Package 'cattonum'

May 2, 2018

Type Package

Version 0.0.2

Title Encode Categorical Features

Description Functions for dummy encoding, frequency encoding, label encoding, leave-one-out encoding, mean encoding, median encoding, and one-hot encoding.

Maintainer Bernie Gray <bfgray3@gmail.com>

URL https://github.com/bfgray3/cattonum

BugReports https://github.com/bfgray3/cattonum/issues

Encoding UTF-8

Depends R (>= 3.2.0)

Imports dplyr (>= 0.7.0), stats, tidyselect (>= 0.2.3)

LazyData true

License MIT + file LICENSE

Suggests covr, knitr, nycflights13, ranger, rmarkdown, testthat, tibble

RoxygenNote 6.0.1

VignetteBuilder knitr

NeedsCompilation no

Author Bernie Gray [aut, cre], Mark Roepke [ctb]

Repository CRAN

Date/Publication 2018-05-02 03:18:21 UTC

R topics documented:

`cattonum` *cattonum: Encode Categorical Features*

Description

Functions for dummy encoding, frequency encoding, label encoding, leave-one-out encoding, mean encoding, median encoding, and one-hot encoding.

`catto_dummy` *Dummy encoding*

Description

Dummy encoding

Usage

```
catto_dummy(train, ..., test, verbose = TRUE)
```

Arguments

`train`	The training data, in a `data.frame` or `tibble`.
`...`	The columns to be encoded. If none are specified, then all character and factor columns are encoded.
`test`	The test data, in a `data.frame` or `tibble`.
`verbose`	Should informative messages be printed? Defaults to `TRUE` (not yet used).

Value

The encoded dataset in a `data.frame` or `tibble`, whichever was input. If a test dataset was provided, a list with names "train" and "test" is returned holding the encoded training and test datasets.

Examples

```
catto_dummy(iris)
```

catto_freq *Frequency encoding*

Description

Frequency encoding

Usage

```
catto_freq(train, ..., test, verbose = TRUE)
```

Arguments

train	The training data, in a data.frame or tibble.
...	The columns to be encoded. If none are specified, then all character and factor columns are encoded.
test	The test data, in a data.frame or tibble.
verbose	Should informative messages be printed? Defaults to TRUE (not yet used).

Value

The encoded dataset in a data.frame or tibble, whichever was input. If a test dataset was provided, a list with names "train" and "test" is returned holding the encoded training and test datasets.

Examples

```
catto_freq(iris)
```

catto_label *Label encoding*

Description

Label encoding

Usage

```
catto_label(train, ..., test, ordering = "increasing", verbose = TRUE,
  seed = 4444)
```

Arguments

train	The training data, in a data.frame or tibble.
...	The columns to be encoded. If none are specified, then all character and factor columns are encoded.
test	The test data, in a data.frame or tibble.
ordering	How should labels be assigned to levels? There are three different ways to pass this argument. First, a length one character vector with value "increasing", "decreasing", "observed", or "random" will apply that ordering to each column being encoded. Second, a character vector of length greater than one may be passed, specifying one of the above four options for each column being encoded. Finally, a list may be passed specifying a user-defined ordering for each column being encoded.
verbose	Should informative messages be printed? Defaults to TRUE (not yet used).
seed	The random seed set before all random ordering encodings if there are any.

Value

The encoded dataset in a data.frame or tibble, whichever was input. If a test dataset was provided, a list with names "train" and "test" is returned holding the encoded training and test datasets.

Examples

```
catto_label(iris)

y <- 2 ^ (0:5)
x1 <- c("a", "b", NA, "b", "a", "a")
x2 <- c("c", "c", "c", "d", "d", "c")
df_fact <- data.frame(y, x1, x2)

catto_label(df_fact,
            ordering = list(c("b", "a"), c("c", "d")))

catto_label(df_fact, ordering = c("increasing", "decreasing"))
```

---

catto_loo *Leave-one-out encoding*

Description

Leave-one-out encoding

Usage

```
catto_loo(train, ..., response, test, verbose = TRUE)
```

Arguments

train	The training data, in a data.frame or tibble.
...	The columns to be encoded. If none are specified, then all character and factor columns are encoded.
response	The response variable used to calculate means.
test	The test data, in a data.frame or tibble.
verbose	Should informative messages be printed? Defaults to TRUE.

Value

The encoded dataset in a data.frame or tibble, whichever was input. If a test dataset was provided, a list with names "train" and "test" is returned holding the encoded training and test datasets.

Examples

```
catto_loo(iris, response = Sepal.Length)
```

---

catto_mean *Mean encoding*

Description

Mean encoding

Usage

```
catto_mean(train, ..., response, test, verbose = TRUE)
```

Arguments

train	The training data, in a data.frame or tibble.
...	The columns to be encoded. If none are specified, then all character and factor columns are encoded.
response	The response variable used to calculate means.
test	The test data, in a data.frame or tibble.
verbose	Should informative messages be printed? Defaults to TRUE.

Value

The encoded dataset in a data.frame or tibble, whichever was input. If a test dataset was provided, a list with names "train" and "test" is returned holding the encoded training and test datasets.

Examples

```
catto_mean(iris, response = Sepal.Length)
```

catto_median *Median encoding*

Description

Median encoding

Usage

```
catto_median(train, ..., response, test, verbose = TRUE)
```

Arguments

train	The training data, in a data.frame or tibble.
...	The columns to be encoded. If none are specified, then all character and factor columns are encoded.
response	The response variable used to calculate medians.
test	The test data, in a data.frame or tibble.
verbose	Should informative messages be printed? Defaults to TRUE.

Value

The encoded dataset in a data.frame or tibble, whichever was input. If a test dataset was provided, a list with names "train" and "test" is returned holding the encoded training and test datasets.

Examples

```
catto_median(iris, response = Sepal.Length)
```

catto_onehot *One-hot encoding*

Description

One-hot encoding

Usage

```
catto_onehot(train, ..., test, verbose = TRUE)
```

Arguments

train	The training data, in a data.frame or tibble.
...	The columns to be encoded. If none are specified, then all character and factor columns are encoded.
test	The test data, in a data.frame or tibble.
verbose	Should informative messages be printed? Defaults to TRUE (not yet used).

Value

The encoded dataset in a data.frame or tibble, whichever was input. If a test dataset was provided, a list with names "train" and "test" is returned holding the encoded training and test datasets.

Examples

```
catto_onehot(iris)
```

Index