Package 'AugmenterR'

October 12, 2022

Type Package

Title Data Augmentation for Machine Learning on Tabular Data

Version 0.1.0

Maintainer Rafael S. Pereira <r.s.p.models@gmail.com>

Description Implementation of a data augmentation technique based on conditional entropy
It was devised by both authors during their masters and is discussed in detail in the second author dissertation.
It is able to create novel samples conditioned on a desired value of a categorical attribute, as a way to augment data for classification tasks
Tests discussed in the dissertation and future paper present that the technique satisfies several statistical assumptions for the novel samples.
It also shows significant improvement for machine learning models trained on small data.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Suggests knitr, ggplot2, markdown

VignetteBuilder knitr

RoxygenNote 6.1.0.9000

NeedsCompilation no

Author Rafael S. Pereira [aut, cre, cph],
Henrique Matheus ferreira da silva [aut, cph],
Fabio A.M Porto [aut, ths, cph]

Repository CRAN

Date/Publication 2021-03-18 09:00:05 UTC

R topics documented:

`Generate` `Generate` *Asks for a dataframe and generates a new sample returns novel sample along with intervals it contained to revalidate it using confidence levels*

Description

`Generate` Asks for a dataframe and generates a new sample returns novel sample along with intervals it contained to revalidate it using confidence levels

Usage

```
Generate(data, regression = FALSE)
```

Arguments

`data`	Dataframe
`regression`	if we are to generate data for regression or classification (will the data be conditioned on a certain class)

Value

if regression is true returns a dataframe, if false returns a list.

Examples

```
# basic usage of Generate
Generate(iris,regression=TRUE)
```

`GenerateASingleCandidate` `GenerateASingleCandidate` *Generates a novel sample from a target class and evaluate it against the other classes to check if it satisfies the confidence level returns NA if the generated sample does not satisfy the condition, otherwise returns novel sample*

Description

`GenerateASingleCandidate` Generates a novel sample from a target class and evaluate it against the other classes to check if it satisfies the confidence level returns NA if the generated sample does not satisfy the condition, otherwise returns novel sample

Usage

```
GenerateASingleCandidate(data, Class, col, Prob)
```

Arguments

data	A dataframe containing available data
Class	The target class
col	Column the target class is located in the dataframe
Prob	Minimum confidence level to generate sample

Value

A dataframe containing a novel sample if it satisfies the confidence given in Prob, otherwise NA

GenerateMultipleCandidates

GenerateMultipleCandidates *Asks for a dataframe and some parameters and returns multiple novel samples from the target class*

Description

GenerateMultipleCandidates Asks for a dataframe and some parameters and returns multiple novel samples from the target class

Usage

```
GenerateMultipleCandidates(data, Class, col, Prob, amount)
```

Arguments

data	Reference dataframe
Class	Value of the target class
col	column of the dataframe which contains the class
Prob	Minimum confidence level to generate the sample
amount	Number of novel samples to be generated

Value

A dataframe containing novel samples of the class Class that satisfies Prob confidence.

Examples

```
# basic usage of GenerateMultipleCandidates
GenerateMultipleCandidates(iris,Class='virginica',col=5,Prob=0.3,amount=10)
```

ObtainCandidate	ObtainCandidate *Asks for a vector and returns a value along with the range it is contained in the attribute Is used alongside other functions when generating a new sample*

Description

ObtainCandidate Asks for a vector and returns a value along with the range it is contained in the attribute Is used alongside other functions when generating a new sample

Usage

```
ObtainCandidate(Dado)
```

Arguments

Dado	vector containing an attribute of your dataframe

Value

Obtains a sample of an attribute based on their possible values, it is a part of the Generate function

Index