Package ‘BigKnn’

May 30, 2024

Type Package

Title Large Scale K-Nearest Neighbor Classifier using the Lucene Search Engine

Version 1.0.2

Date 2025-05-30

Maintainer Martijn Schuemie <schuemie@ohdsi.org>

Description A large scale k-nearest neighbor classifier using the Lucene search engine.

SystemRequirements Java version 8 or higher (https://www.java.com/)

Imports rJava,
Andromeda (>= 0.6.3),
dplyr,
rlang

Suggests testthat

License Apache License

RoxygenNote 7.3.1

URL https://ohdsi.github.io/BigKnn, https://github.com/OHDSI/BigKnn

BugReports https://github.com/OHDSI/BigKnn/issues

Encoding UTF-8

Contents

buildKnn *Build a K-nearest neighbor (KNN) classifier*

Description

buildKnn loads data from two Andromeda tables, and inserts them into a KNN classifier.

Usage

```
buildKnn(outcomes, covariates, indexFolder, overwrite = TRUE)
```

Arguments

outcomes	An Andromeda table containing the outcomes with predefined columns (see below).
covariates	An Andromeda table containing the covariates with predefined columns (see below).
indexFolder	Path to a local folder where the KNN classifier index can be stored.
overwrite	Automatically overwrite if an index already exists?

Details

These columns are expected in the outcome object:

rowId	(integer)	Row ID is used to link multiple covariates (x) to a single outcome (y)
y	(real)	The outcome variable

These columns are expected in the covariates object:

rowId	(integer)	Row ID is used to link multiple covariates (x) to a single outcome (y)
covariateId	(integer)	A numeric identifier of a covariate
covariateValue	(real)	The value of the specified covariate

Value

Nothing

buildKnnFromPlpData *Build a K-nearest neighbor (KNN) classifier from a plpData object*

Description

Build a K-nearest neighbor (KNN) classifier from a plpData object

Usage

```
buildKnnFromPlpData(
  plpData,
  population,
  indexFolder,
  overwrite = TRUE,
  cohortId = NULL,
  outcomeId = NULL
)
```

Arguments

plpData	An object of type plpData.
population	The population.
indexFolder	Path to a local folder where the KNN classifier index can be stored.
overwrite	Automatically overwrite if an index already exists?
cohortId	The ID of the specific cohort for which to fit a model.
outcomeId	The ID of the specific outcome for which to fit a model.

predictKnn *Predict using a K-nearest neighbor (KNN) classifier*

Description

predictKnn uses a KNN classifier to generate predictions.

Usage

```
predictKnn(
  cohorts,
  covariates,
  indexFolder,
  k = 1000,
  weighted = TRUE,
  threads = 1
)
```

Arguments

`cohorts`	An Andromeda table containing the cohorts with predefined columns (see below).
`covariates`	An Andromeda table containing the covariates with predefined columns (see below).
`indexFolder`	Path to a local folder where the KNN classifier index can be stored.
`k`	The number of nearest neighbors to use to predict the outcome.
`weighted`	Should the prediction be weighted by the (inverse of the) distance metric?
`threads`	Number of parallel threads to used for the computation.

Details

These columns are expected in the covariates object:

`rowId`	(integer)	Row ID is used to link multiple covariates (x) to a single outcome (y)
`covariateId`	(integer)	A numeric identifier of a covariate
`covariateValue`	(real)	The value of the specified covariate

This column is expected in the covariates object:

`rowId`	(integer)	Row ID is used to link multiple covariates (x) to a single outcome (y)

Value

A data.frame with two columns:

`rowId`	(integer)	Row ID is used to link multiple covariates (x) to a single outcome (y)
`prediction`	(real)	A number between 0 and 1 representing the probability of the outcome

---

predictKnnUsingPlpData

Create predictive probabilities using KNN.

Description

Create predictive probabilities using KNN.

Usage

```
predictKnnUsingPlpData(
  plpData,
  population,
  indexFolder,
  k = 1000,
  weighted = TRUE,
  threads = 10
)
```

Arguments

`plpData`	An object of type `plpData` as generated using `getDbPlpData`.
`population`	The population to predict for.
`indexFolder`	Path to a local folder where the KNN classifier index is be stored.
`k`	The number of nearest neighbors to use to predict the outcome.
`weighted`	Should the prediction be weigthed by the (inverse of the) distance metric?
`threads`	Number of parallel threads to used for the computation.

Details

Generates predictions for the population specified in plpData.

Value

The value column in the result data.frame is: logistic: probabilities of the outcome, poisson: Poisson rate (per day) of the outcome, survival: hazard rate (per day) of the outcome.

Index