Package ‘picR’

October 24, 2022

Title Predictive Information Criteria for Model Selection

Version 1.0.0

Description Computation of predictive information criteria (PIC) from select model object classes for model selection in predictive contexts. In contrast to the more widely used Akaike Information Criterion (AIC), which are derived under the assumption that target(s) of prediction (i.e. validation data) are independently and identically distributed to the fitting data, the PIC are derived under less restrictive assumptions and thus generalize AIC to the more practically relevant case of training/validation data heterogeneity. The methodology featured in this package is based on Flores (2021) <https://iro.uiowa.edu/esploro/outputs/doctoral/A-new-class-of-information-criteria/9984097169902771?institution=01IOWA_INST> ``A new class of information criteria for improved prediction in the presence of training/validation data heterogeneity".

License GPL (>= 3)

Encoding UTF-8

RoxygenNote 7.2.0

Imports stats

Suggests rmarkdown, knitr, testthat (>= 3.0.0), dplyr

Config/testthat/edition 3

URL https://github.com/javenrflo/picR

BugReports https://github.com/javenrflo/picR/issues

NeedsCompilation no

Author Javier Flores [aut, cre] (<https://orcid.org/0000-0002-1550-1655>)

Maintainer Javier Flores <javenrflo.pro@pm.me>

Repository CRAN

Date/Publication 2022-10-24 17:52:36 UTC

R topics documented:

PIC *Predictive Information Criteria*

Description

PIC is the S3 generic function for computing predictive information criteria (PIC). Depending on the class of the fitted model supplied to object, the function invokes the appropriate method for computing PIC.

Usage

```
PIC(object, newdata, ...)
```

Arguments

object	A fitted model object.
newdata	An optional dataframe to be used as validation data in computing PIC. If omitted, the training data contained within object are used.
...	Further arguments passed to other methods.

Details

The PIC are model selection criteria that may be used to select from among predictive models in a candidate set. The model with the minimum criterion value is preferred.

The PIC asymptotically select the candidate model that minimizes the mean squared error of prediction (MSEP), thus behaving similarly to the the Akaike Information Criterion (AIC). However in contrast to the AIC, the PIC do not assume a panel of validation data that are independent and identically distributed to the set of training data. This effectively enables the PIC to accommodate training/validation data *heterogeneity*, where training and validation data may differ from one another in distribution.

Data heterogeneity is arguably the more typical circumstance in practice, especially when one considers applications where a set of covariates are used to model and predict some response. In these regression contexts, one often predicts values of the response at combinations of covariate values not necessarily used in training the predictive model.

Value

The form of the value returned by PIC depends on the fitted model class and its method-specific arguments. Details may be found in the documentation of each method.

See Also

PIC.lm, PIC.mlm

Examples

```
data(iris)

# Fit a regression model
mod <- lm(Sepal.Length ~ Sepal.Width + Species, data = iris)
PIC(object  = mod,
    newdata = data.frame(Sepal.Width = c(0.25, 1.74, 2.99),
                         Species = factor(c("setosa", "virginica", "virginica"),
                                      levels = c("setosa", "versicolor", "virginica"))))

# Fit a bivariable regression model
mod <- lm(cbind(Sepal.Length, Sepal.Width) ~ Species + Petal.Length, data = iris)
# Note: For multivariable models, response variable columns must be included if
#       newdata is specified. If the values of the validation response(s) are
#       unknown, specify NA. If partially observed, specify NA only where unknown.
PIC(object  = mod,
    newdata = data.frame(Sepal.Length = c(4.1, NA, NA),
                         Sepal.Width  = c(NA,NA,3.2),
                         Petal.Length = c(1.2, 3.5, 7),
                         Species = factor(c("setosa", "virginica", "virginica"),
                                      levels = c("setosa", "versicolor", "virginica"))))
```

PIC.lm *PIC method for Linear Models*

Description

Computation of predictive information criteria for linear models.

Usage

```
## S3 method for class 'lm'
PIC(object, newdata, group_sizes = NULL, bootstraps = NULL, ...)
```

Arguments

`object` A fitted model object of `class` "lm".

`newdata` An optional dataframe to be used as validation data in computing PIC. If omitted, the training data contained within `object` are used.

`group_sizes` An optional scalar or numeric vector indicating the sizes of `newdata` partitions. If omitted, `newdata` is not partitioned. See ’Details’.

`bootstraps` An optional numeric value indicating the number of bootstrap samples to use for a bootstrapped PIC. See ’Details’.

`...` Further arguments passed to or from other methods.

Details

PIC.lm computes PIC values based on the supplied regression model. Candidate models with relatively smaller criterion values are preferred. Depending on the value(s) supplied to group_sizes, one of three implementations of PIC are computed:

- **iPIC**: The individualized predictive information criterion (iPIC) is computed when group_sizes = 1. A value of iPIC is determined for each *individual* observation in newdata. Using iPIC, one may thus select optimal predictive models specific to each particular validation datapoint.
- **gPIC**: The group predictive information criterion (gPIC) is computed when group_sizes > 1 or is.vector(group_sizes) == TRUE. A value of gPIC is determined for each cohort or *group* of observations defined by the partitions of newdata. Using gPIC, one may thus select optimal predictive models specific to each group of validation datapoints. For the class of regression models, the gPIC value of a group of validation observations is equivalent to the sum of their individual iPIC values.
- **tPIC**: The total predictive information criterion (tPIC) is computed when group_sizes = NULL. Computation of the tPIC is the default, and one may use the tPIC to select the optimal predictive model for the entire set of validation datapoints. The tPIC and gPIC are equivalent when group_sizes = m, where m is the number of observations in newdata. When newdata is not supplied, tPIC is exactly equivalent to the Akaike Information Criterion (AIC).

If a numeric value is supplied to bootstraps the total Predictive information criterion (tPIC) is computed bootstraps times, where generated bootstrap samples are each used as sets of validation data in computing the tPIC. The resulting tPIC values are then averaged to generate a single, bootstrapped tPIC value. Model selection based on this bootstrapped tPIC value may lead to the selection of a more generally applicable predictive model whose predictive accuracy is not strictly optimized to a particular set of validation data.

For further details, see *A new class of information criteria for improved prediction in the presence of training/validation data heterogeneity*.

Value

If group_sizes = NULL or bootstraps > 0, a scalar is returned. Otherwise, newdata is returned with an appended column labeled 'PIC' containing either iPIC or gPIC values, depending on the value provided to group_sizes.

References

Flores, J.E. (2021), *A new class of information criteria for improved prediction in the presence of training/validation data heterogeneity* [Unpublished PhD dissertation]. University of Iowa.

See Also

PIC, PIC.mlm, lm

Examples

```
data(iris)
```

```
# Fit a regression model
mod <- lm(Sepal.Length ~ ., data = iris)
class(mod)

# Hypothetical validation data
set.seed(1)
vdat <- iris[sample(1:nrow(iris), 10),]

# tPIC, newdata not supplied
PIC(object = mod)
AIC(mod) # equivalent to PIC since training and validation data are the same above

# tPIC, newdata supplied
PIC(object = mod, newdata = vdat)
AIC(mod) # not equivalent to PIC since training and validation data differ above

# gPIC
PIC(object = mod, newdata = vdat, group_sizes = c(5,3,2))
PIC(object = mod, newdata = vdat, group_sizes = 5)

# iPIC
PIC(object = mod, newdata = vdat, group_sizes = rep(1, 10))
PIC(object = mod, newdata = vdat, group_sizes = 1)

# bootstrapped tPIC (based on 10 bootstrap samples)
set.seed(1)
PIC(object = mod, bootstraps = 10)
```

PIC.mlm *PIC method for Multivariable Linear Models*

Description

Computation of predictive information criteria for multivariable linear models. Currently, computations are supported for only bivariable linear models.

Usage

```
## S3 method for class 'mlm'
PIC(object, newdata, group_sizes = NULL, bootstraps = NULL, ...)
```

Arguments

object	A fitted model object of class "mlm".
newdata	An optional dataframe to be used as validation data in computing PIC. If omitted, the training data contained within object are used. If specified, newdata must contain columns for each model response. See ’Details’.

`group_sizes` An optional scalar or numeric vector indicating the sizes of `newdata` partitions. If omitted, `newdata` is not partitioned. See 'Details'.

`bootstraps` An optional numeric value indicating the number of bootstrap samples to use for a bootstrapped PIC. See 'Details'.

`...` Further arguments passed to or from other methods.

Details

`PIC.mlm` computes PIC values based on the supplied multivariable regression model. Candidate models with relatively smaller criterion values are preferred. Depending on the value(s) supplied to `group_sizes`, one of three implementations of PIC are computed:

- **iPIC**: The individualized predictive information criterion (iPIC) is computed when `group_sizes = 1`. A value of iPIC is determined for each *individual* observation in `newdata`. Using iPIC, one may thus select optimal predictive models specific to each particular validation datapoint.
- **gPIC**: The group predictive information criterion (gPIC) is computed when `group_sizes > 1` or `is.vector(group_sizes) == TRUE`. A value of gPIC is determined for each cohort or *group* of observations defined by the partitions of `newdata`. Using gPIC, one may thus select optimal predictive models specific to each group of validation datapoints. For the class of regression models, the gPIC value of a group of validation observations is equivalent to the sum of their individual iPIC values.
- **tPIC**: The total predictive information criterion (tPIC) is computed when `group_sizes = NULL`. Computation of the tPIC is the default, and one may use the tPIC to select the optimal predictive model for the entire set of validation datapoints. The tPIC and gPIC are equivalent when `group_sizes = m`, where `m` is the number of observations in `newdata`. When `newdata` is not supplied, tPIC is exactly equivalent to the Akaike Information Criterion (AIC).

Distinct from the computation for the class of "lm" models (PIC.lm), the PIC computation for multivariable regression models differs depending on the whether validation data are partially or completely unobserved. If partially unobserved, where only some values of the multivariable response vector are unknown/unobserved, any remaining observed values are used in the PIC computation.

If a numeric value is supplied to `bootstraps` the total Predictive information criterion (tPIC) is computed `bootstraps` times, where generated bootstrap samples are each used as sets of validation data in computing the tPIC. It is assumed that the multivariable response vectors are each completely unobserved. The resulting tPIC values are then averaged to generate a single, bootstrapped tPIC value. Model selection based on this bootstrapped tPIC value may lead to the selection of a more generally applicable predictive model whose predictive accuracy is not strictly optimized to a particular set of validation data.

For further details, see *A new class of information criteria for improved prediction in the presence of training/validation data heterogeneity*.

Value

If `group_sizes = NULL` or `bootstraps > 0`, a scalar is returned. Otherwise, `newdata` is returned with an appended column labeled 'PIC' containing either iPIC or gPIC values, depending on the value provided to `group_sizes`.

References

Flores, J.E. (2021), *A new class of information criteria for improved prediction in the presence of training/validation data heterogeneity* [Unpublished PhD dissertation]. University of Iowa.

See Also

PIC, PIC.lm, lm

Examples

```
require(dplyr, quietly = TRUE)
data(iris)

# Fit a bivariable regression model
mod <- lm(cbind(Sepal.Length, Sepal.Width) ~ ., data = iris)
class(mod)

# Hypothetical validation data
set.seed(1)
vdat <- iris[sample(1:nrow(iris), 10),]

# tPIC, completely unobserved response data
PIC(object = mod, newdata = vdat %>% dplyr::mutate(Sepal.Length = NA, Sepal.Width = NA))

# tPIC, partially unobserved response data
PIC(object = mod, newdata = vdat %>% dplyr::mutate(Sepal.Length = NA))

# tPIC, mix of completely and partially unobserved cases.
PIC(object = mod, newdata = vdat %>%
dplyr::mutate(Sepal.Length = ifelse(Sepal.Length < 6, NA, Sepal.Length),
Sepal.Width = ifelse(Sepal.Width < 3.3, NA, Sepal.Width)))

# tPIC, newdata not supplied
PIC(object = mod)

# gPIC
PIC(object = mod, newdata = vdat, group_sizes = c(5,3,2))
PIC(object = mod, newdata = vdat, group_sizes = 5)

# iPIC
PIC(object = mod, newdata = vdat, group_sizes = rep(1, 10))
PIC(object = mod, newdata = vdat, group_sizes = 1)

# bootstrapped tPIC (based on 10 bootstrap samples)
set.seed(1)
PIC(object = mod, bootstraps = 10)
```

Index