   two   one   one  zero  zero  zero
7   two   one  zero   one   one  zero   two  zero  zero  zero   one  zero
8  zero   one  zero   one  zero  zero   one  zero   two   two  zero  zero
9   one   one   one  zero  zero   one   one  zero   two   one  zero   one

      Y
0   one
1  zero
2  zero
3   one
4  zero
5   one
6   one
7  zero
8  zero
9   one
```

```
a = disc(dat, make_dummy=True)
```

```
coefficients of discriminant functions
          one:zero         F   p value
X1_two   -2.359861  5.310797  0.023787
X1_zero   0.648009  0.441030  0.508535
X2_zero  -1.565254  3.675798  0.058778
X3_zero   2.379234  6.912939  0.010259
X4_two   -1.343930  2.051702  0.155932
X4_zero  -0.167736  0.024722  0.875457
X5_two    0.471475  0.256853  0.613684
X5_zero  -2.699780  7.675514  0.006957
X6_zero   1.315415  2.434396  0.122647
X7_two   -1.555243  2.827480  0.096564
X7_zero   0.396980  0.172979  0.678591
X8_zero  -0.981076  1.551497  0.216550
X9_two   -0.425616  0.189369  0.664614
X9_zero   0.381231  0.150127  0.699443
X10_two   0.371998  0.142256  0.707047
X10_zero  1.575304  2.350639  0.129177
X11_two   0.662651  0.440749  0.508670
X11_zero  0.418863  0.171787  0.679637
X12_zero  1.871882  6.670181  0.011626
constant -0.253238       NaN       NaN

coefficients of classification functions
                one       zero
X1_two   -10.656689  -5.936967
X1_zero   -5.954703  -7.250722
X2_zero   -5.886602  -2.756095
X3_zero   -0.638733  -5.397202
X4_two    -8.075466  -5.387606
X4_zero   -9.714089  -9.378617
X5_two    -4.582641  -5.525592
X5_zero   -8.809773  -3.410212
X6_zero   -7.010496  -9.641326
X7_two   -10.044999  -6.934513
X7_zero   -3.790067  -4.584026
X8_zero   -2.746170  -0.784018
X9_two    -9.709664  -8.858432
X9_zero   -7.204916  -7.967378
```

```
X10_two   -9.522287 -10.266284
X10_zero  -9.277515 -12.428124
X11_two   -5.761038  -7.086341
X11_zero  -7.576423  -8.414149
X12_zero  -2.523230  -6.266994
constant  24.427116  24.933593

result of discrimination
      one  zero
one    40    10
zero    5    45

correct rate = 85.0 %
```