Package 'distrom'

March 29, 2022

Title Distributed Multinomial Regression

Version 1.0.1

Depends R (>= 2.15), Matrix, gamlr, parallel, methods, stats

Suggests MASS, textir

Description Fast distributed/parallel estimation for multinomial logistic regression via Poisson factorization and the 'gamlr' package. For details see: Taddy (2015, AoAS), Distributed Multinomial Regression, <arXiv:1311.6139>.

Maintainer Nelson Rayl <nelsonrayl14@gmail.com>

License GPL-3

URL https://github.com/TaddyLab/distrom

NeedsCompilation no

Author Matt Taddy [aut],
Nelson Rayl [cre]

Repository CRAN

Date/Publication 2022-03-29 00:10:08 UTC

R topics documented:

`collapse` *Data checking and binning*

Description

Collapses counts along equal levels of binned covariates.

Usage

```
collapse(v,counts,mu=NULL,bins=NULL)
```

Arguments

v	Either matrix or Matrix of covariates (matches covars in dmr).
counts	Either matrix or Matrix of multinomial counts, or a factor (matches counts in dmr).
mu	Possible pre-specified fixed effects for dmr; otherwise they are calculated here.
bins	The number of quantile bins into which we collapse v. bins=NULL does no collapsing.

Details

For each column of v, aggregates the observations into bins defined by their average value. Both v and counts are then collapsed according to levels of the interaction across implied bin-factors, and the number of observations in each bin is recorded as n. Look at the code of the dmr function to see collapse used in practice.

Value

A list containing collapsed and formatted v, counts, and nbin, along with mu = log(rowSums(counts)), the plug-in fixed effect estimates for dmr.

Author(s)

Matt Taddy <mataddy@gmail.com>

See Also

we8there

---

dmr *Distributed Multinomial Regression*

Description

Gamma-lasso path estimation for a multinomial logistic regression factorized into independent Poisson log regressions.

Usage

```
dmr(cl, covars, counts, mu=NULL, bins=NULL, verb=0, cv=FALSE, ...)
## S3 method for class 'dmr'
coef(object, ...)
## S3 method for class 'dmr'
predict(object, newdata,
type=c("link","response","class"), ...)
```

Arguments

cl	A parallel library socket cluster. If is.null(cl), everything is done in serial. See help(parallel), help(makeCluster), and our examples here for details.
covars	A dense matrix or sparse Matrix of covariates. This should not include the intercept.
counts	A dense matrix or sparse Matrix of response counts.
mu	Pre-specified fixed effects for each observation in the Poisson regression linear equation. If mu=NULL, then we use log(rowSums(x)). Note that if bins is non-null then this argument is ignored and mu is recalculated on the collapsed data.
bins	Number of bins into which we will attempt to collapse each column of covars. Since sums of multinomials with equal probabilities are also multinomial, the model is then fit to these collapsed 'observations'. bins=NULL does no collapsing.
verb	Whether to print some info. max(0,verb-1) is passed on to gamlr and will print if you created an outfile when specifying cl.
cv	A flag for whether to use cv.gamlr instead of gamlr for each Poisson regression.
type	For predict.dmr, this is the scale upon which you want prediction. Under "link", just the linear map newdata times object, under "response" the fitted multinomial probabilities, under "class" the max-probability class label. For sufficient reductions see the srproj function of the textir library.
newdata	A Matrix with the same number of columns as covars.
...	Additional arguments to gamlr, cv.gamlr, and their associated methods.
object	A dmr list of fitted gamlr models for each response category.

Details

dmr fits multinomial logistic regression by assuming that, unconditionally on the 'size' (total count across categories) each individual category count has been generated as a Poisson

$$x_{ij} \sim Po(exp[\mu_i + \alpha_j + \beta v_i]).$$

We [default] plug-in estimate $\hat{\mu}_i = log(m_i)$, where $m_i = \sum_j x_{ij}$ and p is the dimension of x_i. Then each individual is outsourced to Poisson regression in the gamlr package via the parLapply function of the parallel library. The output from dmr is a list of gamlr fitted models.

coef.dmr builds a matrix of multinomial logistic regression coefficients from the length(object) list of gamlr fits. Default selection under cv=FALSE uses an information criteria via AICc on Poisson deviance for each individual response dimension (see gamlr). Combined coefficients across all dimensions are then returned as a dmrcoef s4-class object.

predict.dmr takes either a dmr or dmrcoef object and returns predicted values for newdata on the scale defined by the type argument.

Value

dmr returns the dmr s3 object: an ncol(counts)-length list of fitted gamlr objects, with the added attributes nlambda, mu, and nobs.

Author(s)

Matt Taddy <mataddy@gmail.com>

References

Taddy (2015 AoAS) Distributed Multinomial Regression

Taddy (2017 JCGS) One-step Estimator Paths for Concave Regularization, the Journal of Computational and Graphical Statistics

Taddy (2013 JASA) Multinomial Inverse Regression for Text Analysis

See Also

dmrcoef-class, cv.dmr, AICc, and the gamlr and textir packages.

Examples

```
library(MASS)
data(fgl)

## make your cluster
## FORK is faster but memory heavy, and doesn't work on windows.
cl <- makeCluster(2,type=ifelse(.Platform$OS.type=="unix","FORK","PSOCK"))
print(cl)

## fit in parallel
fits <- dmr(cl, fgl[,1:9], fgl$type, verb=1)

## its good practice stop the cluster once you're done
stopCluster(cl)

## Individual Poisson model fits and AICc selection
par(mfrow=c(3,2))
for(j in 1:6){
plot(fits[[j]])
mtext(names(fits)[j],font=2,line=2) }

##  AICc model selection
B <- coef(fits)

## Fitted probability by true response
par(mfrow=c(1,1))
P <- predict(B, fgl[,1:9], type="response")
boxplot(P[cbind(1:214,fgl$type)]~fgl$type,
ylab="fitted prob of true class")
```

dmrcoef-class *Class* "dmrcoef"

Description

The extended dgCMatrix class for output from coef.dmr.

Details

This is the class for a covariate matrix from dmr regression; it inherits the dgCMatrix class as defined in the Matrix library. In particular, this is the ncol(covars) by ncol(counts) matrix of logistic regression coefficients chosen in coef.dmr from the regularization paths for each category.

Objects from the Class

Objects can be created only by a call to the coef.dmr function.

Slots

i: From dgCMatrix: the row indices.

p: From dgCMatrix: the column pointers.

Dim: From dgCMatrix: the dimensions.

Dimnames: From dgCMatrix: the list of labels.

x: From dgCMatrix: the nonzero entries.

factors: From dgCMatrix.

Extends

Class dgCMatrix, directly.

Methods

predict signature(object = "dmrcoef"): Prediction for a given dmrcoef matrix. Takes the same arguments as predict.dmr, but will be faster (since coef.dmr is called inside predict.dmr).

Author(s)

Matt Taddy <mataddy@gmail.com>

See Also

dmr, coef.dmr, predict.dmr

Examples

```
showClass("dmrcoef")
```

Index