Package ‘auctestr’

October 12, 2022

Title Statistical Testing for AUC Data
Version 1.0.0
Maintainer Josh Gardner <jpgard@umich.edu>
Description Performs statistical testing to compare predictive models based on multiple observations of the A' statistic (also known as Area Under the Receiver Operating Characteristic Curve, or AUC). Specifically, it implements a testing method based on the equivalence between the A' statistic and the Wilcoxon statistic.

Imports dplyr, tidyr
Depends R (>= 3.3.1)
License MIT + file LICENSE
LazyData true
RoxygenNote 6.0.1
Suggests knitr, rmarkdown, testthat
VignetteBuilder knitr
NeedsCompilation no
Author Josh Gardner [aut, cre]
Repository CRAN
Date/Publication 2017-11-13 09:46:18 UTC

R topics documented:

  auctestr ................................................................. 2
  auc_compare ........................................................... 2
  fbh_test ............................................................... 4
  sample_experiment_data .............................................. 5
  se_auc ................................................................. 6
  stouffer_z ............................................................ 7

Index 8
auc_test

*auctestr*  *Statistical Testing for AUC data.*

**Description**

*auctestr* currently provides four main useful functions for statistical testing of the AUC, or $A'$, statistic: *fbh_auc_compare*, *stouffer_z*, *fbh_test*, and *se_auc*.

**auc_compare**  *Compare AUC values using the FBH method.*

**Description**

Apply the FBH method to compare `outcome_col` by `compare_col`, averaging over `time_col` (due to non-independence) and then over `over_col` by using Stouffer's Method.

**Usage**

```r
auc_compare(df, compare_values, filter_value, time_col = "time", outcome_col = "auc", compare_col = "model_id", over_col = "dataset", n_col = "n", n_p_col = "n_p", n_n_col = "n_n", filter_col = "model_variant")
```

**Arguments**

- `df`: DataFrame containing `time_col`, `outcome_col`, `compare_col`, and `over_col`.
- `compare_values`: names of models to compare (character vector of length 2). These should match exactly the names as they appear in `compare_col`.
- `filter_value`: (optional) keep only observations which contain `filter_value` for `filter_col`.
- `time_col`: name of column in `df` representing time of observations ($z$-scores are averaged over `time_col` within each model/dataset due to non-independence). These can also be other dependent groupings, such as cross-validation folds.
- `outcome_col`: name of column in `df` representing outcome to compare; this should be Area Under the Receiver Operating Characteristic or $A'$ statistic (this method applies specifically to AUC and not other metrics (i.e., sensitivity, precision, F1)).
- `compare_col`: name of column in `df` representing two conditions to compare (should contain at least 2 unique values; these two values are specified as `compare_values`).
- `over_col`: identifier for independent experiments, iterations, etc. over which $z$-scores for models are to be compared (using Stouffer’s $Z$).
- `n_col`: name of column in `df` with total number of observations in the sample tested by each row.
- `n_p_col`: name of column in `df` with `n_p`, number of positive observations.
- `n_n_col`: name of column in `df` with `n_n`, number of negative observations.
- `filter_col`: (optional) name of column in `df` to filter observations on; keep only observations which contain `filter_value` for `filter_col`.  

**Value**

numeric, overall z-score of comparison using the FBH method.

**References**


**See Also**

Other fbh method: fbh_test, se_auc

**Examples**

```r
## load sample experiment data
data(sample_experiment_data)
## compare VariantA of ModelA and ModelB
auc_compare(sample_experiment_data,
            compare_values = c('ModelA', 'ModelB'),
            filter_value = c('VariantA'),
            time_col = 'time',
            outcome_col = 'auc',
            compare_col = 'model_id',
            over_col = 'dataset',
            filter_col = 'model_variant')
## compare VariantC of ModelA and ModelB
auc_compare(sample_experiment_data,
            compare_values = c('ModelA', 'ModelB'),
            filter_value = c('VariantC'),
            time_col = 'time',
            outcome_col = 'auc',
            compare_col = 'model_id',
            over_col = 'dataset',
            filter_col = 'model_variant')
## compare ModelC, VariantA and VariantB
auc_compare(sample_experiment_data,
            compare_values = c('VariantA', 'VariantB'),
            filter_value = c('ModelC'),
            time_col = 'time',
            outcome_col = 'auc',
            compare_col = 'model_variant',
            over_col = 'dataset',
            filter_col = 'model_id')
```
fbh_test

Apply z-test for difference between auc_1 and auc_2 using FBH method.

Description

Apply z-test for difference between auc_1 and auc_2 using FBH method.

Usage

fbh_test(auc_1, auc_2, n_p, n_n)

Arguments

- **auc_1**: value of A’ statistic (or AUC, or Area Under the Receiver operating characteristic curve) for the first group (numeric).
- **auc_2**: value of A’ statistic (or AUC, or Area Under the Receiver operating characteristic curve) for the second group (numeric).
- **n_p**: number of positive observations (needed for calculation of standard error of Wilcoxon statistic) (numeric).
- **n_n**: number of negative observations (needed for calculation of standard error of Wilcoxon statistic) (numeric).

Value

numeric, single aggregated z-score of comparison A’_1 - A’_2.

References


See Also

Other fbh method: auc_compare, se_auc

Examples

```r
## Two models with identical AUC return z-score of zero
fbh_test(0.56, 0.56, 1000, 2500)
## Compare two models; note that changing order changes sign of z-statistic
fbh_test(0.56, 0.59, 1000, 2500)
fbh_test(0.59, 0.56, 1000, 2500)
```
Performance of several predictive models over three different datasets, using multiple cutoff points for time within each dataset.

Description

A dataset containing the performance of several predictive models over three different datasets, where models are built using data from multiple time points (where time 1 has less data than time 2, but each subsequent time point T also uses data from all prior time points up to that time t <= T.) This represents the typical output of a machine learning experiment where several models are being considered across multiple datasets, often with different variants of each model type being considered (i.e., different hyperparameter settings of each model).

Usage

sample_experiment_data

Format

A data frame with 180 rows and 10 variables:

- **auc**: Area Under the Receiver Operating Characteristic Curve, or AUC, for this model configuration.
- **precision**: Precision for this model configuration.
- **accuracy**: Accuracy for this model configuration.
- **n**: Number of observations in this dataset.
- **n_n**: Number of negative observations (i.e., outcome == 0) in this dataset (required for standard error estimation of AUC statistic).
- **n_p**: Number of positive observations (i.e., outcome == 1) in this dataset (required for standard error estimation of AUC statistic).
- **dataset**: indicator for different datasets.
- **time**: indicator for different time points used to build each dataset; these represent dependent observations of model performance.
- **model_id**: Indicator for the statistical algorithm used (this could be 'Logistic Regression', 'SVM', etc.).
- **model_variant**: Indicator for different variants of each model which are not equivalent and should be used individually (model should not be averaged over these, and instead should be held fixed when comparing to other model). Example of this could be various hyperparameter settings for a given model (i.e., cost for an SVM).
se_auc

Compute standard error of AUC score, using its equivalence to the Wilcoxon statistic.

Description

Compute standard error of AUC score, using its equivalence to the Wilcoxon statistic.

Usage

se_auc(auc, n_p, n_n)

Arguments

- auc: value of A' statistic (or AUC, or Area Under the Receiver operating characteristic curve) (numeric).
- n_p: number of positive cases (integer).
- n_n: number of negative cases (integer).

References


See Also

Other fbh method: auc_compare, fbh_test

Examples

se_auc(0.75, 20, 200)
# standard error decreases when data become more balanced over positive/negative outcome class, holding sample size fixed
se_auc(0.75, 110, 110)
# standard error increases when sample size shrinks
se_auc(0.75, 20, 20)
**Description**

Compute aggregate z-score using Stouffer’s method.

**Usage**

```r
stouffer_z(z_vec, ignore.na = TRUE)
```

**Arguments**

- `z_vec`: vector of z-scores (numeric).
- `ignore.na`: should NA values be ignored? defaults to TRUE.

**Value**

numeric, Z-score using Stouffer’s method aggregated over `z_vec`.

**References**

Index

* datasets
  sample_experiment_data, 5

auc_compare, 2, 4, 6
auc_teststr, 2
auc_teststr-package (auc_teststr), 2

fbh_test, 3, 4, 6

sample_experiment_data, 5
se_auc, 3, 4, 6
stouffer_z, 7