Package 'maxent'

July 2, 2014

Type Package

Title Low-memory Multinomial Logistic Regression with Support for Text Classification

Version 1.3.3.1

Date 2013-04-06

Author Timothy P. Jurka, Yoshimasa Tsuruoka

Maintainer Timothy P. Jurka <tpjurka@ucdavis.edu>

Depends R (>= 2.13.0), methods, SparseM, tm

Imports Rcpp

LinkingTo Rcpp

Description maxent is an R package with tools for low-memory
multinomial logistic regression, also known as maximum entropy.
The focus of this maximum entropy classifier is to minimize
memory consumption on very large datasets, particularly sparse
document-term matrices represented by the tm package. The
classifier is based on an efficient C++ implementation written by Dr. Yoshimasa Tsuruoka.

License GPL-3

LazyLoad yes

NeedsCompilation yes

Repository CRAN

Date/Publication 2013-11-14 18:03:19

R topics documented:

maxent-package . . . 2
as.compressed.matrix . . . 3
load.model . . . 4
maxent . . . 5
maxent-class . . . 7
NYTimes . . . 8
predict.maxent . . . 8
save.model . . . 9
tune.maxent . . . 10
USCongress . . . 11

Index **13**

`maxent-package` *Low-memory Multinomial Logistic Regression with Support for Text Classification*

Description

maxent is an R package with tools for low-memory multinomial logistic regression, also known as maximum entropy. The focus of this maximum entropy classifier is to minimize memory consumption on very large datasets, particularly sparse document-term matrices represented by the **tm** package. The library is built on top of an efficient C++ implementation written by Yoshimasa Tsuruoka.

Details

Package:	maxent
Type:	Package
Version:	1.3.3
Date:	2013-04-06
License:	GPL-3
LazyLoad:	yes

Author(s)

Timothy P. Jurka <tpjurka@ucdavis.edu>

References

Y. Tsuruoka. "A simple C++ library for maximum entropy classification." University of Tokyo Department of Computer Science (Tsujii Laboratory), 2011. URL http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/maxent/.

Examples

```
# LOAD LIBRARY
library(maxent)

# READ THE DATA, PREPARE THE CORPUS, and CREATE THE MATRIX
data <- read.csv(system.file("data/NYTimes.csv.gz",package="maxent"))
corpus <- Corpus(VectorSource(data$Title[1:150]))
matrix <- DocumentTermMatrix(corpus)

# TRAIN/PREDICT USING SPARSEM REPRESENTATION
sparse <- as.compressed.matrix(matrix)
model <- maxent(sparse[1:100,],data$Topic.Code[1:100])
results <- predict(model,sparse[101:150,])
```

`as.compressed.matrix` *converts a tm DocumentTermMatrix or TermDocumentMatrix into a matrix.csr representation.*

Description

Converts a DocumentTermMatrix or TermDocumentMatrix (package tm), Matrix (package Matrix), matrix.csr (SparseM), data.frame, or matrix into a `matrix.csr` representation to be used in the `maxent` and `predict.maxent` functions.

Usage

```
as.compressed.matrix(DocumentTermMatrix)
```

Arguments

`DocumentTermMatrix`
: A class of type DocumentTermMatrix or TermDocumentMatrix (package tm), Matrix (package Matrix), matrix.csr (SparseM), data.frame, or matrix.

Value

A `matrix.csr` representation of the DocumentTermMatrix or TermDocumentMatrix (package tm), Matrix (package Matrix), matrix.csr (SparseM), data.frame, or matrix.

Author(s)

Timothy P. Jurka <tpjurka@ucdavis.edu>

Examples

```
# LOAD LIBRARY
library(maxent)

# READ THE DATA, PREPARE THE CORPUS, and CREATE THE MATRIX
data <- read.csv(system.file("data/NYTimes.csv.gz",package="maxent"))
corpus <- Corpus(VectorSource(data$Title[1:150]))
matrix <- DocumentTermMatrix(corpus)

# CREATE A MATRIX.CSR (SPARSEM) REPRESENTATION
sparse <- as.compressed.matrix(matrix)
```

`load.model` *loads a maximum entropy model from a file.*

Description

Loads a multinomial logistic regression model of class `maxent-class` given a `file` created by function `save.model`.

Usage

```
load.model(file)
```

Arguments

`file`	The path to a file created by function `save.model`.

Value

Returns an object of class `maxent-class` with two slots.

`model`	A `character vector` containing the trained maximum entropy model.
`weights`	A `data.frame` listing all the weights in three columns: `Weight`, `Label`, and `Feature`.

Author(s)

Timothy P. Jurka <tpjurka@ucdavis.edu>

Examples

```
# LOAD LIBRARY
library(maxent)

# READ THE DATA, PREPARE THE CORPUS, and CREATE THE MATRIX
data <- read.csv(system.file("data/NYTimes.csv.gz",package="maxent"))
corpus <- Corpus(VectorSource(data$Title[1:150]))
matrix <- DocumentTermMatrix(corpus)
```

```
# TRAIN USING SPARSEM REPRESENTATION
sparse <- as.compressed.matrix(matrix)
model <- maxent(sparse[1:100,],as.factor(data$Topic.Code)[1:100])

save.model(model,"myModel")
saved_model <- load.model("myModel")
results <- predict(saved_model,sparse[101:150,])
```

`maxent` *trains a maximum entropy model given a training matrix and a vector or factor of labels.*

Description

Trains a multinomial logistic regression model of class `maxent-class` given a `matrix` or `matrix.csr` with training data, and a `vector` or `factor` with corresponding labels. Additional parameters such as `feature_cutoff`, `gaussian_prior`, `inequality_constraints`, and `set_heldout` help prevent model overfitting.

Usage

```
maxent(feature_matrix, code_vector, l1_regularizer=0.0, l2_regularizer=0.0,
       use_sgd=FALSE, set_heldout=0, verbose=FALSE)
```

Arguments

`feature_matrix`	A DocumentTermMatrix or TermDocumentMatrix (package tm), Matrix (package Matrix), matrix.csr (SparseM), data.frame, or matrix.
`code_vector`	A `factor` or `vector` of labels corresponding to each document in the `feature_matrix`.
`l1_regularizer`	An `numeric` turning on L1 regularization and setting the regularization parameter. A value of 0 will disable L1 regularization.
`l2_regularizer`	An `numeric` turning on L2 regularization and setting the regularization parameter. A value of 0 will disable L2 regularization.
`use_sgd`	A `logical` indicating that SGD parameter estimation should be used. Defaults to `FALSE`.
`set_heldout`	An `integer` specifying the number of documents to hold out. Sets a held-out subset of your data to test against and prevent overfitting.
`verbose`	A `logical` specifying whether to provide descriptive output about the training process. Defaults to `FALSE`, or no output.

Details

Yoshimasa Tsuruoka recommends using one of following three methods if you see overfitting.

1. Set the `l1_regularizer` parameter to `1.0`, leaving `l2_regularizer` and `set_heldout` as default.

2. Set the `l2_regularizer` parameter to `1.0`, leaving `l1_regularizer` and `set_heldout` as default.

3. Set the `set_heldout` parameter to hold-out a portion of your data, leaving `l1_regularizer` and `l2_regularizer` as default.

If you are using a large number of training samples, try setting the `use_sgd` parameter to `TRUE`.

Value

Returns an object of class `maxent-class` with two slots.

`model`	A `character vector` containing the trained maximum entropy model.
`weights`	A `data.frame` listing all the weights in three columns: `Weight`, `Label`, and `Feature`.

Author(s)

Timothy P. Jurka <tpjurka@ucdavis.edu>

References

Y. Tsuruoka. "A simple C++ library for maximum entropy classification." University of Tokyo Department of Computer Science (Tsujii Laboratory), 2011. URL http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/maxent/.

Examples

```
# LOAD LIBRARY
library(maxent)

# READ THE DATA, PREPARE THE CORPUS, and CREATE THE MATRIX
data <- read.csv(system.file("data/NYTimes.csv.gz",package="maxent"))
corpus <- Corpus(VectorSource(data$Title[1:150]))
matrix <- DocumentTermMatrix(corpus)

# TRAIN USING SPARSEM REPRESENTATION
sparse <- as.compressed.matrix(matrix)
model <- maxent(sparse[1:100,],as.factor(data$Topic.Code)[1:100])

# A DIFFERENT EXAMPLE (taken from package e10711)
# CREATE DATA
x <- seq(0.1, 5, by = 0.05)
y <- log(x) + rnorm(x, sd = 0.2)

# ESTIMATE MODEL AND PREDICT INPUT VALUES
m <- maxent(x, y)
```

```
new <- predict(m, x)

# VISUALIZE
plot(x, y)
points(x, log(x), col = 2)
points(x, new[,1], col = 4)
```

---

maxent-class *an S4 class containing the trained maximum entropy model.*

Description

An S4 class containing the trained maximum entropy model and its corresponding weights as a data.frame with three columns: Weight, Label, and Feature.

Objects from the Class

Objects could in principle be created by calls of the form new("maxent", ...). The preferred form is to have them created via a call to maxent.

Slots

model Object of class "character": stores the trained maximum entropy model as returned from maxent

weights Object of class "data.frame": contains the weights of the trained maximum entropy model, with three columns: Weight, Label, and Feature.

Author(s)

Timothy P. Jurka <tpjurka@ucdavis.edu>

Examples

```
# LOAD LIBRARY
library(maxent)

# READ THE DATA, PREPARE THE CORPUS, and CREATE THE MATRIX
data <- read.csv(system.file("data/NYTimes.csv.gz",package="maxent"))
corpus <- Corpus(VectorSource(data$Title[1:150]))
matrix <- DocumentTermMatrix(corpus)

# TRAIN USING SPARSEM REPRESENTATION
sparse <- as.compressed.matrix(matrix)
model <- maxent(sparse[1:100,],as.factor(data$Topic.Code)[1:100])
class(model)
model@model
model@weights
```

NYTimes — *a sample dataset containing labeled headlines from The New York Times.*

Description

A sample dataset containing labeled headlines from The New York Times, compiled by Professor Amber E. Boydstun at the University of California, Davis.

Usage

```
data(NYTimes)
```

Format

A `data.frame` containing five columns.

1. `Article_ID` - A unique identifier for the headline from The New York Times.
2. `Date` - The date the headline appeared in The New York Times.
3. `Title` - The headline as it appeared in The New York Times.
4. `Subject` - A manually classified subject of the headline.
5. `Topic.Code` - A manually labeled topic code corresponding to the subject.

Source

http://www.amberboydstun.com/

Examples

```
# READ THE CSV
data <- read.csv(system.file("data/NYTimes.csv.gz",package="maxent"))
# ALTERNATIVELY, USE THE data() FUNCTION
data(NYTimes)
```

predict.maxent — *predicts the expected label of a document given a trained model.*

Description

Predicts the expected labels and probability scores of a `matrix` of documents given a trained model of class `maxent-class` generated by function `maxent`.

Usage

```
## S3 method for class 'maxent'
predict(object, feature_matrix, ...)
```

Arguments

object	An object of class maxent-class, as returned by the maxent function.
feature_matrix	Either a regular matrix of class DocumentTermMatrix or TermDocumentMatrix from package **tm**, a matrix.csr representation generated by as.compressed.matrix, Matrix (package Matrix), matrix.csr (SparseM), data.frame, or matrix.
...	Not used but needed for compatibility with generic predict method.

Value

Returns a matrix with the first column containing predicted labels, and the remaining columns containing probability scores for each unique label.

Author(s)

Timothy P. Jurka <tpjurka@ucdavis.edu>

References

Y. Tsuruoka. "A simple C++ library for maximum entropy classification." University of Tokyo Department of Computer Science (Tsujii Laboratory), 2011. URL http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/maxent/.

Examples

```
# LOAD LIBRARY
library(maxent)

# READ THE DATA, PREPARE THE CORPUS, and CREATE THE MATRIX
data <- read.csv(system.file("data/NYTimes.csv.gz",package="maxent"))
corpus <- Corpus(VectorSource(data$Title[1:150]))
matrix <- DocumentTermMatrix(corpus)

# TRAIN/PREDICT USING SPARSEM REPRESENTATION
sparse <- as.compressed.matrix(matrix)
model <- maxent(sparse[1:100,],as.factor(data$Topic.Code)[1:100])
results <- predict(model,sparse[101:150,])
```

---

save.model *saves a maximum entropy model to a file.*

Description

Saves a multinomial logistic regression model of class maxent-class to a specified file. This model can then be loaded using function load.model.

Usage

```
save.model(model,file)
```

Arguments

model	An object of class maxent-class, as returned by the maxent function.
file	The path to a file used to save the model.

Author(s)

Timothy P. Jurka <tpjurka@ucdavis.edu>

Examples

```
# LOAD LIBRARY
library(maxent)

# READ THE DATA, PREPARE THE CORPUS, and CREATE THE MATRIX
data <- read.csv(system.file("data/NYTimes.csv.gz",package="maxent"))
corpus <- Corpus(VectorSource(data$Title[1:150]))
matrix <- DocumentTermMatrix(corpus)

# TRAIN USING SPARSEM REPRESENTATION
sparse <- as.compressed.matrix(matrix)
model <- maxent(sparse[1:100,],as.factor(data$Topic.Code)[1:100])
save.model(model,"myModel")

# TRAIN USING REGULAR MATRIX REPRESENTATION
model <- maxent(as.matrix(matrix)[1:100,],as.factor(data$Topic.Code)[1:100])
save.model(model,"myModel")
```

---

tune.maxent *fits a maximum entropy model given a training matrix and a vector or factor of labels.*

Description

Fits a multinomial logistic regression model of class maxent-class given a matrix or matrix.csr with training data, and a vector or factor with corresponding labels.

Usage

```
tune.maxent(feature_matrix, code_vector, nfold=3, showall=FALSE, verbose=FALSE)
```

Arguments

feature_matrix	A DocumentTermMatrix or TermDocumentMatrix (package tm), Matrix (package Matrix), matrix.csr (SparseM), data.frame, or matrix.
code_vector	A factor or vector of labels corresponding to each document in the feature_matrix.
nfold	An integer specifying the number of folds to perform for cross-validation. Defaults to 3.

showall	A logical specifying whether to show the accuracy results of all tested parameter configurations. Defaults to FALSE.
verbose	A logical specifying whether to provide descriptive output about the fitting process. Defaults to FALSE, or no output.

Value

Returns an object of class matrix with configurations along the y-axis and parameters along the x-axis.

Author(s)

Timothy P. Jurka <tpjurka@ucdavis.edu>

Examples

```
# LOAD LIBRARY
library(maxent)

# A DIFFERENT EXAMPLE
data(iris)
attach(iris)

x <- subset(iris, select = -Species)
y <- Species

f <- tune.maxent(x,y,nfold=3,showall=TRUE)
```

USCongress *a sample dataset containing labeled bills from the United State Congress.*

Description

A sample dataset containing labeled bills from the United States Congress, compiled by Professor John D. Wilkerson at the University of Washington, Seattle and E. Scott Adler at the University of Colorado, Boulder.

Usage

```
data(USCongress)
```

Format

A data.frame containing five columns.

1. ID - A unique identifier for the bill.
2. cong - The session of congress that the bill first appeared in.

3. billnum - The number of the bill as it appears in the congressional docket.

4. h_or_sen - A field specifying whether the bill was introduced in the House (HR) or the Senate (S).

5. major - A manually labeled topic code corresponding to the subject of the bill.

Source

http://www.congressionalbills.org/

Examples

```
# READ THE CSV
data <- read.csv(system.file("data/USCongress.csv.gz",package="maxent"))
# ALTERNATIVELY, USE THE data() FUNCTION
data(USCongress)
```

Index

∗Topic **classes**
    maxent-class, 7
∗Topic **datasets**
    NYTimes, 8
    USCongress, 11
∗Topic **methods**
    as.compressed.matrix, 3
    load.model, 4
    maxent, 5
    predict.maxent, 8
    save.model, 9
    tune.maxent, 10
∗Topic **package**
    maxent-package, 2

as.compressed.matrix, 3, *9*

load.model, 4, *9*

maxent, *3*, 5, *7–10*
maxent-class, 7
maxent-package, 2

NYTimes, 8

predict.maxent, *3*, 8

save.model, *4*, 9

tune.maxent, 10

USCongress, 11