Package ‘irboost’

April 18, 2024

Type Package

Title Iteratively Reweighted Boosting for Robust Analysis

Version 0.1-1.5

Date 2024-04-18

Author Zhu Wang [aut, cre] (<https://orcid.org/0000-0002-0773-0052>)

Maintainer Zhu Wang <zhuwang@gmail.com>

Description Fit a predictive model using iteratively reweighted boosting (IRBoost) to minimize robust loss functions within the CC-family (concave-convex). This constitutes an application of iteratively reweighted convex optimization (IRCO), where convex optimization is performed using the functional descent boosting algorithm. IRBoost assigns weights to facilitate outlier identification. Applications include robust generalized linear models and robust accelerated failure time models. Wang (2021) <doi:10.48550/arXiv.2101.07718>.

Depends R (>= 3.5.0)

Imports mpath (>= 0.4-2.21), xgboost

Suggests R.rsp, DiagrammeR, survival, Hmisc

VignetteBuilder R.rsp

License GPL (>= 3)

Encoding UTF-8

LazyLoad yes

RoxygenNote 7.3.1

NeedsCompilation no

Repository CRAN

Date/Publication 2024-04-18 17:52:58 UTC

R topics documented:

`dataLS` *generate random data for classification as in Long and Servedio (2010)*

Description

generate random data for classification as in Long and Servedio (2010)

Usage

```
dataLS(ntr, ntu = ntr, nte, percon)
```

Arguments

`ntr`	number of training data
`ntu`	number of tuning data, default is the same as `ntr`
`nte`	number of test data
`percon`	proportion of contamination, must between 0 and 1. If `percon` > `0`, the labels of the corresponding percenrage of response variable in the training and tuning data are flipped.

Value

a list with elements xtr, xtu, xte, ytr, ytu, yte for predictors of disjoint training, tuning and test data, and response variable -1/1 of training, tuning and test data.

Author(s)

Zhu Wang
Maintainer: Zhu Wang <`zhuwang@gmail.com`>

References

P. Long and R. Servedio (2010), *Random classification noise defeats all convex potential boosters*, *Machine Learning Journal*, 78(3), 287–304.

Examples

```
dat <- dataLS(ntr=100, nte=100, percon=0)
```

irb.train — *fit a robust predictive model with iteratively reweighted boosting algorithm*

Description

Fit a predictive model with the iteratively reweighted convex optimization (IRCO) that minimizes the robust loss functions in the CC-family (concave-convex). The convex optimization is conducted by functional descent boosting algorithm in the R package **xgboost**. The iteratively reweighted boosting (IRBoost) algorithm reduces the weight of the observation that leads to a large loss; it also provides weights to help identify outliers. Applications include the robust generalized linear models and extensions, where the mean is related to the predictors by boosting, and robust accelerated failure time models. `irb.train` is an advanced interface for training an irboost model. The `irboost` function is a simpler wrapper for `irb.train`. See `xgboost::xgb.train`.

Usage

```
irb.train(
  params = list(),
  data,
  z_init = NULL,
  cfun = "ccave",
  s = 1,
  delta = 0.1,
  iter = 10,
  nrounds = 100,
  del = 1e-10,
  trace = FALSE,
  ...
)
```

Arguments

`params` the list of parameters, `params` is passed to function xgboost::xgb.train which requires the same argument. The list must include `objective`, a convex component in the CC-family, the second C, or convex down. It is the same as `objective` in the `xgboost::xgb.train`. The following objective functions are currently implemented:

- `reg:squarederror` Regression with squared loss.
- `binary:logitraw` logistic regression for binary classification, predict linear predictor, not probabilies.
- `binary:hinge` hinge loss for binary classification. This makes predictions of -1 or 1, rather than producing probabilities.
- `multi:softprob` softmax loss function for multiclass problems. The result contains predicted probabilities of each data point in each class, say p_k, k=0, ..., nclass-1. Note, `label` is coded as in [0, ..., nclass-1]. The loss

function cross-entropy for the i-th observation is computed as -log(p_k) with k=lable_i, i=1, ..., n.

- `count:poisson`: Poisson regression for count data, predict mean of poisson distribution.
- `reg:gamma`: gamma regression with log-link, predict mean of gamma distribution. The implementation in `xgboost::xgb.train` takes a parameterization in the exponential family:
  xgboost/src/src/metric/elementwise_metric.cu.
  In particularly, there is only one parameter psi and set to 1. The implementation of the IRCO algorithm follows this parameterization. See Table 2.1, McCullagh and Nelder, Generalized linear models, Chapman & Hall, 1989, second edition.
- `reg:tweedie`: Tweedie regression with log-link. See also `tweedie_variance_power` in range: (1,2). A value close to 2 is like a gamma distribution. A value close to 1 is like a Poisson distribution.
- `survival:aft`: Accelerated failure time model for censored survival time data. `irb.train` invokes `irb.train_aft`.

`data` training dataset. `irb.train` accepts only an `xgboost::xgb.DMatrix` as the input. `irboost`, in addition, also accepts `matrix`, `dgCMatrix`, or name of a local data file. See `xgboost::xgb.train`.

`z_init` vector of nobs with initial convex component values, must be non-negative with default values = weights if data has provided, otherwise z_init = vector of 1s

`cfun` concave component of CC-family, can be `"hacve"`, `"acave"`, `"bcave"`, `"ccave"`, `"dcave"`, `"ecave"`, `"gcave"`, `"hcave"`.
See Table 2 https://arxiv.org/pdf/2010.02848.pdf

`s` tuning parameter of `cfun`. `s > 0` and can be equal to 0 for `cfun="tcave"`. If `s` is too close to 0 for `cfun="acave", "bcave", "ccave"`, the calculated weights can become 0 for all observations, thus crash the program

`delta` a small positive number provided by user only if `cfun="gcave"` and `0 < s <1`

`iter` number of iteration in the IRCO algorithm

`nrounds` boosting iterations within each IRCO iteration

`del` convergency criteria in the IRCO algorithm, no relation to `delta`

`trace` if `TRUE`, fitting progress is reported

`...` other arguments passing to `xgb.train`

Value

An object with S3 class `xgb.train` with the additional elments:

- `weight_update_log` a matrix of `nobs` row by `iter` column of observation weights in each iteration of the IRCO algorithm
- `weight_update` a vector of observation weights in the last IRCO iteration that produces the final model fit

- loss_log sum of loss value of the composite function in each IRCO iteration. Note, cfun requires objective non-negative in some cases. Thus care must be taken. For instance, with objective="reg:gamma", the loss value is defined by gamma-nloglik - (1+log(min(y))), where y=label. The second term is introduced such that the loss value is non-negative. In fact, gamma-nloglik=y/ypre + log(ypre) in the xgboost::xgb.train, where ypre is the mean prediction value, can be negative. It can be derived that for fixed y, the minimum value of gamma-nloglik is achived at ypre=y, or 1+log(y). Thus, among all label values, the minimum of gamma-nloglik is 1+log(min(y)).

Author(s)

Zhu Wang
Maintainer: Zhu Wang <zhuwang@gmail.com>

References

Wang, Zhu (2021), *Unified Robust Boosting*, arXiv eprint, https://arxiv.org/abs/2101.07718

Examples

```
# logistic boosting
data(agaricus.train, package='xgboost')
data(agaricus.test, package='xgboost')

dtrain <- with(agaricus.train, xgboost::xgb.DMatrix(data, label = label))
dtest <- with(agaricus.test, xgboost::xgb.DMatrix(data, label = label))
watchlist <- list(train = dtrain, eval = dtest)

# A simple irb.train example:
param <- list(max_depth = 2, eta = 1, nthread = 2,
objective = "binary:logitraw", eval_metric = "auc")
bst <- xgboost::xgb.train(params=param, data=dtrain, nrounds = 2,
                          watchlist=watchlist, verbose=2)
bst <- irb.train(params=param, data=dtrain, nrounds = 2)
summary(bst$weight_update)
# a bug in xgboost::xgb.train
#bst <- irb.train(params=param, data=dtrain, nrounds = 2,
#                 watchlist=watchlist, trace=TRUE, verbose=2)

# time-to-event analysis
X <- matrix(1:5, ncol=1)
# Associate ranged labels with the data matrix.
# This example shows each kind of censored labels.
# uncensored  right  left  interval
y_lower = c(10,  15, -Inf, 30, 100)
y_upper = c(Inf, Inf,   20, 50, Inf)
dtrain <- xgboost::xgb.DMatrix(data=X, label_lower_bound=y_lower,
                               label_upper_bound=y_upper)
param <- list(objective="survival:aft", aft_loss_distribution="normal",
              aft_loss_distribution_scale=1, max_depth=3, min_child_weight=0)
watchlist <- list(train = dtrain)
```

```
bst <- xgboost::xgb.train(params=param, data=dtrain, nrounds=15,
                          watchlist=watchlist)
predict(bst, dtrain)
bst_cc <- irb.train(params=param, data=dtrain, nrounds=15, cfun="hcave",
                    s=1.5, trace=TRUE, verbose=0)
bst_cc$weight_update
```

irb.train_aft — *fit a robust accelerated failure time model with iteratively reweighted boosting algorithm*

Description

Fit an accelerated failure time model with the iteratively reweighted convex optimization (IRCO) that minimizes the robust loss functions in the CC-family (concave-convex). The convex optimization is conducted by functional descent boosting algorithm in the R package **xgboost**. The iteratively reweighted boosting (IRBoost) algorithm reduces the weight of the observation that leads to a large loss; it also provides weights to help identify outliers. For time-to-event data, an accelerated failure time model (AFT model) provides an alternative to the commonly used proportional hazards models. Note, function `irboost_aft` was developed to facilitate a data input format used with function `xgb.train` for `objective=survival:aft` in package `xgboost`. In other ojective functions, the input format is different with function `xgboost` at the time.

Usage

```
irb.train_aft(
  params = list(),
  data,
  z_init = NULL,
  cfun = "ccave",
  s = 1,
  delta = 0.1,
  iter = 10,
  nrounds = 100,
  del = 1e-10,
  trace = FALSE,
  ...
)
```

Arguments

`params` the list of parameters used in `xgb.train` of **xgboost**.
Must include `aft_loss_distribution`, `aft_loss_distribution_scale`, but there is no need to include `objective`. The complete list of parameters is available in the online documentation.

`data` training dataset. `irboost_aft` accepts only an `xgb.DMatrix` as the input.

z_init vector of nobs with initial convex component values, must be non-negative with default values = weights if provided, otherwise z_init = vector of 1s

cfun concave component of CC-family, can be "hacve", "acave", "bcave", "ccave", "dcave", "ecave", "gcave", "hcave".
See Table 2 at https://arxiv.org/pdf/2010.02848.pdf

s tuning parameter of cfun. s > 0 and can be equal to 0 for cfun="tcave". If s is too close to 0 for cfun="acave", "bcave", "ccave", the calculated weights can become 0 for all observations, thus crash the program

delta a small positive number provided by user only if cfun="gcave" and 0 < s <1

iter number of iteration in the IRCO algorithm

nrounds boosting iterations in xgb.train within each IRCO iteration

del convergency criteria in the IRCO algorithm, no relation to delta

trace if TRUE, fitting progress is reported

... other arguments passing to xgb.train

Value

An object of class xgb.Booster with additional elements:

- weight_update_log a matrix of nobs row by iter column of observation weights in each iteration of the IRCO algorithm
- weight_update a vector of observation weights in the last IRCO iteration that produces the final model fit
- loss_log sum of loss value of the composite function cfun(survival_aft_distribution) in each IRCO iteration

Author(s)

Zhu Wang
Maintainer: Zhu Wang <zhuwang@gmail.com>

References

Wang, Zhu (2021), *Unified Robust Boosting*, arXiv eprint, https://arxiv.org/abs/2101.07718

See Also

irboost

Examples

```
library("xgboost")
X <- matrix(1:5, ncol=1)

# Associate ranged labels with the data matrix.
# This example shows each kind of censored labels.
#          uncensored  right  left  interval
y_lower = c(10,  15, -Inf, 30, 100)
y_upper = c(Inf, Inf,   20, 50, Inf)
dtrain <- xgb.DMatrix(data=X, label_lower_bound=y_lower, label_upper_bound=y_upper)
                 params = list(objective="survival:aft", aft_loss_distribution="normal",
                      aft_loss_distribution_scale=1, max_depth=3, min_child_weight= 0)
watchlist <- list(train = dtrain)
bst <- xgb.train(params, data=dtrain, nrounds=15, watchlist=watchlist)
predict(bst, dtrain)
bst_cc <- irb.train_aft(params, data=dtrain, nrounds=15, watchlist=watchlist, cfun="hcave",
                       s=1.5, trace=TRUE, verbose=0)
bst_cc$weight_update
predict(bst_cc, dtrain)
```

irboost — *fit a robust predictive model with iteratively reweighted boosting algorithm*

Description

Fit a predictive model with the iteratively reweighted convex optimization (IRCO) that minimizes the robust loss functions in the CC-family (concave-convex). The convex optimization is conducted by functional descent boosting algorithm in the R package **xgboost**. The iteratively reweighted boosting (IRBoost) algorithm reduces the weight of the observation that leads to a large loss; it also provides weights to help identify outliers. Applications include the robust generalized linear models and extensions, where the mean is related to the predictors by boosting, and robust accelerated failure time models.

Usage

```
irboost(
  data,
  label,
  weights,
  params = list(),
  z_init = NULL,
  cfun = "ccave",
  s = 1,
  delta = 0.1,
  iter = 10,
  nrounds = 100,
  del = 1e-10,
  trace = FALSE,
  ...
)
```

Arguments

`data`
: input data, if `objective="survival:aft"`, it must be an `xgb.DMatrix`; otherwise, it can be a matrix of dimension nobs x nvars; each row is an observation vector. Can accept `dgCMatrix`

`label`
: response variable. Quantitative for `objective="reg:squarederror"`, `objective="count:poisson"` (non-negative counts) or `objective="reg:gamma"` (positive). For `objective="binary:logitraw"` or `"binary:hinge"`, `label` should be a factor with two levels

`weights`
: vector of nobs with non-negative weights

`params`
: the list of parameters, `params` is passed to function xgboost::xgboost which requires the same argument. The list must include `objective`, a convex component in the CC-family, the second C, or convex down. It is the same as `objective` in the `xgboost::xgboost`. The following objective functions are currently implemented:

- `reg:squarederror` Regression with squared loss.
- `binary:logitraw` logistic regression for binary classification, predict linear predictor, not probabilies.
- `binary:hinge` hinge loss for binary classification. This makes predictions of -1 or 1, rather than producing probabilities.
- `multi:softprob` softmax loss function for multiclass problems. The result contains predicted probabilities of each data point in each class, say p_k, k=0, ..., nclass-1. Note, `label` is coded as in [0, ..., nclass-1]. The loss function cross-entropy for the i-th observation is computed as -log(p_k) with k=lable_i, i=1, ..., n.
- `count:poisson`: Poisson regression for count data, predict mean of poisson distribution.
- `reg:gamma`: gamma regression with log-link, predict mean of gamma distribution. The implementation in `xgboost` takes a parameterization in the exponential family:
  xgboost/src/src/metric/elementwise_metric.cu.
  In particularly, there is only one parameter psi and set to 1. The implementation of the IRCO algorithm follows this parameterization. See Table 2.1, McCullagh and Nelder, Generalized linear models, Chapman & Hall, 1989, second edition.
- `reg:tweedie`: Tweedie regression with log-link. See also `tweedie_variance_power` in range: (1,2). A value close to 2 is like a gamma distribution. A value close to 1 is like a Poisson distribution.
- `survival:aft`: Accelerated failure time model for censored survival time data. `irboost` invokes `irb.train_aft`.

`z_init`
: vector of nobs with initial convex component values, must be non-negative with default values = weights if provided, otherwise z_init = vector of 1s

`cfun`
: concave component of CC-family, can be `"hacve"`, `"acave"`, `"bcave"`, `"ccave"`, `"dcave"`, `"ecave"`, `"gcave"`, `"hcave"`.
See Table 2 at https://arxiv.org/pdf/2010.02848.pdf

s — tuning parameter of cfun. s > 0 and can be equal to 0 for cfun="tcave". If s is too close to 0 for cfun="acave", "bcave", "ccave", the calculated weights can become 0 for all observations, thus crash the program

delta — a small positive number provided by user only if cfun="gcave" and 0 < s <1

iter — number of iteration in the IRCO algorithm

nrounds — boosting iterations within each IRCO iteration

del — convergency criteria in the IRCO algorithm, no relation to delta

trace — if TRUE, fitting progress is reported

... — other arguments passing to xgboost

Value

An object with S3 class xgboost with the additional elments:

- weight_update_log a matrix of nobs row by iter column of observation weights in each iteration of the IRCO algorithm
- weight_update a vector of observation weights in the last IRCO iteration that produces the final model fit
- loss_log sum of loss value of the composite function in each IRCO iteration. Note, cfun requires objective non-negative in some cases. Thus care must be taken. For instance, with objective="reg:gamma", the loss value is defined by gamma-nloglik - (1+log(min(y))), where y=label. The second term is introduced such that the loss value is non-negative. In fact, gamma-nloglik=y/ypre + log(ypre) in the xgboost, where ypre is the mean prediction value, can be negative. It can be derived that for fixed y, the minimum value of gamma-nloglik is achived at ypre=y, or 1+log(y). Thus, among all label values, the minimum of gamma-nloglik is 1+log(min(y)).

Author(s)

Zhu Wang
Maintainer: Zhu Wang <zhuwang@gmail.com>

References

Wang, Zhu (2021), *Unified Robust Boosting*, arXiv eprint, https://arxiv.org/abs/2101.07718

Examples

```
# regression, logistic regression, Poisson regression
x <- matrix(rnorm(100*2),100,2)
g2 <- sample(c(0,1),100,replace=TRUE)
fit1 <- irboost(data=x, label=g2, cfun="acave",s=0.5,
                params=list(objective="reg:squarederror", max_depth=1), trace=TRUE,
                verbose=0, nrounds=50)
fit2 <- irboost(data=x, label=g2, cfun="acave",s=0.5,
                params=list(objective="binary:logitraw", max_depth=1), trace=TRUE,
                verbose=0, nrounds=50)
```

```
fit3 <- irboost(data=x, label=g2, cfun="acave",s=0.5,
                params=list(objective="binary:hinge", max_depth=1), trace=TRUE,
                verbose=0, nrounds=50)
fit4 <- irboost(data=x, label=g2, cfun="acave",s=0.5,
                params=list(objective="count:poisson", max_depth=1), trace=TRUE,
                verbose=0, nrounds=50)

# Gamma regression
x <- matrix(rnorm(100*2),100,2)
g2 <- sample(rgamma(100, 1))
library("xgboost")
param <- list(objective="reg:gamma", max_depth=1)
fit5 <- xgboost(data=x, label=g2, params=param, nrounds=50)
fit6 <- irboost(data=x, label=g2, cfun="acave",s=5, params=param, trace=TRUE,
                verbose=0, nrounds=50)
plot(predict(fit5, newdata=x), predict(fit6, newdata=x))
hist(fit6$weight_update)
plot(fit6$loss_log)
summary(fit6$weight_update)

# Tweedie regression
param <- list(objective="reg:tweedie", max_depth=1)
fit6t <- irboost(data=x, label=g2, cfun="acave",s=5, params=param,
                 trace=TRUE, verbose=0, nrounds=50)
# Gamma vs Tweedie regression
hist(fit6$weight_update)
hist(fit6t$weight_update)
plot(predict(fit6, newdata=x), predict(fit6t, newdata=x))

# multiclass classification in iris dataset:
lb <- as.numeric(iris$Species)-1
num_class <- 3
set.seed(11)

param <- list(objective="multi:softprob", max_depth=4, eta=0.5, nthread=2,
subsample=0.5, num_class=num_class)
fit7 <- irboost(data=as.matrix(iris[, -5]), label=lb, cfun="acave", s=50,
                params=param, trace=TRUE, verbose=0, nrounds=10)
# predict for softmax returns num_class probability numbers per case:
pred7 <- predict(fit7, newdata=as.matrix(iris[, -5]))
# reshape it to a num_class-columns matrix
pred7 <- matrix(pred7, ncol=num_class, byrow=TRUE)
# convert the probabilities to softmax labels
pred7_labels <- max.col(pred7) - 1
# classification error: 0!
sum(pred7_labels != lb)/length(lb)
table(lb, pred7_labels)
hist(fit7$weight_update)
```

Index