Package 'BigKnn'

March 27, 2020

Type Package

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

Version 0.0.3

Date 2020-03-27

Maintainer Martijn Schuemie <schuemie@ohdsi.org>

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

Imports rJava,
Cyclops,
ParallelLogger,
ff,
ffbase,
bit

License Apache License

RoxygenNote 7.0.2

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

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

Encoding UTF-8

R topics documented:

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

Description

`buildKnn` loads data from two ffdf objects, and inserts them into a KNN classifier.

Usage

```
buildKnn(
  outcomes,
  covariates,
  indexFolder,
  overwrite = TRUE,
  checkSorting = TRUE,
  checkRowIds = TRUE,
  quiet = FALSE
)
```

Arguments

outcomes	A ffdf object containing the outcomes with predefined columns (see below).
covariates	A ffdf object 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?
checkSorting	Check if the data are sorted appropriately, and if not, sort.
checkRowIds	Check if all rowIds in the covariates appear in the outcomes.
quiet	If true, (warning) messages are suppressed.

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

Note: If checkSorting is turned off, the covariate table should be sorted by rowId.

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,
  indexFolder,
  overwrite = TRUE,
  removeDropouts = TRUE,
  cohortId = NULL,
  outcomeId = NULL
)
```

Arguments

plpData	An object of type plpData.
indexFolder	Path to a local folder where the KNN classifier index can be stored.
overwrite	Automatically overwrite if an index already exists?
removeDropouts	If TRUE subjects that do not have the full observation window (i.e. are censored earlier) and do not have the outcome are removed prior to fitting the model.
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(
  covariates,
  cohorts,
  indexFolder,
  k = 1000,
  weighted = TRUE,
  checkSorting = TRUE,
  quiet = FALSE,
  threads = 1
)
```

Arguments

covariates	A ffdf object containing the covariates with predefined columns (see below).
cohorts	A ffdf object containing the cohorts 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?
checkSorting	Check if the data are sorted appropriately, and if not, sort.
quiet	If true, (warning) messages are surpressed.
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)

Note: If checkSorting is turned off, the covariate table should be sorted by rowId.

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

Arguments

`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.
`plpData`	An object of type `plpData` as generated using `getDbPlpData`.

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