Package 'fairml'

September 10, 2022

Type Package

Title Fair Models in Machine Learning

Version 0.7

Date 2022-09-10

Depends R (>= 3.5.0)

Imports methods, optiSolve, CVXR, glmnet

Suggests lattice, parallel

Author Marco Scutari [aut, cre]

Maintainer Marco Scutari <scutari@bnlearn.com>

Description Fair machine learning regression models which take sensitive attributes into account in model estimation. Currently implementing Komiyama et al. (2018) <http://proceedings.mlr.press/v80/komiyama18a/komiyama18a.pdf>, Zafar et al. (2019) <https://www.jmlr.org/papers/volume20/18-262/18-262.pdf> and my own approach that uses ridge regression to enforce fairness.

License MIT + file LICENSE

LazyData yes

NeedsCompilation no

Repository CRAN

Date/Publication 2022-09-10 16:22:54 UTC

R topics documented:

`fairml-package` *Fair models in machine learning*

Description

Fair machine learning models: estimation, tuning and prediction.

Details

fairml implements key algorithms for learning machine learning models while enforcing fairness with respect to a set of observed sensitive (or protected) attributes.

Currently **fairml** implements the following algorithms (references below):

- `nclm()`: the non-convex formulation of fair linear regression model from Komiyama et al. (2018).
- `frrm()`: my fair (linear) ridge regression model.
- `fgrrm()`: my fair generalized (linear) ridge regression model.
- `zlrm()`: the fair logistic regression with covariance constraints from Zafar et al. (2019).
- `zlrm()`: a fair linear regression with covariance constraints following Zafar et al. (2019).

Author(s)

Marco Scutari
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)

Maintainer: Marco Scutari <scutari@bnlearn.com>

References

Komiyama J, Takeda A, Honda J, Shimao H (2018). "Nonconvex Optimization for Regression with Fairness Constraints". Proceedints of the 35th International Conference on Machine Learning (ICML), PMLR **80**:2737–2746.
`http://proceedings.mlr.press/v80/komiyama18a/komiyama18a.pdf`

Zafar BJ, Valera I, Gomez-Rodriguez M, Gummadi KP (2019). "Fairness Constraints: a Flexible Approach for Fair Classification". Journal of Machine Learning Research, 30:1–42.
`https://www.jmlr.org/papers/volume20/18-262/18-262.pdf`

adult *Census Income*

Description

Predict whether income exceeds $50K per year using the U.S. 1994 Census data.

Usage

```
data(adult)
```

Format

The data contains 30162 observations and 14 variables. See the UCI Machine Learning Repository for details.

Note

The data set has been pre-processed as in Zafar et al. (2019), with the following exceptions:

- the data do not include the test sample from the UCI repository;
- the variables `"capital_gain"` and `"capital_loss"` have been scaled by `1/1000`.

In that paper, `income` is the response variable, `sex` and `race` are the sensitive attributes and the remaining variables are used as predictors.

References

UCI Machine Learning Repository.
`https://archive.ics.uci.edu/ml/datasets/adult`

Examples

```
data(adult)

# short-hand variable names.
r = adult[, "income"]
s = adult[, c("sex", "race")]
p = adult[, setdiff(names(adult), c("income", "sex", "race"))]

## Not run:
m = zlrm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)

## End(Not run)
```

bank *Bank Marketing*

Description

Direct marketing campaigns (phone calls) of a Portuguese banking institution to make clients subscribe a term deposit.

Usage

```
data(bank)
```

Format

The data contains 41188 observations and 19 variables. See the UCI Machine Learning Repository for details.

Note

The data set has been pre-processed as in Zafar et al. (2019), with the following exceptions:

- the variable `duration` has been dropped in order to learn as realistic predictive model;
- the variable `pdays` has been dropped because it is not defined for the vast majority of samples;
- observations where `loan` is `"unknown"` have been dropped because the corresponding regression coefficient estimated by `glm()` is `NA`;
- the three observations where `default` is `"yes"` have been dropped to avoid errors in cross-validation (if all those three observations are in the test fold it is impossible to compute predictions from them).

In that paper, `subscribed` is the response variable, `age` is the sensitive attribute and the remaining variables are used as predictors.

References

UCI Machine Learning Repository.
`https://archive.ics.uci.edu/ml/datasets/bank+marketing`

Examples

```
data(bank)

# remove loans with unknown status, the corresponding coefficient is NA in glm().
bank = bank[bank$loan != "unknown", ]

# short-hand variable names.
r = bank[, "subscribed"]
s = bank[, c("age")]
p = bank[, setdiff(names(bank), c("subscribed", "age"))]
```

```
m = zlrm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)
```

`communities.and.crime` *Communities and Crime Data Set*

Description

Combined socio-economic data from the 1990 Census, law enforcement data from the 1990 LEMAS survey, and crime data from the 1995 FBI UCR for various communities in the United States.

Usage

```
data(communities.and.crime)
```

Format

The data contains 1969 observations and 104 variables. See the UCI Machine Learning Repository for details.

Note

The data set has been pre-processed as in Komiyama et al. (2018), with the following exceptions:

- the variable `community` has been dropped, as it is non-predictive and contains a sizeable number of missing values;
- the variables `LemasSwornFT`, `LemasSwFTPerPop`, `LemasSwFTFieldOps`, `LemasSwFTFieldPerPop`, `LemasTotalReq`, `LemasTotReqPerPop`, `PolicReqPerOffic`, `PolicPerPop`, `RacialMatchCommPol`, `PctPolicWhite`, `PctPolicBlack`, `PctPolicHisp`, `PctPolicAsian`, `PctPolicMinor`, `OfficAssgnDrugUnits`, `NumKindsDrugsSeiz`, `PolicAveOTWorked`, `PolicCars`, `PolicOperBudg`, `LemasPctPolicOnPatr`, `LemasGangUnitDeploy` and `PolicBudgPerPop` have been dropped because they have more than 80% missing values.

In that paper, `ViolentCrimesPerPop` is the response variable, `racepctblack` and `PctForeignBorn` are the sensitive attributes and the remaining variables are used as predictors.

References

UCI Machine Learning Repository:
`http://archive.ics.uci.edu/ml/datasets/communities+and+crime`

Examples

```
data(communities.and.crime)

# short-hand variable names.
cc = communities.and.crime[complete.cases(communities.and.crime), ]
r = cc[, "ViolentCrimesPerPop"]
s = cc[, c("racepctblack", "PctForeignBorn")]
p = cc[, setdiff(names(cc), c("ViolentCrimesPerPop", names(s)))]

m = nclm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)

m = frrm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)
```

---

compas *Criminal Offenders Screened in Florida*

Description

A collection of criminal offenders screened in Florida (US) during 2013-14.

Usage

```
data(compas)
```

Format

The data contains 5855 observations and the following variables:

- `age`, a continuous variable containing the age (in years) of the person;
- `juv_fel_count`, a continuous variable containing the number of juvenile felonies;
- `decile_score`, a continuous variable, the decile of the COMPAS score;
- `juv_misd_count`, a continuous variable containing the number of juvenile misdemeanors;
- `juv_other_count`, a continuous variable containing the number of prior juvenile convictions that are not considered either felonies or misdemeanors;
- `v_decile_score`, a continuous variable containing the predicted decile of the COMPAS score;
- `priors_count`, a continuous variable containing the number of prior crimes committed;
- `sex`, a factor with levels `"Female"` and `"Male"`;
- `two_year_recid`, a factor with two levels `"Yes"` and `"No"` (if the person has recidivated within two years);
- `race`, a factor encoding the race of the person;
- `c_jail_in`, a numeric variable containing the date in which the person entered jail (normalized between 0 and 1);

- c_jail_out, a numeric variable containing the date in which the person was released from jail (normalized between 0 and 1);
- c_offense_date, a numeric variable containing the date the offense was committed;
- screening_date, a numeric variable containing the date in which the person was screened (normalized between 0 and 1);
- in_custody, a numeric variable containing the date in which the person was placed in custody (normalized between 0 and 1);
- out_custody, a numeric variable containing the date in which the person was released from custody (normalized between 0 and 1);

Note

The data set has been pre-processed as in Komiyama et al. (2018), with the following exceptions:

- the race variable has not been reduced to a binary factor with levels "African-American" and "not African-American";
- the variables type_of_assessment, v_type_of_assessment have been dropped from the analysis because they take the same value for all observations;
- variables like c_jail_in and c_jail_out that encode dates have been jointly rescaled to preserve the temporal ordering of events.

In that paper, two_year_recid is the response variable, sex and race are the sensitive attributes and the remaining variables are used as predictors.

References

Angwin J, Larson J, Mattu S, Kirchner L (2016). "Machine Bias: Theres Software Used Around the Country to Predict Future Criminals."
https://www.propublica.org

Examples

```
data(compas)

# convert the response back to a numeric variable.
compas$two_year_recid = as.numeric(compas$two_year_recid) - 1

# short-hand variable names.
r = compas[, "two_year_recid"]
s = compas[, c("sex", "race")]
p = compas[, setdiff(names(compas), c("two_year_recid", "sex", "race"))]

m = nclm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)

m = frrm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)
```

`fairml.cv` *Cross-Validation for Fair Models*

Description

Cross-validation for the models in the **fairml** package.

Usage

```
fairml.cv(response, predictors, sensitive, method = "k-fold", ..., unfairness,
  model, model.args = list(), cluster)

cv.loss(x)
cv.unfairness(x)
cv.folds(x)
```

Arguments

`response`	a numeric vector, the response variable.
`predictors`	a numeric matrix or a data frame containing numeric and factor columns; the predictors.
`sensitive`	a numeric matrix or a data frame containing numeric and factor columns; the sensitive attributes.
`method`	a character string, either `k-fold`, `custom-folds` or `hold-out`. See below for details.
`...`	additional arguments for the cross-validation `method`.
`unfairness`	a positive number in [0, 1], the proportion of the explained variance that can be attributed to the sensitive attributes.
`model`	a character string, the label of the model. Currently `"nclm"`, `"frrm"`, `"fgrrm"`, `"zlm"` and `"zlrm"` are available.
`model.args`	additional arguments passed to the model.
`cluster`	an optional cluster object from package **parallel**, to process folds or subsamples in parallel.
`x`	an object of class `fair.kcv` or `fair.kcv.list`.

Details

The following cross-validation methods are implemented:

- *k-fold*: the data are split in `k` subsets of equal size. For each subset in turn, `model` is fitted on the other `k - 1` subsets and the loss function is then computed using that subset. Loss estimates for each of the `k` subsets are then combined to give an overall loss for data.
- *custom-folds*: the data are manually partitioned by the user into subsets, which are then used as in k-fold cross-validation. Subsets are not constrained to have the same size, and every observation must be assigned to one subset.

- *hold-out*: `k` subsamples of size `m` are sampled independently without replacement from the data. For each subsample, `model` is fitted on the remaining `m - length(response)` samples and the loss function is computed on the `m` observations in the subsample. The overall loss estimate is the average of the `k` loss estimates from the subsamples.

Cross-validation methods accept the following optional arguments:

- `k`: a positive integer number, the number of groups into which the data will be split (in k-fold cross-validation) or the number of times the data will be split in training and test samples (in hold-out cross-validation).
- `m`: a positive integer number, the size of the test set in hold-out cross-validation.
- `runs`: a positive integer number, the number of times k-fold or hold-out cross-validation will be run.
- `folds`: a list in which element corresponds to one fold and contains the indices for the observations that are included to that fold; or a list with an element for each run, in which each element is itself a list of the folds to be used for that run.

If cross-validation is used with multiple `runs`, the overall loss is the average of the loss estimates from the different runs.

The predictive performance of the models is measured using the mean square error as the loss function.

Value

`fairml.cv()` returns an object of class `fair.kcv.list` if `runs` is at least 2, an object of class `fair.kcv` if `runs` is equal to 1.

`cv.loss()` returns a numeric vector or a numeric matrix containing the values of the loss function computed for each run of cross-validation.

`cv.unfairness()` returns a numeric vectors containing the values of the unfairness criterion computed on the validation folds for each run of cross-validation.

`cv.folds()` returns a list containing the indexes of the observations in each of the cross-validation folds. In the cakse of k-fold cross-validation, if `runs` is larger than `1`, each element of the list is itself a list with the indexes for the observations in each fold in each run.

Author(s)

Marco Scutari

Examples

```
kcv = fairml.cv(response = vur.test$y, predictors = vur.test$X,
        sensitive = vur.test$S, unfairness = 0.10, model = "nclm",
        method = "k-fold", k = 10, runs = 10)
kcv
cv.loss(kcv)
cv.unfairness(kcv)

# run a second cross-validation with the same folds.
fairml.cv(response = vur.test$y, predictors = vur.test$X,
```

```
        sensitive = vur.test$S, unfairness = 0.10, model = "nclm",
        method = "custom-folds", folds = cv.folds(kcv))

# run cross-validation in parallel.
## Not run:
library(parallel)
cl = makeCluster(2)
fairml.cv(response = vur.test$y, predictors = vur.test$X,
  sensitive = vur.test$S, unfairness = 0.10, model = "nclm",
  method = "k-fold", k = 10, runs = 10, cluster = cl)
stopCluster(cl)

## End(Not run)
```

`fairness.profile.plot` *Profile Fair Models with Respect to Tuning Parameters*

Description

Visually explore various aspect of a model over the range of possible values of the tuning parameters that control its fairness.

Usage

```
fairness.profile.plot(response, predictors, sensitive, unfairness,
  legend = FALSE, type = "coefficients", model, model.args = list(), cluster)
```

Arguments

`response`	a numeric vector, the response variable.
`predictors`	a numeric matrix or a data frame containing numeric and factor columns; the predictors.
`sensitive`	a numeric matrix or a data frame containing numeric and factor columns; the sensitive attributes.
`unfairness`	a vector of positive numbers in [0, 1], how unfair is the model allowed to be. The default value is `seq(from = 0.00, to = 1, by = 0.02)`.
`legend`	a logical value, whether to add a legend to the plot.
`type`	a character string, either `"coefficients"` (the default) or `"constraints"`.
`model`	a character string, the label of the model. Currently `"nclm"`, `"frrm"`, `"fgrrm"`, `"zlm"` and `"zlrm"` are available.
`model.args`	additional arguments passed to the model.
`cluster`	an optional cluster object from package **parallel**, to fit models in parallel.

Details

`fairness.profile.plot()` fits the model for all the values of the argument `unfairness`, and produces a profile plot of the regression coefficients or the proportion of explained variance.

If `type = "coefficients"`, the coefficients of the model are plotted against the values of `unfairness`.

If `type = "constraints"`, the following quantities are plotted against the values of `unfairness`:

1. For model `"nclm"`, and model `"frrm"` with `definition = "sp-komiyama"`:
   (a) the proportion of variance explained by the sensitive attributes (with respect to the response);
   (b) the proportion of variance explained by the predictors (with respect to the response);
   (c) the proportion of variance explained by the sensitive attributes (with respect to the combined sensitive attributes and predictors).
2. For model `"frrm"` with `definition = "eo-komiyama"`:
   (a) the proportion of variance explained by the sensitive attributes (with respect to the fitted values);
   (b) the proportion of variance explained by the response (with respect to the fitted values);
   (c) the proportion of variance explained by the sensitive attributes (with respect to the combined sensitive attributes and response).
3. For model `"fgrrm"`: same as for `"frrm"` with the same `definition`.
4. For models `"zlm"` and `"zlrm"`: the correlations between the fitted values (from `fitted()` with `type = "link"`) and the sensitive attributes.

If `type = "precision-recall"` and the `model` is a classifier, the precision, recall and F1 measures are plotted against the values of `unfairness`.

If `type = "rmse"` and the `model` is a linear regression, the residuals mean square error are plotted against the values of `unfairness`.

Value

A `trellis` object containing a **lattice** plot.

Author(s)

Marco Scutari

Examples

```
data(vur.test)
fairness.profile.plot(response = vur.test$y, predictors = vur.test$X,
  sensitive = vur.test$S, type = "coefficients", model = "nclm", legend = TRUE)
fairness.profile.plot(response = vur.test$y, predictors = vur.test$X,
  sensitive = vur.test$S, type = "constraints", model = "nclm", legend = TRUE)
fairness.profile.plot(response = vur.test$y, predictors = vur.test$X,
  sensitive = vur.test$S, type = "rmse", model = "nclm", legend = TRUE)

# profile plots fitting models in parallel.
## Not run:
```

```
library(parallel)
cl = makeCluster(2)
fairness.profile.plot(response = vur.test$y, predictors = vur.test$X,
  sensitive = vur.test$S, model = "nclm", cluster = cl)
stopCluster(cl)

## End(Not run)
```

`frrm` *Fair Ridge Regression Model*

Description

A regression model enforcing fairness with a ridge penalty.

Usage

```
# a fair ridge regression model.
frrm(response, predictors, sensitive, unfairness,
  definition = "sp-komiyama", lambda = 0, save.auxiliary = FALSE)
# a fair generalized ridge regression model.
fgrrm(response, predictors, sensitive, unfairness,
  definition = "sp-komiyama", family = "binomial", lambda = 0,
  save.auxiliary = FALSE)
```

Arguments

`response`	a numeric vector, the response variable.
`predictors`	a numeric matrix or a data frame containing numeric and factor columns; the predictors.
`sensitive`	a numeric matrix or a data frame containing numeric and factor columns; the sensitive attributes.
`unfairness`	a positive number in [0, 1], how unfair is the model allowed to be. A value of `0` means the model is completely fair, while a value of `1` means the model is not constrained to be fair at all.
`definition`	a character string, the label of the definition of fairness used in fitting the model. Currently either `"sp-komiyama"` or `"eo-komiyama"`. See below for details.
`family`	a character string, either `"binomial"` to fit a logistic regression or `"gaussian"` to fit a linear regression.
`lambda`	a non-negative number, a ridge-regression penalty coefficient. It defaults to zero.
`save.auxiliary`	a logical value, whether to save the fitted values and the residuals of the auxiliary model that constructs the decorrelated predictors. The default value is `FALSE`.

Details

`frrm()` can accommodate different definitions of fairness, which can be selected via the `definition` argument.

- `"sp-komiyama"` uses the same definition of fairness as `nclm()`: the model bounds the proportion of the variance that is explained by the sensitive attributes over the total explained variance. This falls within the definition of statistical parity.
- `"eo-komiyama"` enforces equality of opportunity in a similar way: it regresses the fitted values against the sensitive attributes and the response, and it bounds the proportion of the variance explained by the sensitive attributes over the total explained variance in that model.

The algorithm works like this:

1. regresses the predictors against the sensitive attributes;
2. constructs a new set of predictors that are decorrelated from the sensitive attributes using the residuals of this regression;
3. regresses the response against the decorrelated predictors and the sensitive attributes; while
4. using a ridge penalty to control the proportion of variance the sensitive attributes can explain with respect to the overall explained variance of the model.

Both `sensitive` and `predictors` are standardized internally before estimating the regression coefficients, which are then rescaled back to match the original scales of the variables.

`fgrrm()` is the extension of `frrm()` to generalized linear models, currently implementing linear (`family = "gaussian"`) and logistic (`family = "binomial"`) regressions. `fgrrm()` is equivalent to `frrm()` with `family = "gaussian"`. The definition of fairness are identical between `frrm()` and `fgrrm()`.

Value

`frrm()` returns an object of class `c("frrm", "fair.model")`. `fgrrm()` returns an object of class `c("fgrrm", "fair.model")`.

Author(s)

Marco Scutari

See Also

nclm, zlm, zlrm

`german.credit` *German Credit Data*

Description

A credit scoring data set that can be used to predict defaults on consumer loans in the German market.

Usage

```
data(german.credit)
```

Format

The data contains 1000 observations (700 good loans, 300 bad loans) and the following variables:

- `Account_status`: a factor with four levels representing the amount of money in the account or `"no chcking account"`.
- `Duration`: a continuous variable, the duration in months.
- `Credit_history`: a factor with five levels representing possible credit history backgrounds.
- `Purpose`: a factor with ten levels representing possible reasons for taking out a loan.
- `Credit_amount`: a continuous variable.
- `Savings_bonds`: a factor with five levels representing amount of money available in savings and bonds or `"unknown / no savings account"`.
- `Present_employment_since`: a factor with five levels representing the length of tenure in the current employment or `"unemployed"`.
- `Installment_rate`: a continuous variable, the installment rate in percentage of disposable income.
- `Other_debtors_guarantors`: a factor with levels `"none"`, `"co-applicant"` and `"guarantor"`.
- `Resident_since`: a continuous variable, number of years in the current residence.
- `Property`: a factor with four levels describing the type of property to be bought or `"unknown / no property"`.
- `Age`: a continuous variable, the age in years.
- `Other_installment_plans`: a factor with levels `"bank"`, `"none"` and `"stores"`.
- `Housing`: a factor with levels `"rent"`, `"own"` and `"for free"`.
- `Existing_credits`: a continuous variable, the number of existing credit lines at this bank.
- `Job`: a factor with four levels for different job descriptions.
- `People_maintenance_for`: a continuous variable, the number of people being liable to provide maintenance for.
- `Telephone`: a factor with levels `"none"` and `"yes"`.
- `Foreign_worker`: a factor with levels `"no"` and `"yes"`.
- `Credit_risk`: a factor with levels `"BAD"` and `"GOOD"`.
- `Gender`: a factor with levels `"Male"` and `"Female"`.

Note

The variable "Personal status and sex" in the original data has been transformed into `Gender` by dropping the personal status information.

References

UCI Machine Learning Repository:
`https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`

`law.school.admissions` *Law School Admission Council data*

Description

Survey among students attending law school in the U.S. in 1991.

Usage

```
data(law.school.admissions)
```

Format

The data contains 20800 observations and the following variables:

- `age`, a continuous variable containing the student's age in years;
- `decile1`, a continuous variable containing the student's decile in the school given his grades in Year 1;
- `decile3`, a continuous variable containing the student's decile in the school given his grades in Year 3;
- `fam_inc`, a continuous variable containing student's family income bracket (from 1 to 5);
- `lsat`, a continuous variable containing the student's LSAT score;
- `ugpa`, a continuous variable containing the student's undergraduate GPA;
- `gender`, a factor with levels `"female"` and `"male"`;
- `race1`, a factor with levels `"asian"`, `"black"`, `"hisp"`, `"other"` and `"white"`;
- `cluster`, a factor with levels `"1"`, `"2"`, `"3"`, `"4"`, `"5"` and `"6"` encoding the tiers of law school prestige;
- `fulltime`, a factor with levels `"FALSE"` and `"TRUE"`, whether the student will work full-time or part-time;
- `bar`, a factor with levels `"FALSE"` and `"TRUE"`, whether the student passed the bar exam on the first try.

Note

The data set has been pre-processed as in Komiyama et al. (2018), with the following exceptions:

- `DOB_yr`, the year of birth, has been dropped because it is (nearly) perfectly collinear with `age`, and thus it is redundant;
- `decile1b` has been dropped because it is (nearly) perfectly collinear with `decile1`, and thus it is redundant.

In that paper, `ugpa` is the response variable, `age` and `race1` are the sensitive attributes and the remaining variables are used as predictors.

References

Sander RH (2004). "A Systemic Analysis of Affirmative Action in American Law Schools". Stanford Law Review, 57:367–483.

Examples

```
data(law.school.admissions)

# short-hand variable names.
ll = law.school.admissions
r = ll[, "ugpa"]
s = ll[, c("age", "race1")]
p = ll[, setdiff(names(ll), c("ugpa", "age", "race1"))]

m = nclm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)

m = frrm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)
```

---

methods for fair.model objects

Extract information from fair.model objects

Description

Extract various quantities of interest from an object of class `fair.model`.

Usage

```
# methods for all fair.model objects.
## S3 method for class 'fair.model'
coef(object, ...)
## S3 method for class 'fair.model'
residuals(object, ...)
## S3 method for class 'fair.model'
```

```
fitted(object, type = "response", ...)
## S3 method for class 'fair.model'
sigma(object, ...)
## S3 method for class 'fair.model'
deviance(object, ...)
## S3 method for class 'fair.model'
logLik(object, ...)
## S3 method for class 'fair.model'
nobs(object, ...)
## S3 method for class 'fair.model'
print(x, digits, ...)
## S3 method for class 'fair.model'
summary(object, ...)
## S3 method for class 'fair.model'
all.equal(target, current, ...)
## S3 method for class 'fair.model'
plot(x, diagonal = FALSE, regression = FALSE, ...)

# predict() methods.
## S3 method for class 'nclm'
predict(object, new.predictors, new.sensitive, type = "response", ...)
## S3 method for class 'zlm'
predict(object, new.predictors, type = "response", ...)
## S3 method for class 'zlrm'
predict(object, new.predictors, type = "response", ...)
## S3 method for class 'frrm'
predict(object, new.predictors, new.sensitive, type = "response", ...)
## S3 method for class 'fgrrm'
predict(object, new.predictors, new.sensitive, type = "response", ...)
```

Arguments

object,x,target,current	an object of class `fair.model` or `nclm`.
type	a character string, the type of fitted value. If `"response"`, `fitted()` and `predict()` will return the fitted values (if the response in the model is continuous) or the classification probabilities (if it was discrete). If `"class"` and `object` is a classifier, `fitted()` and `predict()` will return the class labels as a factor. If `"link"` and `object` is a classifier, `fitted()` and `predict()` will return the linear component of the fitted or predicted value, on the scale of the link function.
digits	a non-negative integer, the number of significant digits.
new.predictors	a numeric matrix or a data frame containing numeric and factor columns; the predictors for the new observations.
new.sensitive	a numeric matrix or a data frame containing numeric and factor columns; the sensitive attributes for the new observations.
diagonal	a logical value, whether to draw the diagonal of the first quadrant in `plot()`.
regression	a logical value, whether to draw the regression line of the observed values on the fitted values from the model in `plot()`.

... additional arguments, currently ignored.

`national.longitudinal.survey`
Income and Labour Market Activities

Description

Survey results from the U.S. Bureau of Labor Statistics to gather information on the labour market activities and other life events of several groups.

Usage

```
data(national.longitudinal.survey)
```

Format

The data contains 4908 observations and the following variables:

- `age`, a numeric variable containing the interviewee's age in years;
- `race`, a factor with 20 levels denoting various racial/ethnic origins;
- `gender`, a factor with levels `"Male"` and `"Female"`.
- `grade90`, a factor containing the highest completed school grade from "3RD GRADE" to "8TH YR COL OR MORE", with 18 levels;
- `income06`, a numeric variable, income in 2006 in 10000-USD units;
- `income96`, a numeric variable, income in 1996 in 10000-USD units;
- `income90`, a numeric variable, income in 1990 in 10000-USD units;
- `partner`, a factor encoding whether the interviewee has a partner, with levels `"No"` and `"Yes"`;
- `height`, a numeric variable, the height of the interviewee;
- `weight`, a numeric variable, the weight of the interviewee;
- `famsize`, a numeric variable, the number of family members;
- `genhealth`, a factor with levels `"Excellent"`, `"Very Good"`, `"Good"`, `"Fair"`, `"Poor"` encoding the general health status of the interviewee;
- `illegalact`, a numeric variable containing the number of illegal acts committed by the interviewee;
- `charged`, a numeric variable containing the number of illegal acts for which the interviewee has been charged;
- `jobsnum90`, a numeric value, the number of different jobs ever reported;
- `afqt89`, a numeric value, the percentile score of the "Profiles, Armed Forces Qualification Test" (AFQT);
- `typejob90`, a factor with 13 levels encoding different job types;
- `jobtrain90`, a factor with levels `"No"` and `"Yes"` encoding whether the job was classified as training.

Note

The data set has been pre-processed differently from Komiyama et al. (2018). In particular:

- the variables `income96` and `income06` have been retained as alternative responses;
- the variables `height`, `weight`, `race`, `partner` and `famsize` have been retained;
- the variables `grade90` and `genhealth` are coded as ordered factors because they do not make sense on a numeric scale.

In that paper, `income90` is the response variable, `gender` and `age` are the sensitive attributes.

References

U.S. Bureau of Labor Statistics.
`https://www.bls.gov/nls/`

Examples

```
data(national.longitudinal.survey)

# short-hand variable names.
nn = national.longitudinal.survey
# remove alternative response variables.
nn = nn[, setdiff(names(nn), c("income96", "income06"))]
# short-hand variable names.
r = nn[, "income90"]
s = nn[, c("gender", "age")]
p = nn[, setdiff(names(nn), c("income90", "gender", "age"))]

m = nclm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)

m = frrm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
summary(m)
```

---

nclm *Nonconvex Optimization for Regression with Fairness Constraints*

Description

Fair regression model based on nonconvex optimization from Komiyama et al. (2018).

Usage

```
nclm(response, predictors, sensitive, unfairness, covfun, lambda = 0,
  save.auxiliary = FALSE)
```

Arguments

response	a numeric vector, the response variable.
predictors	a numeric matrix or a data frame containing numeric and factor columns; the predictors.
sensitive	a numeric matrix or a data frame containing numeric and factor columns; the sensitive attributes.
unfairness	a positive number in [0, 1], how unfair is the model allowed to be. A value of 0 means the model is completely fair, while a value of 1 means the model is not constrained to be fair at all.
covfun	a function computing covariance matrices. It defaults to the cov() function from the **stats** package.
lambda	a non-negative number, a ridge-regression penalty coefficient. It defaults to zero.
save.auxiliary	a logical value, whether to save the fitted values and the residuals of the auxiliary model that constructs the decorrelated predictors. The default value is FALSE.

Details

nclm() defines fairness as statistical parity. The model bounds the proportion of the variance that is explained by the sensitive attributes over the total explained variance.

The algorithm proposed by Komiyama et al. (2018) works like this:

1. regresses the predictors against the sensitive attributes;
2. constructs a new set of predictors that are decorrelated from the sensitive attributes using the residuals of this regression;
3. regresses the response against the decorrelated predictors and the sensitive attributes, while
4. bounding the proportion of variance the sensitive attributes can explain with respect to the overall explained variance of the model.

Both sensitive and predictors are standardized internally before estimating the regression coefficients, which are then rescaled back to match the original scales of the variables. response is only standardized if it has a variance smaller than 1, as that seems to improve the stability of the solutions provided by the optimizer (as far as the data included in **fairml** are concerned).

The covfun argument makes it possible to specify a custom function to compute the covariance matrices used in the constrained optimization. Some examples are the kernel estimators described in Komiyama et al. (2018) and the shrinkage estimators in the **corpcor** package.

Value

nclm() returns an object of class c("nclm", "fair.model").

Author(s)

Marco Scutari

References

Komiyama J, Takeda A, Honda J, Shimao H (2018). "Nonconvex Optimization for Regression with Fairness Constraints". Proceedints of the 35th International Conference on Machine Learning (ICML), PMLR **80**:2737–2746.
http://proceedings.mlr.press/v80/komiyama18a/komiyama18a.pdf

See Also

frrm, zlm

`synthetic data sets` *Synthetic data sets to test fair models*

Description

Synthetic data sets used as test cases in the **fairml** package.

Usage

```
data(vur.test)
data(vuc.test)
```

Format

Each data set is a list with following three elements:

- `y`, the response variable;
- `X`, a numeric matrix containing 3 predictors called `X1`, `X2` and `X3`;
- `S`, a numeric matrix containing 3 sensitive attributes called `S1`, `S2` and `S3`.

Note

This data sets are called `vur.test` and `vuc.test` because they are generated from *v*ery *u*nfair *r*egression and *classification* models in which sensitive attributes explain the lion's share of the overall explained variance or deviance.

The code used to generate the predictors and the sensitive attributes is as follows.

```
library(mvtnorm)
sigma = matrix(0.3, nrow = 6, ncol = 6)
diag(sigma) = 1
n = 1000
X = rmvnorm(n, mean = rep(0, 6), sigma = sigma)
S = X[, 4:6]
X = X[, 1:3]
colnames(X) = c("X1", "X2", "X3")
colnames(S) = c("S1", "S2", "S3")
```

The continuous response in `vur.test` is produced as follows.

```
y = 2 + 2 * X[, 1] + 3 * X[, 2] + 4 * X[, 3] + 5 * X[, 4] +
        6 * X[, 5] + 7 * X[, 6] + rnorm(n, sd = 10)
```

The discrete response in `vuc.test` is produced as follows.

```
nu = 1 + 0.5 * X[, 1] + 0.6 * X[, 2] + 0.7 * X[, 3] + 0.8 * X[, 4] +
         0.9 * X[, 5] + 1.0 * X[, 6]
y = rbinom(n = nrow(X), size = 1, prob = exp(nu) / (1 + exp(nu)))
```

Examples

```
data(vur.test)
sensitive.attributes.model = lm(y ~ S, data = vur.test)
summary(sensitive.attributes.model)$r.squared
overall.model = lm(y ~ X + S, data = vur.test)
summary(overall.model)$r.squared

data(vuc.test)
sensitive.attributes.model = glm(y ~ S, data = vuc.test, family = "binomial")
deviance(sensitive.attributes.model)
overall.model = glm(y ~ X + S, data = vuc.test, family = "binomial")
deviance(overall.model)
```

---

zlm *Zafar's Linear and Logistic Regressions*

Description

Linear and logistic regression models enforcing fairness by bounding the covariance between sensitive attributes and predictors.

Usage

```
# a fair linear regression model.
zlm(response, predictors, sensitive, unfairness)
# a fair logistic regression model.
zlrm(response, predictors, sensitive, unfairness)
```

Arguments

`response`	a numeric vector, the response variable.
`predictors`	a numeric matrix or a data frame containing numeric and factor columns; the predictors.
`sensitive`	a numeric matrix or a data frame containing numeric and factor columns; the sensitive attributes.
`unfairness`	a positive number in [0, 1], how unfair is the model allowed to be. A value of `0` means the model is completely fair, while a value of `1` means the model is not constrained to be fair at all.

Details

`zlm()` and `zlrm()` define fairness as statistical parity.

Estimation minimizes the log-likelihood of the regression models under the constraint that the correlation between each sensitive attribute and the fitted values (on the linear predictor scale, in the case of logistic regression) is smaller than `unfairness` in absolute value. Both models include `predictors` as explanatory variables; the variables `sensitive` only appear in the constraints.

Value

`zlm()` returns an object of class `c("zlm", "fair.model")`. `zlrm()` returns an object of class `c("zlrm", "fair.model")`.

Author(s)

Marco Scutari

References

Zafar BJ, Valera I, Gomez-Rodriguez M, Gummadi KP (2019). "Fairness Constraints: a Flexible Approach for Fair Classification". Journal of Machine Learning Research, 30:1–42.
`https://www.jmlr.org/papers/volume20/18-262/18-262.pdf`

See Also

nclm, frrm, fgrrm

Index