Package ‘pranker’

May 5, 2015

Type Package

Title Significance assessment of precision and recall curves

Version 1.0

Date 2015-05-01

Author Miguel Lopes

Maintainer Miguel Lopes <miguelaglopes@gmail.com>

Description Computes the average precision of a binary classification, and returns a p-value on the null hypothesis of a random classifier.

License GPL3

Imports Rcpp (>= 0.10.3)

LinkingTo Rcpp

URL https://github.com/miguelaglopes/pranker

Archs i386, x86_64

R topics documented:

`pranker-package` *Pranker - R package for significancy assessment of the area under the precision and recall curve (AUPRC)*

Description

A common problem in information retrieval and machine learning in general is to select positive instances (P) out of a pool of both positive and non-positive instances (N). In this problem, instances are typically scored and ranked (higher scored instances are selected first). When the gold standard is available, selection rankings are commonly assessed with the area under the precision recall curve (AUPRC), plotting precision as a function of recall. This area can be approximated as the average maximum precision for all possible values of recall. This particular version of the AUPRC is adopted here.

Ideally, positive instances are selected first, resulting in a maximum AUPRC (area under the curve) of 1. This package computes the AUPRC and assigns a p-value to it, relative to the null hypothesis of random selection.

The null (relative to random) AUPRC distribution is useful to estimate AUPRC significancy and to compare AUPRC values obtained in different configurations of N and P (as the null distribution depends on these parameters).

In this implementation, the variance and covariance of the null AUPRC is computed and used to fit a beta distribution approximating the true null AUPRC distribution. When the number of instances is low (P < few dozens) the approximation exhibits some bias. However, as the number of instances increases the approximation becomes more accurate. On the other hand, when the number of instances is very high (> few hundreds of thousands), the computation of the AUPRC variance may be too computationally intensive. Between these extremes, this package is a useful alternative to Monte Carlo simulations. An approximation to speed up the AUPRC covariance computation is implemented (described in detail in the documentation for the function "pranker").

The main function of this package is "pranker", which takes as input a vector of scores of instances (the instances with the highest scores are selected first), and a gold standard vector (non-zero elements represent positive instances). The area under the precision recall curve and the respective p-value are then computed and returned. The function "pranker" calls the functions "auprc.ap" (which computes the AUPRC) and "null.params", which computes the mean and variance of the null AUPRC distribution, for the number of instances (N) and positives (P) of the gold standard. The functions "pranker", "auprc.ap" and "null.params" are described in more detail in the associated documentation.

The methods implemented in this package are described in "On the null distribution of the precision-recall curve", Lopes and Bontempi, ECML KDD 2014.

Details

Package:	pranker
Type:	Package
Version:	1.0
Date:	2015-05-04
License:	GPL3

Author(s)

Miguel Lopes Maintainer: Miguel Lopes <miguelaglopes@gmail.com>

References

On the null distribution of the precision-recall curve, Lopes and Bontempi, ECML KDD 2014

See Also

pranker null.params auprc.ap

---

auprc.ap *auprc.ap - AUPRC score*

Description

This function computes the AUPRC (average maximum precision) for a vector of scores and gold standard. Scores vector and gold standard must be of the same size. Elements in the score vector are incrementally selected (from the highest to lowest vector) and precision and recall points are computed. The AUPRC is computed as the average maximum precision for all points of recall (there may be multiple values of precision associated with each recall).

Usage

```
auprc.ap(scores, y)
```

Arguments

scores	Numeric vector or matrix of scores. No NA or NULL values allowed.
y	The gold standard (numeric vector or matrix, of the same size as "scores"). Non-zero elements represent positive instances. No NA or NULL values allowed.

Value

The AUPRC (average maximum precision) of the score vector.

Author(s)

Miguel Lopes

References

On the null distribution of the precision-recall curve, Lopes and Bontempi, ECML KDD 2014

See Also

prankerpackage pranker null.params

`null.params` *null.params - mean and variance of the AUPRC (average maximum precision) distribution.*

Description

This function computes the mean and variance of the null (of random selection) AUPRC (average maximum precision), for a given number of total instances (N) and number of positive instances (P). An approximation may be implemented, which is described in the documentation for the function "pranker".

Usage

```
null.params(N, P, approx, approxN)
```

Arguments

`N`	Integer, number of total instances.
`P`	Integer, number of positive instances.
`approx`	Logical, TRUE (default) or FALSE. If TRUE, the covariance spline approximation is used.
`approxN`	Integer, the parameter of the covariance spline approximation (see above). Default is P/10.

Value

Vector of two numeric elements, the mean (the first) and the variance (the second) of the null AUPRC.

Author(s)

Miguel Lopes

References

On the null distribution of the precision-recall curve, Lopes and Bontempi, ECML KDD 2014

See Also

`prankerpackage` `pranker` `auprc.ap`

pranker *pranker - significancy assessment of the area under the precision-recall curve.*

Description

This is the main function of the package (see the documentation on prankerpackage). It takes as input a vector of instance scores of an information retrieval task (higher scores are selected first) and a gold standard vector (non-zero elements represent positive instances). N represents the number of instances and P the number of positives.

This function returns a list of elements: the first is the obtained AUPRC; the second is a p-value relative to the null hypothesis of random selection; the third is a numeric vector of the mean and variance of the null AUPRC; the fourth is the number of total and positive instances.

The null AUPRC distribution is obtained by first computing its mean and variance, and then by fitting a beta distribution to these two parameters, additionally to the maximum and the minimum of the distribution (note that while the maximum is 1, the minimum is not). A p-value is then obtained.

When N and P are large, the AUPRC variance may be too intensive to compute. It requires the covariance matrix of the maximum precision at different points of recall, and an approximation is implemented which skips elements in each row/column in this matrix. These are then interpolated with splines. The optional parameter "approx" defines whether this approximation is used, and the parameter "approxN" (integer) defines the quantity of skipped elements. In particular, only one out of consecutive approxN elements is computed (plus the first and the last in each row of the upper diagonal of the covariance matrix, plus the diagonal).

Note that if they are available, the mean and variance of the null AUPRC may be given as input (as a numeric vector). These parameters may be computed using the function "null.params" (which takes as input the number of instances and number of positives). If only the AUPRC value is necessary (and not the p-value), it can be computed using the function "auprc.ap".

Usage

```
pranker(scores, y, params, approx, approxN)
```

Arguments

`scores` Numeric vector or matrix of scores. No NA or NULL values allowed.

`y` The gold standard (numeric vector or matrix, of the same size as "prediction"). Non-zero elements represent positive instances. No NA or NULL values allowed.

`params` (optional) Vector of two numeric elements, the mean (the first) and the variance (the second) of the null AUPRC. If these are given there is no need to compute them. These may be obtained in the function null.params.

`approx` (optional) Logical, TRUE (default) or FALSE. If TRUE, the covariance spline approximation is used (see function null.params).

`approxN` (optional) Integer, the parameter of the covariance spline approximation (see above). Default is P/10 (see function null.params).

Value

auprc	AUPRC value
pvalue	AUPRC p-value
nullparams	Mean and variance of the AUPRC null distribution
instances	Number of total instances and number of positive instances

Author(s)

Miguel Lopes

References

On the null distribution of the precision-recall curve, Lopes and Bontempi, ECML KDD 2014

See Also

prankerpackage null.params auprc.ap

Examples

```
# scores=abs(rnorm(100,0,1)) # generate scores for 100 instances
# test=numeric(100) # generate gold standard (the first 10 are positive)
# test[1:10]=1 # run pranker
# pranker(scores,test)
```

Index