Package ‘partitionMap’

February 20, 2015

Type Package

Title Partition Maps

Version 0.5

Date 2013-01-18

Author Nicolai Meinshausen

Maintainer Nicolai Meinshausen <meinshausen@stats.ox.ac.uk>

Depends randomForest

Imports stats, graphics

Description Low-dimensional embedding, using Random Forests for multiclass classification

License GPL

LazyLoad no

URL http://www.stats.ox.ac.uk/~meinshau

Repository CRAN

Date/Publication 2013-01-18 15:55:12

NeedsCompilation no

R topics documented:

partitionMap ................................................................. 2
Soybean ............................................................... 3

Index 6
**partitionMap**

**Partition Maps**

**Description**

using Random Forest multiclass output, embed observations in low-dimensional space

**Usage**

```
partitionMap(x, y, XTEST = NULL, YTEST = NULL, method = "pm", dimen = 2,
force = TRUE, ntree = 100,
plottrain = TRUE, addjitter = 0.03, ...)
```

**Arguments**

- **x**: matrix with predictor variables in the training dataset
- **y**: response variable, a factor with multiple classes
- **XTEST**: The matrix of predictor variables for the test dataset (optional)
- **YTEST**: Class labels of test observations, used for coloring the test embeddings in the plot. If not supplied, test observations are shown in grey (optional)
- **method**: pm for "partitionMap" and ha for "Homogeneity Analysis"
- **dimen**: dimension of embedding, typically 2 or 3
- **force**: use force-based variation of "partitionMap" algorithm? no effect if method="ha"
- **ntree**: number of trees to use for randomForest prediction
- **plottrain**: plot embedding for training data?
- **addjitter**: amount of jitter to add to the plots to avoid overlapping observations (set addjitter=0 for no jitter)
- **...**: other arguments to be passed to randomForest

**Value**

A list with values

- **Samples**: low-dimensional co-ordinates of embedded training samples
- **Rules**: low-dimensional co-ordinates of embedded Rules (nodes in the trees)
- **Z**: a binary matrix, with as many rows as training samples and as many columns as rules. a value 1 in row i and column j indicates that observation i is part of rule j
- **SamplesTest**: low-dimensional co-ordinates of embedded test samples
- **ZTest**: a binary matrix, with as many rows as test samples and as many columns as rules. a value 1 in row i and column j indicates that observation i in the test data is part of rule j
- **rf**: the trained Random Forest classifier
Examples

```r
##---- load Soybean data ----
data(Soybean)
X <- Soybean[,,-1]
Y <- Soybean$Y

##---- divide into training and test data ----
indtrain <- rep(0,nrow(X))
indtrain[sample(1:length(indtrain), ceiling(nrow(X)/3*2))] <- 1
XTEST <- X[indtrain==0]
YTEST <- Y[indtrain==0]
X <- X[indtrain==1]
Y <- Y[indtrain==1]

##---- compute Partition Map solution ----
pm <- partitionMap(X,Y,XTEST=YTEST,method="pm",force=TRUE,
dimen=2,ntree=80,plottrain=TRUE)

##---- plot the embedded training and test samples ----
par(mfrow=c(1,1))
plot(pm$Samples,col=Y,pch=20,cex=1.5,main="Training Data",
  xlab="Dimension 1",ylab="Dimension 2")
points(pm$Rules,pch="." )
plot(pm$Samplestest,col=YTEST,pch=20,cex=1.5,main="Test Data",
  xlab="Dimension 1",ylab="Dimension 2")
points(pm$Rules,pch=".")
plot(pm$Samples,col=Y,pch=20,cex=1.5,xlab="",ylab="",type="n",axes=FALSE)
legend(quantile(pm$Samples[,1],0),quantile(pm$Samples[,2],1),unique(Y),
  col=1:length(unique(Y)),fill=1:length(unique(Y)),border=0)
par(mfrow=c(1,1))
```
Description
There are 19 classes, only the first 15 of which have been used in prior work. The folklore seems to be that the last four classes are unjustified by the data since they have so few examples. There are 35 categorical attributes, some nominal and some ordered. The values for attributes are encoded numerically, with the first value encoded as “0,” the second as “1,” and so forth. Observations with missing values in the original dataset have been removed.

Usage
data(Soybean)

Format
A data frame with 562 observations on the following 36 variables.

Y the 19 classes
date apr(0),may(1),june(2),july(3),aug(4),sept(5),oct(6)
plant.stand normal(0),lt-normal(1)
precip lt-norm(0),norm(1),gt-norm(2)
temp lt-norm(0),norm(1),gt-norm(2)
hail yes(0),no(1)
crop.hist dif-lst-yr(0),s-l-yr(1),s-l-2-yr(2),s-l-7-yr(3)
area.dam scatter(0),low-area(1),upper-ar(2),whole-field(3)
sever minor(0),pot-severe(1),severe(2)
seed.tmt none(0),fungicide(1),other(2)
germ 90-100(0),80-89(1),lt-80(2)
plant.growth norm(0),abnorm(1)
leaves norm(0),abnorm(1)
leaf.halo absent(0),yellow-halos(1),no-yellow-halos(2)
leaf.marg w-s-marg(0),no-w-s-marg(1),dna(2)
leaf.size lt-1/8(0),gt-1/8(1),dna(2)
leaf.shread absent(0),present(1)
leaf.malf absent(0),present(1)
leaf.mild absent(0),upper-surf(1),lower-surf(2)
stem norm(0),abnorm(1)
lodging yes(0),no(1)
stem.cankers absent(0),below-soil(1),above-s(2),ab-sec-nde(3)
canker.lesion dna(0),brown(1),dk-brown-blk(2),tan(3)
fruiting.bodies absent(0),present(1)
ext.decay absent(0),firm-and-dry(1),watery(2)
mycelium absent(0),present(1)
Soybean

int.discolor none(0),brown(1),black(2)
sclerotia absent(0),present(1)
fruit.pods norm(0),diseased(1),few-present(2),dna(3)
fruit.spots absent(0),col(1),br-w/blk-speck(2),distort(3),dna(4)
seed norm(0),abnorm(1)
mold.growth absent(0),present(1)
seed.discolor absent(0),present(1)
seed.size norm(0),lt-norm(1)
shriveling absent(0),norm(1)
roots norm(0),rotted(1),galls-cysts(2)

Source

Donor: Ming Tan & Jeff Schlimmer (Jeff.Schlimmer@cs.cmu.edu)
These data have been taken from the UCI Repository Of Machine Learning Databases at
* <URL: http://www.ics.uci.edu/~mlearn/MLRepository.html>
and were converted to R format by Evgenia Dimitriadou, as were copied from the mlbench package.

References


Examples

data(Soybean)
Index

*Topic datasets
  Soybean, 3

partitionMap, 2

Soybean, 3