

# Package ‘visualpred’

October 24, 2020

**Title** Visualization 2D of Binary Classification Models

**Version** 0.1.0

**Description** Visual 2D point and contour plots for binary classification modeling under algorithms such as `glm()`, `randomForest()`, `gbm()`, `nnet()` and `svm()`, presented over two dimensions generated by FAMD and MCA methods. Package 'FactoMineR' for multivariate reduction functions and package 'MBA' for interpolation functions are used. The package can be used to visualize the discriminant power of input variables and algorithmic modeling, explore outliers, compare algorithm behaviour, etc. It has been created initially for teaching purposes, but it has also many practical uses.

**License** GPL (>= 3)

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.1.1

**Imports** gbm, randomForest, nnet (>= 7.3.12), e1071, MASS (>= 7.3.51.4), magrittr, FactoMineR (>= 2.3), ggplot2 (>= 3.3.0), mltools, dplyr, data.table, MBA, pROC, ggrepel

**Suggests** knitr, markdown,egg

**VignetteBuilder** knitr

**Depends** R (>= 3.5.0)

**NeedsCompilation** no

**Author** Javier Portela [aut, cre]

**Maintainer** Javier Portela <javipgm@gmail.com>

**Repository** CRAN

**Date/Publication** 2020-10-24 09:40:02 UTC

## R topics documented:

breastwisconsin1 . . . . .	2
famdcontour . . . . .	2
famdcontourlabel . . . . .	5
Hmda . . . . .	6

mcacontour . . . . .	7
mcacontourjit . . . . .	9
mcamodelobis . . . . .	10
nba . . . . .	11
pima . . . . .	11
spiral . . . . .	12

<b>Index</b>	<b>13</b>
--------------	-----------

---

breastwisconsin1	<i>Breast Cancer Wisconsin dataset</i>
------------------	--

---

### Description

Breast Cancer Wisconsin dataset

### Usage

```
data(breastwisconsin1)
```

### Format

An object of class `data.frame` with 699 rows and 10 columns.

### Source

[https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original))

---

famdcontour	<i>Contour plots and FAMD function for classification modeling</i>
-------------	--

---

### Description

This function presents visual graphics by means of FAMD. FAMD function is Factorial Analysis for Mixed Data (interval and categorical) Dependent classification variable is set as supplementary variable. Machine learning algorithm predictions are presented in a filled contour setting

### Usage

```
famdcontour(dataf=dataf,listconti,listclass,vardep,proba="",
title="",title2="",depcol="",listacol="",alpha1=0.7,alpha2=0.7,alpha3=0.7,
classvar=1,intergrid=0,selec=0,modelo="glm",nodos=3,maxit=200,decay=0.01,
sampsize=400,mtry=2,nodesize=10,ntree=400,ntreegbm=500,shrink=0.01,
bag.fraction=1,n.minobsinnode=10,C=100,gamma=10,Dime1="Dim.1",Dime2="Dim.2")
```

**Arguments**

dataf	data frame.
listconti	Interval variables to use, in format c("var1","var2",...).
listclass	Class variables to use, in format c("var1","var2",...).
vardep	Dependent binary classification variable.
proba	vector of probability predictions obtained externally (optional)
title	plot main title
title2	plot subtitle
depcol	vector of two colors for points
listacol	vector of colors for labels
alpha1	alpha transparency for majoritary class
alpha2	alpha transparency for minority class
alpha3	alpha transparency for fit probability plots
classvar	1 if dependent variable categories are plotted as supplementary
intergrid	scale of grid for contour:0 if automatic
selec	1 if stepwise logistic variable selection is required, 0 if not.
modelo	name of model: "glm","gbm","rf","nnet","svm".
nodos	nnet: nodes
maxit	nnet: iterations
decay	nnet: decay
sampsize	rf: sampsize
mtry	rf: mtry
nodesize	rf: nodesize
ntree	rf: ntree
ntreegbm	gbm: ntree
shrink	gbm: shrink
bag.fraction	gbm: bag.fraction
n.minobsinnode	gbm:n.minobsinnode
C	svm Radial: C
gamma	svm Radial: gamma
Dime1, Dime2	FAMD Dimensions to consider. Dim.1 and Dim.2 by default.

**Details**

FAMD algorithm from FactoMineR package is used to compute point coordinates on dimensions (Dim.1 and Dim.2 by default). Minority class on dependent variable category is represented as red, majority category as green. Color scheme can be altered using depcol and listacol, as well as alpha transparency values.

**Predictive modeling:**

For predictive modeling, `selec=1` selects variables with a simple stepwise logistic regression. By default `select=0`. Logistic regression is used by default. Basic parameter setting is supported for algorithms `nnet`, `rf`, `gbm` and `svm-RBF`. A vector of fitted probabilities obtained externally from other algorithms can be imported in parameter `proba=nameofvector`. Contour curves are then computed based on this vector.

**Contour curves:**

Contour curves are build by the following process: i) the chosen algorithm model is trained and all observations are predicted-fitted. ii) A grid of points on the two chosen FAMD dimensions is built iii) package `MBA` is used to interpolate probability estimates over the grid, based on previously fitted observations.

**Variable representation:**

In order to represent interval variables, categories of class variables, and points in the same plot, a proportional projection of interval variables coordinates over the two dimensions range is applied. Since space of input variables is frequently larger than two dimensions, sometimes overlapping of points is produced; a frequency variable is used, and alpha values may be adjusted to avoid wrong interpretations of the presence of dependent variable category/color.

**Troubleshooting:**

- Check missings. Missing values are not allowed.
- By default `selec=0`. Setting `selec=1` may sometimes imply that no variables are selected; an error message is shown in this case.
- Models with only two input variables could lead to plot generation problems.
- Be sure that variables named in `listconti` are all numeric.
- If some numeric variable is constant at one single value, process is stopped since numeric Min-max standardization is performed, and NaN values are generated.

**Value**

A list with the following objects:

**graph1** plot of points on FAMD first two dimensions

**graph2** plot of points and contour curves

**graph3** plot of points and variables

**graph4** plot of points variable and contour curves

**graph5** plot of points colored by fitted probability

**graph6** plot of points colored by abs difference

**df1** data frame used for graph1

**df2** data frame used for contour curves

**df3** data frame used for variable names

**listconti** interval variables used-selected

**listclass** class variables used-selected

## References

Pages J. (2004). Analyse factorielle de donnees mixtes. Revue Statistique Appliquee. LII (4). pp. 93-111.

## Examples

```
data(breastwisconsin1)
dataf<-breastwisconsin1
listconti=c( "clump_thickness", "uniformity_of_cell_shape", "mitosis")
listclass=c("")
vardep="classes"
result<-famdcontour(dataf=dataf, listconti, listclass, vardep)
```

---

famdcontourlabel	<i>Outliers in Contour plots and FAMD function for classification modeling</i>
------------------	--

---

## Description

This function adds outlier marks to famdcontour using ggrepel package.

## Usage

```
famdcontourlabel(
  dataf = dataf,
  Idt = "",
  inf = 0.1,
  sup = 0.9,
  cutprob = 0.5,
  ...
)
```

## Arguments

dataf	data frame.
Idt	Identification variable, default "", row number
inf, sup	Quantiles for x,y outliers
cutprob	cut point for outliers based on prob. estimation error
...	options to be passed from famdcontour

## Details

An identification variable can be set in Idt parameter. By default, number of row is used. There are two source of outliers: i) outliers in the two FAMD dimension space, where the cutpoints are set as quantiles given (inf=0.1 and sup=0.9 in both dimensions by default) and ii) outliers with respect to the fitted probability. The dependent variable is set to 1 for the minority class, and 0 for the majority class. Points considered outliers are those for which  $\text{abs}(\text{vardep}-\text{fittedprob})$  exceeds parameter cutprob.

**Value**

A list with the following objects:

**graph1\_graph6** plots for dimension outliers

**graph7\_graph12** plots for fit outliers

**Examples**

```
data(breastwisconsin1)
dataf<-breastwisconsin1
listconti=c( "clump_thickness", "uniformity_of_cell_shape", "mitosis")
listclass=c("")
vardep="classes"
result<-famdcntourlabel(dataf=dataf, listconti=listconti,
listclass=listclass, vardep=vardep)
```

---

Hmda

*Home Mortgage Disclosure Act dataset*

---

**Description**

Home Mortgage Disclosure Act dataset

**Usage**

```
data(Hmda)
```

**Format**

An object of class `data.frame` with 2380 rows and 13 columns.

**Source**

Stock, J. H. and Watson, M. W. (2007). Introduction to Econometrics, 2nd ed. Boston: Addison Wesley.

**Description**

This function presents visual graphics by means of Multiple correspondence Analysis projection. Interval variables are categorized to bins. Dependent classification variable is set as supplementary variable. Machine learning algorithm predictions are presented in a filled contour setting.

**Usage**

```
mcacontour(dataf=dataf,listconti,listclass,vardep,proba="",bins=8,
Dime1="Dim.1",Dime2="Dim.2",classvar=1,intergrid=0,selec=0,
title="",title2="",listacol="",depcol="",alpha1=0.8,alpha2=0.8,alpha3=0.7,modelo="glm",
nodos=3,maxit=200,decay=0.01,sampsize=400,mtry=2,nodesize=5,
ntree=400,ntreegbm=500,shrink=0.01,bag.fraction=1,n.minobsinnode=10,C=100,gamma=10)
```

**Arguments**

dataf	data frame.
listconti	Interval variables to use, in format c("var1","var2",...).
listclass	Class variables to use, in format c("var1","var2",...).
vardep	Dependent binary classification variable.
proba	vector of probability predictions obtained externally (optional)
bins	Number of bins for categorize interval variables .
Dime1	FAMD Dimensions to consider. Dim.1 and Dim.2 by default.
Dime2	FAMD Dimensions to consider. Dim.1 and Dim.2 by default.
classvar	1 if dependent variable categories are plotted as supplementary
intergrid	scale of grid for contour:0 if automatic
selec	1 if stepwise logistic variable selection is required, 0 if not.
title	plot main title
title2	plot subtitle
listacol	vector of colors for labels
depcol	vector of two colors for points
alpha1	alpha transparency for majoritary class
alpha2	alpha transparency for minority class
alpha3	alpha transparency for fit probability plots
modelo	name of model: "glm","gbm","rf","nnet","svm".
nodos	nnet: nodes
maxit	nnet: iterations

decay	nnet: decay
sampsize	rf: sampsize
mtry	rf: mtry
nodesize	rf: nodesize
ntree	rf: ntree
ntreegbm	gbm: ntree
shrink	gbm: shrink
bag.fraction	gbm: bag.fraction
n.minobsinnode	gbm:n.minobsinnode
C	svm Radial: C
gamma	svm Radial: gamma

### Details

This function applies MCA (Multiple Correspondence Analysis) in order to project points and categories of class variables in the same plot. In addition, interval variables listed in `listconti` are categorized to the number given in `bins` parameter (by default 8 bins). Further explanation about machine learning classification and contour curves, see the `famdcontour` function documentation.

### Value

A list with the following objects:

- graph1** plot of points on MCA two dimensions
- graph2** plot of points and variables
- graph3** plot of points and contour curves
- graph4** plot of points, contour curves and variables
- graph5** plot of points colored by fitted probability
- graph6** plot of points colored by abs difference
- df1** dataset used for graph1
- df2** dataset used for graph2
- df3** dataset used for graph3
- df4** dataset used for graph4
- listconti** interval variables used
- listclass** class variables used
- ... color schemes and other parameters

### Examples

```
data(breastwisconsin1)
dataf<-breastwisconsin1
listconti=c( "clump_thickness", "uniformity_of_cell_shape", "mitosis")
listclass=c("")
vardep="classes"
result<-mcacontour(dataf=dataf, listconti, listclass, vardep)
```



---

mcacontourjit

*Contour plots and MCA function for classification modeling*


---

### Description

This function is similar to mcacontour but points are jittered in every plot

### Usage

```
mcacontourjit(dataf=dataf,jit=0.1,alpha1=0.8,alpha2=0.8,alpha3=0.7,title="",...)
```

### Arguments

dataf	data frame.
jit	jit distance. Default 0.1.
alpha1	alpha transparency for majoritary class
alpha2	alpha transparency for minority class
alpha3	alpha transparency for fit probability plots
title	plot main title
...	options to be passed from mcacontour

### Value

A list with the following objects:

- graph1** plot of points on MCA two dimensions
- graph2** plot of points and variables
- graph3** plot of points and contour curves
- graph4** plot of points, contour curves and variables
- graph5** plot of points colored by fitted probability
- graph6** plot of points colored by abs difference

### Examples

```
data(breastwisconsin1)
dataf<-breastwisconsin1
listconti=c( "clump_thickness", "uniformity_of_cell_shape", "mitosis")
listclass=c("")
vardep="classes"
result<-mcacontourjit(dataf=dataf,listconti=listconti,listclass=listclass,vardep=vardep,jit=0.1)
```

---

mcamodelobis

*Basic MCA function for clasification*


---

### Description

This function presents visual graphics by means of Multiple correspondence Analysis projection. Interval variables are categorized to bins. Dependent classification variable is set as supplementary variable. It is used as base for mcacontour function.

### Usage

```
mcamodelobis(dataf=dataf,listconti,listclass, vardep,bins=8,selec=1,
Dime1="Dim.1",Dime2="Dim.2")
```

### Arguments

dataf	data frame.
listconti	Interval variables to use, in format c("var1","var2",...).
listclass	Class variables to use, in format c("var1","var2",...).
vardep	Dependent binary classification variable.
bins	Number of bins for categorize interval variables .
selec	1 if stepwise logistic variable selection is required, 0 if not.
Dime1, Dime2	MCA Dimensions to consider. Dim.1 and Dim.2 by default.

### Value

A list with the following objects:

- df1** dataset used for graph1
- df2** dataset used for graph2
- df3** dataset used for graph2
- listconti** interval variables used
- listclass** class variables used
- axisx** axis definition in plot
- axisy** axis definition in plot

### Examples

```
data(breastwisconsin1)
dataf<-breastwisconsin1
listconti=c( "clump_thickness","uniformity_of_cell_shape","mitosis")
listclass=c("")
vardep="classes"
result<-mcacontour(dataf=dataf,listconti,listclass,vardep,bins=8,title="",selec=1)
```

---

nba

*nba dataset*

---

**Description**

nba dataset

**Usage**

```
data(nba)
```

**Format**

An object of class `data.frame` with 1340 rows and 21 columns.

**Source**

<https://data.world/exercises/logistic-regression-exercise-1>

---

pima

*Pima indian diabetes dataset*

---

**Description**

Pima indian diabetes dataset

**Usage**

```
data(pima)
```

**Format**

An object of class `data.frame` with 768 rows and 9 columns.

**Source**

<https://sci2s.ugr.es/keel/dataset.php?cod=21>

---

spiral	<i>spiral sample data</i>
--------	---------------------------

---

**Description**

spiral sample data

**Usage**

```
data(spiral)
```

**Format**

An object of class `data.frame` with 803 rows and 3 columns.

# Index

- \* **FAMD**
    - famdcontour, 2
    - famdcontourlabel, 5
  - \* **MCA**
    - mcacontour, 7
    - mcacontourjit, 9
    - mcamodelobis, 10
  - \* **classification**
    - famdcontour, 2
    - famdcontourlabel, 5
    - mcacontour, 7
    - mcacontourjit, 9
  - \* **contour\_curves**
    - famdcontour, 2
    - famdcontourlabel, 5
    - mcacontour, 7
    - mcacontourjit, 9
  - \* **datasets**
    - breastwisconsin1, 2
    - Hmda, 6
    - nba, 11
    - pima, 11
    - spiral, 12
  - \* **outliers**
    - famdcontourlabel, 5
- breastwisconsin1, 2
- famdcontour, 2
- famdcontourlabel, 5
- Hmda, 6
- mcacontour, 7
- mcacontourjit, 9
- mcamodelobis, 10
- nba, 11
- pima, 11
- spiral, 12