Package 'dsos'

January 18, 2022

Title Dataset Shift with Outlier Scores

Version 0.1.0

Description Test for no adverse shift in two-sample comparison when we have a training set, the reference distribution, and a test set. The approach is flexible and relies on a robust and powerful test statistic, the weighted AUC. Technical details are in Kamulete, V. M. (2021) <arXiv:1908.04000>. Modern notions of outlyingness such as trust scores and prediction uncertainty can be used as the underlying scores for example.

License GPL (>= 3)

URL https://github.com/vathymut/dsos

BugReports https://github.com/vathymut/dsos/issues

Imports data.table (>= 1.14.0), ggplot2 (>= 3.3.3), isotree (>= 0.2.7), ranger (>= 0.12.1), scales (>= 1.1.1), simctest (>= 2.6), stats (>= 3.6.1), WeightedROC (>= 2020.1.31)

Suggests fdrtool (>= 1.2.16), knitr (>= 1.33), rmarkdown (>= 2.7), testthat (>= 3.0.2)

VignetteBuilder knitr

Encoding UTF-8

Language en-US

RoxygenNote 7.1.2

NeedsCompilation no

Author Vathy M. Kamulete [aut, cre] (<https://orcid.org/0000-0002-4451-3743>), Royal Bank of Canada (RBC) [cph] (Research supported by RBC)

Maintainer Vathy M. Kamulete <vathymut@gmail.com>

Repository CRAN

Date/Publication 2022-01-18 08:32:51 UTC

R topics documented:

`cp_at` *Dataset Shift via Class Probabilities*

Description

Test for no adverse shift via class probabilities for two-sample comparison. The scores are out-of-bag predictions from random forests with the package **ranger**. The prefix *cp* stands for class probability, whether the instance belongs to the training or test set. The probability of belonging to the test set is the relevant notion of outlyingness.

Usage

```
cp_at(x_train, x_test, R = 1000, num_trees = 500, sub_ratio = 1/2)
```

Arguments

`x_train`	Training sample.
`x_test`	Test sample.
`R`	The number of permutations. May be ignored.
`num_trees`	The number of trees in random forests.
`sub_ratio`	Subsampling ratio for sample splitting. May be ignored.

Details

The suffix *at* refers to the asymptotic test statistic. This variant uses the asymptotic null distribution for the weighted AUC (WAUC), the test statistic. Li & Fine (2010) derives its null distribution. This approximation is reliable in large sample; otherwise, prefer permutations for inference. The example below uses datasets with small samples, which is generally not advisable, is for illustration only.

Value

A named list or object of class `outlier.test` containing:

- `statistic`: observed WAUC statistic
- `seq_mct`: sequential Monte Carlo test, if applicable
- `p_value`: p-value
- `outlier_scores`: outlier scores from training and test set

Notes

Please see references for the classifier two-sample test, the inspiration behind this approach. Note that Ciemencon et al. (2009) uses both sample splitting for inference and the AUC, rather than the WAUC. Most supervised method for binary classification can replace random forests, the default in this implementation.

References

Kamulete, V. M. (2021). *Test for non-negligible adverse shifts*. arXiv preprint arXiv:2107.02990.

Ciemencon, S., Depecker, M., & Vayatis, N. (2009, December). *AUC optimization and the two-sample problem*. In Proceedings of the 22nd International Conference on Neural Information Processing Systems (pp. 360-368).

Lopez-Paz, D., & Oquab, M. (2016). *Revisiting classifier two-sample tests*. arXiv preprint arXiv:1610.06545.

Friedman, J. (2004). *On multivariate goodness-of-fit and two-sample testing*.

Gandy, A. (2009). *Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk*. Journal of the American Statistical Association, 104(488), 1504-1511.

Li, J., & Fine, J. P. (2010). *Weighted area under the receiver operating characteristic curve and its application to gene selection*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(4), 673-692.

Rinaldo, A., Wasserman, L., & G'Sell, M. (2019). *Bootstrapping and sample splitting for high-dimensional, assumption-lean inference*. Annals of Statistics, 47(6), 3438-3469.

See Also

[cp_ss()] for asymptotic p-value via sample splitting. [cp_pt()] for p-value approximation via permutations.

Other classifiers: `cp_pt()`, `cp_ss()`

Examples

```
library(dsos)
set.seed(12345)
data(iris)
x_train <- iris[1:50, 1:4] # Training sample: Species == 'setosa'
x_test <- iris[51:100, 1:4] # Test sample: Species == 'versicolor'
iris_test <- cp_at(x_train, x_test) # Can also use: cp_ss and cp_pt
str(iris_test)
```

cp_pt *Dataset Shift via Class Probabilities*

Description

Test for no adverse shift via class probabilities for two-sample comparison. The scores are out-of-bag predictions from random forests with the package **ranger**. The prefix *cp* stands for class probability, whether the instance belongs to the training or test set. The probability of belonging to the test set is the relevant notion of outlyingness.

Usage

```
cp_pt(x_train, x_test, R = 1000, num_trees = 500, sub_ratio = 1/2)
```

Arguments

x_train	Training sample.
x_test	Test sample.
R	The number of permutations. May be ignored.
num_trees	The number of trees in random forests.
sub_ratio	Subsampling ratio for sample splitting. May be ignored.

Details

The empirical null distribution uses R permutations to estimate the p-value. For speed, this is implemented as a sequential Monte Carlo test with the **simctest** package. See Gandy (2009) for details. The suffix *pt* refers to permutation test. It does not use the asymptotic (theoretical) null distribution for the weighted AUC (WAUC), the test statistic. This is the recommended approach for small samples.

Value

A named list or object of class `outlier.test` containing:

- `statistic`: observed WAUC statistic
- `seq_mct`: sequential Monte Carlo test, if applicable
- `p_value`: p-value
- `outlier_scores`: outlier scores from training and test set

Notes

Please see references for the classifier two-sample test, the inspiration behind this approach. Note that Ciemencon et al. (2009) uses both sample splitting for inference and the AUC, rather than the WAUC. Most supervised method for binary classification can replace random forests, the default in this implementation.

References

Kamulete, V. M. (2021). *Test for non-negligible adverse shifts*. arXiv preprint arXiv:2107.02990.

Ciemencon, S., Depecker, M., & Vayatis, N. (2009, December). *AUC optimization and the two-sample problem*. In Proceedings of the 22nd International Conference on Neural Information Processing Systems (pp. 360-368).

Lopez-Paz, D., & Oquab, M. (2016). *Revisiting classifier two-sample tests*. arXiv preprint arXiv:1610.06545.

Friedman, J. (2004). *On multivariate goodness-of-fit and two-sample testing*.

Gandy, A. (2009). *Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk*. Journal of the American Statistical Association, 104(488), 1504-1511.

Li, J., & Fine, J. P. (2010). *Weighted area under the receiver operating characteristic curve and its application to gene selection*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(4), 673-692.

Rinaldo, A., Wasserman, L., & G'Sell, M. (2019). *Bootstrapping and sample splitting for high-dimensional, assumption-lean inference*. Annals of Statistics, 47(6), 3438-3469.

See Also

Other classifiers: `cp_at()`, `cp_ss()`

Examples

```
library(dsos)
set.seed(12345)
data(iris)
x_train <- iris[1:50, 1:4] # Training sample: Species == 'setosa'
x_test <- iris[51:100, 1:4] # Test sample: Species == 'versicolor'
iris_test <- cp_at(x_train, x_test) # Can also use: cp_ss and cp_pt
str(iris_test)
```

`cp_ss` *Dataset Shift via Class Probabilities*

Description

Test for no adverse shift via class probabilities for two-sample comparison. The scores are out-of-bag predictions from random forests with the package **ranger**. The prefix *cp* stands for class probability, whether the instance belongs to the training or test set. The probability of belonging to the test set is the relevant notion of outlyingness.

Usage

```
cp_ss(x_train, x_test, sub_ratio = 1/2, R = 1000, num_trees = 500)
```

Arguments

x_train	Training sample.
x_test	Test sample.
sub_ratio	Subsampling ratio for sample splitting. May be ignored.
R	The number of permutations. May be ignored.
num_trees	The number of trees in random forests.

Details

This approach uses sample splitting to compute the p-value for inference. sub_ratio splits each sample into two parts: one half for estimation (calibration) and the half for inference, if sub_ratio is 1/2. In other words, it sacrifices some predictive accuracy for inferential robustness as in Rinaldo et al. (2019). The suffix *ss* refers to sample splitting. Sample splitting relies on the asymptotic null distribution for the weighted AUC (WAUC), the test statistic. Li & Fine (2010) derives its null distribution.

Value

A named list or object of class outlier.test containing:

- statistic: observed WAUC statistic
- seq_mct: sequential Monte Carlo test, if applicable
- p_value: p-value
- outlier_scores: outlier scores from training and test set

Notes

Please see references for the classifier two-sample test, the inspiration behind this approach. Note that Ciemencon et al. (2009) uses both sample splitting for inference and the AUC, rather than the WAUC. Most supervised method for binary classification can replace random forests, the default in this implementation.

References

Kamulete, V. M. (2021). *Test for non-negligible adverse shifts*. arXiv preprint arXiv:2107.02990.

Ciemencon, S., Depecker, M., & Vayatis, N. (2009, December). *AUC optimization and the two-sample problem*. In Proceedings of the 22nd International Conference on Neural Information Processing Systems (pp. 360-368).

Lopez-Paz, D., & Oquab, M. (2016). *Revisiting classifier two-sample tests*. arXiv preprint arXiv:1610.06545.

Friedman, J. (2004). *On multivariate goodness-of-fit and two-sample testing*.

Gandy, A. (2009). *Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk*. Journal of the American Statistical Association, 104(488), 1504-1511.

Li, J., & Fine, J. P. (2010). *Weighted area under the receiver operating characteristic curve and its application to gene selection*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(4), 673-692.

Rinaldo, A., Wasserman, L., & G'Sell, M. (2019). *Bootstrapping and sample splitting for high-dimensional, assumption-lean inference*. Annals of Statistics, 47(6), 3438-3469.

See Also

Other classifiers: cp_at(), cp_pt()

Examples

```
library(dsos)
set.seed(12345)
data(iris)
x_train <- iris[1:50, 1:4] # Training sample: Species == 'setosa'
x_test <- iris[51:100, 1:4] # Test sample: Species == 'versicolor'
iris_test <- cp_at(x_train, x_test) # Can also use: cp_ss and cp_pt
str(iris_test)
```

---

od_pt *Dataset Shift via Isolation Scores*

Description

Test for no adverse shift via isolation scores for two-sample comparison. The scores are predictions from extended isolation forest with the package **isotree**. The prefix *od* stands for outlier detection, the relevant notion of outlyingness.

Usage

```
od_pt(x_train, x_test, R = 1000, num_trees = 500, sub_ratio = 1/2)
```

Arguments

x_train	Training sample.
x_test	Test sample.
R	The number of permutations. May be ignored.
num_trees	The number of trees in random forests.
sub_ratio	Subsampling ratio for sample splitting. May be ignored.

Details

The empirical null distribution uses R permutations to estimate the p-value. For speed, this is implemented as a sequential Monte Carlo test with the **simctest** package. See Gandy (2009) for details. The suffix *pt* refers to permutation test. It does not use the asymptotic (theoretical) null distribution for the weighted AUC (WAUC), the test statistic. This is the recommended approach for small samples.

Value

A named list or object of class `outlier.test` containing:

- `statistic`: observed WAUC statistic
- `seq_mct`: sequential Monte Carlo test, if applicable
- `p_value`: p-value
- `outlier_scores`: outlier scores from training and test set

Notes

Isolation forest detects *isolated* points, instances that are typically out-of-distribution relative to the high-density regions of the data distribution. Any performant method for density-based out-of-distribution detection can replace isolation forest, the default in this implementation.

References

Kamulete, V. M. (2021). *Test for non-negligible adverse shifts*. arXiv preprint arXiv:2107.02990.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). *Isolation forest*. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413-422). IEEE.

Gandy, A. (2009). *Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk*. Journal of the American Statistical Association, 104(488), 1504-1511.

Li, J., & Fine, J. P. (2010). *Weighted area under the receiver operating characteristic curve and its application to gene selection*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(4), 673-692.

Examples

```
library(dsos)
set.seed(12345)
data(iris)
x_train <- iris[1:50, 1:4] # Training sample: Species == 'setosa'
x_test <- iris[51:100, 1:4] # Test sample: Species == 'versicolor'
iris_test <- od_pt(x_train, x_test) # Can also use: od_ss
str(iris_test)
```

---

plot.outlier.test *Plot the result of the D-SOS test.*

Description

Plot the result of the D-SOS test.

Usage

```
## S3 method for class 'outlier.test'
plot(x, ...)
```

Arguments

x	A `outlier.test` object from a D-SOS test.
...	Placeholder to be comptatible with S3 ‘plot‘ generic.

Value

A **ggplot2** plot with outlier scores and p-value.

Examples

```
set.seed(12345)
data(iris)
x_train <- iris[1:50, 1:4] # Training sample: Species == 'setosa'
x_test <- iris[51:100, 1:4] # Test sample: Species == 'versicolor'
iris_test <- od_pt(x_train, x_test)
plot(iris_test)
```

---

rd_pt *Dataset Shift via Residuals*

Description

Test for no adverse shift via residuals for multivariate two-sample comparison. The scores are obtained using out-of-bag predictions from random forest with the package **ranger** to get the residuals. The prefix *rd* stands for residual diagnostics, the relevant notion of outlier. This test assumes that both training and test sets are labeled.

Usage

```
rd_pt(
  x_train,
  x_test,
  R = 1000,
  num_trees = 500L,
  sub_ratio = 1/2,
  response_name = "label"
)
```

Arguments

x_train	Training sample.
x_test	Test sample.
R	The number of permutations. May be ignored.
num_trees	The number of trees in random forests.
sub_ratio	Subsampling ratio for sample splitting. May be ignored.
response_name	The column name of the categorical outcome to predict.

Details

The empirical null distribution uses R permutations to estimate the p-value. For speed, this is implemented as a sequential Monte Carlo test with the **simctest** package. See Gandy (2009) for details. The suffix *pt* refers to permutation test. It does not use the asymptotic (theoretical) null distribution for the weighted AUC (WAUC), the test statistic. This is the recommended approach for small samples.

Value

A named list or object of class `outlier.test` containing:

- `statistic`: observed WAUC statistic
- `seq_mct`: sequential Monte Carlo test, if applicable
- `p_value`: p-value
- `outlier_scores`: outlier scores from training and test set

Notes

Residuals traditionally underpin diagnostics (misspecification) tests in supervised learning. For a contemporaneous example of this approach also using machine learning, see see Janková et al. (2020) and references therein.

References

Kamulete, V. M. (2021). *Test for non-negligible adverse shifts*. arXiv preprint arXiv:2107.02990.

Janková, J., Shah, R. D., Bühlmann, P., & Samworth, R. J. (2020). *Goodness-of-fit testing in high dimensional generalized linear models*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(3), 773-795.

Li, J., & Fine, J. P. (2010). *Weighted area under the receiver operating characteristic curve and its application to gene selection*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(4), 673-692.

Gandy, A. (2009). *Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk*. Journal of the American Statistical Association, 104(488), 1504-1511.

Examples

```
library(dsos)
set.seed(12345)
data(iris)
idx <- sample(nrow(iris), 2 / 3 * nrow(iris))
xy_train <- iris[idx, ]
xy_test <- iris[-idx, ]
iris_test <- rd_pt(xy_train, xy_test, response_name = "Species")
str(iris_test)
```

---

rue_pt *Dataset Shift via Resampling (Prediction) Uncertainty*

Description

Test for no adverse shift via prediction uncertainty for two-sample comparison. The scores are out-of-bag predictions from random forests with the package **ranger**. The prefix *rue* stands for resampling uncertainty, the relevant notion of outlier. This uncertainty is the standard error of the mean predictions. This assumes that both training and test sets are labeled.

Usage

```
rue_pt(
  x_train,
  x_test,
  R = 1000,
  sub_ratio = 1/2,
  num_trees = 500L,
  response_name = "label"
)
```

Arguments

x_train	Training sample.
x_test	Test sample.
R	The number of permutations. May be ignored.
sub_ratio	Subsampling ratio for sample splitting. May be ignored.
num_trees	The number of trees in random forests.
response_name	The column name of the categorical outcome to predict.

Details

The empirical null distribution uses `R` permutations to estimate the p-value. For speed, this is implemented as a sequential Monte Carlo test with the **simctest** package. See Gandy (2009) for details. The suffix *pt* refers to permutation test. It does not use the asymptotic (theoretical) null distribution for the weighted AUC (WAUC), the test statistic. This is the recommended approach for small samples.

Value

A named list or object of class `outlier.test` containing:

- `statistic`: observed WAUC statistic
- `seq_mct`: sequential Monte Carlo test, if applicable
- `p_value`: p-value
- `outlier_scores`: outlier scores from training and test set

Notes

For resampling uncertainty, we essentially implement the approach in Schulam & Saria (2019) with random forests. The standard errors of the mean predictions are the underlying scores. Any performant method for confidence-based out-of-distribution detection can replace random forests, the default in this implementation.

References

Kamulete, V. M. (2021). *Test for non-negligible adverse shifts*. arXiv preprint arXiv:2107.02990.

Schulam, P., & Saria, S. (2019, April). Can you trust this prediction? Auditing pointwise reliability after learning. In The 22nd International Conference on Artificial Intelligence and Statistics (pp. 1022-1031). PMLR.

Berger, C., Paschali, M., Glocker, B., & Kamnitsas, K. (2021). Confidence-based Out-of-Distribution Detection: A Comparative Study and Analysis. arXiv preprint arXiv:2107.02568.

Gandy, A. (2009). *Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk*. Journal of the American Statistical Association, 104(488), 1504-1511.

Li, J., & Fine, J. P. (2010). *Weighted area under the receiver operating characteristic curve and its application to gene selection*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(4), 673-692.

Examples

```
library(dsos)
set.seed(12345)
data(iris)
idx <- sample(nrow(iris), 2 / 3 * nrow(iris))
xy_train <- iris[idx, ]
xy_test <- iris[-idx, ]
iris_test <- rue_pt(xy_train, xy_test, response_name = "Species")
str(iris_test)
```

wauc_from_os *Weighted AUC from Outlier Scores*

Description

Computes the weighted AUC with the weighting scheme described in Kamulete, V. M. (2021). This assumes that the training set is the reference distribution and specifies a particular functional form to derive weights from threshold scores.

Usage

```
wauc_from_os(os_train, os_test)
```

Arguments

os_train	Outlier scores in training set.
os_test	Outlier scores in test set.

Value

The value (scalar) of the weighted AUC given the weighting scheme.

References

Kamulete, V. M. (2021). *Test for non-negligible adverse shifts*. arXiv preprint arXiv:2107.02990.

Examples

```
library(dsos)
set.seed(12345)
os_train <- runif(n = 100)
os_test <- runif(n = 100)
test_stat <- wauc_from_os(os_train, os_test)
```

Index