Package 'clustMD'

July 2, 2014

Type Package

Title Model based clustering for mixed data.

Version 1.0

Date 2014-05-30

Author Damien McParland

Maintainer Damien McParland <damien.mcparland@ucd.ie>

Description Model-based clustering of mixed data (i.e. data which consist of continuous, binary, ordinal or nominal variables) using a parsimonious mixture of latent Gaussian variable models.

Imports tmvtnorm, mvtnorm, truncnorm, MASS, mclust, msm

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2014-06-12 15:14:48

R topics documented:

clustMD-package *Model based clustering for mixed data*

Description

Model-based clustering of mixed data (i.e. data that consist of continuous, binary, ordinal or nominal variables) using a parsimonious mixture of latent Gaussian variable models.

Details

Package:	clustMD
Type:	Package
Version:	1.0
Date:	2014-05-30
License:	GPL-2

Author(s)

Damien McParland

Damien McParland <damien.mcparland@ucd.ie> Isobel Claire Gormley <claire.gormley@ucd.ie>

References

McParland, D. and Gormley, I.C. (2014). Model based clustering for mixed data: clustMD. Technical report, University College Dublin.

Byar *Byar prostate cancer data set.*

Description

A data set consisting of variables of mixed type measured on a group of prostate cancer patients.

Usage

```
data(Byar)
```

Format

A data frame with 475 observations on the following 15 variables.

`Age` a numeric vector indicating the age of the patient.

`Weight` a numeric vector indicating the weight of the patient.

`Performance.rating` an ordinal variable indicating how active the patient is: 0 - normal activity, 1 - in bed less than 50% of daytime, 2 - in bed more than 50% of daytime, 3 - confined to bed.

`Cardiovascular.disease.history` a binary variable indicating if the patient has a history of cardiovascular disease: 0 - no, 1 - yes.

`Systolic.Blood.pressure` a numeric vector indicating the systolic blood pressure of the patient in units of ten.

`Diastolic.blood.pressure` a numeric vector indicating the diastolic blood pressure of the patient in units of ten.

`Electrocardiogram.code` a nominal variable indicating the electorcardiogram code: 0 - normal, 1 - benign, 2 - rythmic disturbances and electrolyte changes, 3 - heart blocks or conduction defects, 4 - heart strain, 5 - old myocardial infarct, 6 - recent myocardial infarct.

`Serum.haemoglobin` a numeric vector indicating the serum haemoglobin levels of the patient measured in g/100ml.

`Size.of.primary.tumour` a numeric vector indicating the estimated size of the patient's primary tumour in centimeters squared.

`Index.of.tumour.stage.and.histolic.grade` a numeric vector indicating the combined index of tumour stage and histolic grade of the patient.

`Serum.prostatic.acid.phosphatase` a numeric vector indicating the serum prostatic acid phosphatase levels of the patient in King-Armstong units.

`Bone.metastases` a binary vector indicating the presence of bone metastasis: 0 - no, 1 - yes.

`Stage` the stage of the patient's prostate cancer.

`Observation` a patient ID number.

`SurvStat` the post trial survival status of the patient: 0 - alive, 1 - dead from prostatic cancer, 2 - dead from heart or vascular disease, 3 - dead from cerebrovascular accident, 3 - dead form pulmonary ebolus, 5 - dead from other cancer, 6 - dead from respiratory disease, 7 - dead from other specific non-cancer cause, 8 - dead from other unspecified non-cancer cause, 9 - dead from unknown cause.

Details

A data set consisting of variables of mixed type measured on a group of prostate cancer patients. Patients have either stage 3 or stage 4 prostate cancer.

Source

Andrews, D.A., Herzberg, A.M. (1985). Data: A collection of Problems from Many Fields for the Student and Research Worker. Springer.

References

Byar, D.P. and Green, S.B. (1980). The choice of treatment for cancer patients based on covariate information: applications to prostate cancer. Bulletin du Cancer 67: 477-490.

Hunt, L., Jorgensen, M. (1999). Mixture model clustering using the multimix program. Australia and New Zealand Journal of Statistics 41: 153-171.

Examples

```
data(Byar)
```

`clustMD` *Model-Based Clustering for Mixed Data*

Description

A function which fits the clustMD model to a data set consisting of any combination of continuous, binary, ordinal and nominal variables.

Usage

```
clustMD(X, G, CnsIndx, OrdIndx, Nnorms, MaxIter, model, store.params = FALSE)
```

Arguments

`X`	A data matrix where the variables are ordered so that the continuous variables come first, the binary (coded 1 and 2) and ordinal variables (coded 1, 2,...) come second and the nominal variables (coded 1, 2,...) are in last position.
`G`	The number of mixture components to be fitted.
`CnsIndx`	The number of continuous variables in the data set.
`OrdIndx`	The sum of the number of continuous, binary and ordinal variables in the data set.
`Nnorms`	The number of Monte Carlo samples to be used for the intractable E-step in the presence of nominal data.
`MaxIter`	The number of iterations for which the (MC)EM algorithm should run.
`model`	A string indicating which clustMD model is to be fitted. This may be one of: EII, VII, EEI, VEI, EVI or VVI.
`store.params`	A logical variable indicating if the parameter estimates at each iteration should be saved and returned by the clustMD function.

Details

Model-based clustering of mixed data using a parsimonious mixture of latent Gaussian variables.

Value

A list is returned:

`cl`	The cluster to which each observation belongs.
`tau`	A N x G matrix of the conditional probabilities of each observation blonging to each cluster.
`means`	A D x G matrix of the cluster means.
`A`	A G x D matrix containing the diagonal entries of the A matrix corresponding to each cluster.
`Lambda`	A G x D matrix of volume parameters corresponding to each observed or latent dimension for each cluster.
`Sigma`	A D x D x G array of the covariance matrices for each cluster.
`BIChat`	The estimated Bayesian information criterion for the model fitted.
`paramlist`	If store.params is true then paramlist is a list of the stored parameter values in the order given above with the saved estimated likelihood values in last position.

Author(s)

Damien McParland

References

McParland, D. and Gormley, I.C. (2014). Model based clustering for mixed data: clustMD. Technical report, University College Dublin.

Examples

```
data(Byar)

# Transformation skewed variables
Byar$Size.of.primary.tumour <- sqrt(Byar$Size.of.primary.tumour)
Byar$Serum.prostatic.acid.phosphatase <- log(Byar$Serum.prostatic.acid.phosphatase)

# Order variables (Continuous, ordinal, nominal)
Y <- as.matrix(Byar[, c(1, 2, 5, 6, 8, 9, 10, 11, 3, 4, 12, 7)])

# Start categorical variables at 1 rather than 0
Y[, 9:12] <- Y[, 9:12] + 1

# Standardise continuous variables
Y[, 1:8] <- scale(Y[, 1:8])

# Merge categories of EKG variable for efficiency
Yekg <- rep(NA, nrow(Y))
Yekg[Y[,12]==1] <- 1
Yekg[(Y[,12]==2)|(Y[,12]==3)|(Y[,12]==4)] <- 2
Yekg[(Y[,12]==5)|(Y[,12]==6)|(Y[,12]==7)] <- 3
Y[, 12] <- Yekg
```

```
## Not run:
res <- clustMD(X=Y, G=3, CnsIndx=8, OrdIndx=11, Nnorms=20000,
MaxIter=100, model="EVI", store.params=FALSE)

## End(Not run)
```

Index