Package 'multiselect'

January 25, 2018

Type Package

Title Selecting Combinations of Predictors by Leveraging Multiple AUCs for an Ordered Multilevel Outcome

Version 0.1.0

Date 2018-01-24

Author Allison Meisner

Maintainer Allison Meisner <allison.meisner@gmail.com>

Description Uses multiple AUCs to select a combination of predictors when the outcome has multiple (ordered) levels and the focus is discriminating one particular level from the others. This method is most naturally applied to settings where the outcome has three levels. (Meisner, A, Parikh, CR, and Kerr, KF (2017) <http://biostats.bepress.com/uwbiostat/paper423/>.)

License GPL-2

LazyData TRUE

Imports Hmisc

Suggests MASS

NeedsCompilation no

Repository CRAN

Date/Publication 2018-01-25 15:36:25 UTC

R topics documented:

`multiselect` *Selecting Combinations of Predictors by Leveraging Multiple AUCs for an Ordered Multilevel Outcome*

Description

When several predictors are available, there is often interest in combining a subset of predictors to diagnose disease or predict risk of a clinical outcome, D. In the context of an ordered outcome with K levels, where interest is in predicting $D = K$, there are multiple ways to select a combination. The traditional approach involves dichotomizing the outcome and using logistic regression to construct the combinations, then selecting a combination based on the estimated AUC for $D = K$ vs. $D < K$ for each fitted combination. An alternative approach, implemented here, constructs the combinations in the same way, but uses both the AUC for $D = K$ vs. $D < K$ and the AUC for $D = K - 1$ vs. $D < K - 1$. The combination with the best combined performance is then chosen. This function provides (i) the best combination defined solely by the AUC for $D = K$ vs. $D < K$ and (ii) the best combination defined by both the AUC for $D = K$ vs. $D < K$ and the AUC for $D = K - 1$ vs. $D < K - 1$. In the context where D indicates no, mild, or severe disease (K=3), this is equivalent to (i) selecting a combination in terms of its ability to discriminate between individuals with severe vs. no or mild disease and (ii) selecting a combination in terms of its ability to discriminate between individuals with severe vs. no or mild disease and its ability to discriminate between individuals with mild vs. no disease.

Usage

```
multiselect(data, size=2, Breps=40, nummod=10)
```

Arguments

`data`	The name of the dataset to be used. An object of class 'data.frame' where the first column is the outcome, and the subsequent columns are the predictors. All columns must be numeric. The outcome must be take values 1,...,K, where $K \geq 3$. Missing observations are not allowed. If the columns of `data` are not named, the outcome (first column) will be named "D", and the predictors (subsequent columns) will be named "V1","V2",....
`size`	The size of the combinations. The function considers all possible subsets of the predictors of size `size`. Default 2 (all possible pairs).
`Breps`	The number of bootstrap replicates used to estimate the optimism due to resubstitution bias in the AUCs. For each combination, the function estimates the apparent AUCs for each fitted combination. These apparent AUCs are then corrected by substracting the optimism due to resubstitution bias, which is estimated using a bootstrap procedure. Default 40.
`nummod`	The number of predictor combinations to return. Using the optimism-corrected estimate of the AUC for $D = K$ vs. $D < K$, the function returns the top `nummod` predictor combinations. Default 10.

Details

For each possible predictor combination of size `size`, the function fits the predictor combination using logistic regression comparing outcome $D = K$ to $D < K$. The apparent AUCs for (a) $D = K$ vs. $D < K$ and (b) $D = K - 1$ vs. $D < K - 1$ are calculated. A bootstrapping procedure is then used to estimate the optimism due to resubstitution bias in these apparent AUCs. The AUCs are corrected by subtracting the estimated optimism due to resubstitution bias. Two combinations are then selected: the combination with the highest AUC for $D = K$ vs. $D < K$ ("single AUC" approach) and the combination with the best sum of ranks for the AUC for $D = K$ vs. $D < K$ and the AUC for $D = K - 1$ vs. $D < K - 1$ ("multi-AUC" approach). The selected combinations may be the same for the two approaches. The top `nummod` combinations, in terms of the AUC for $D = K$ vs. $D < K$ (corrected for optimism due to resubstitution bias), are also provided.

If more than one combination is "best" in terms of either the AUC for $D = K$ vs. $D < K$ or the sum of ranks for the AUC for $D = K$ vs. $D < K$ and the AUC for $D = K - 1$ vs. $D < K - 1$ (i.e., in the event of ties) the first combination is returned. The order of the combinations for p candidate predictors is given by `combn(1:p, size)`. If ties occur for either (i) the AUC for $D = K$ vs. $D < K$ or (ii) the sum of ranks for the AUC for $D = K$ vs. $D < K$ and the AUC for $D = K - 1$ vs. $D < K - 1$, a warning is given.

A given bootstrap sample may not have observations from each of the K outcome levels; if this occurs, a warning is given and the estimated optimism for that bootstrap sample for both the AUC for $D = K$ vs. $D < K$ and the AUC for $D = K - 1$ vs. $D < K - 1$ will be NA. NAs are removed in the calculation of the mean optimism (used to correct the AUC estimates for resubstitution bias), and the total number of NAs across the `Breps` (for either the AUC for $D = K$ vs. $D < K$ or the AUC for $D = K - 1$ vs. $D < K - 1$) is indicated by "numNA" in the output.

Value

A list with the following components:

`Best.Single` The best predictor combination as chosen by the "single AUC" approach. The first `size` elements give the names of the included predictors (under "Var1","Var2",...), the next is the estimated AUC $D = K$ vs. $D < K$ ("AUC1"), the next is the estimated AUC for $D = K - 1$ vs. $D < K - 1$ ("AUC2"), the next is number of NAs across the `Brep` bootstrap replicates ("numNA"; see 'Details'), and the final `size` elements give the estimated coefficients for each of the included predictors ("Coef1","Coef2",...). Both AUCs are corrected for optimism due to resubstitution bias. Recall that if the columns of `data` are unnamed, the predictors will be named "V1","V2",....

`Best.Multi` The best predictor combination as chosen by the "multi-AUC" approach. The elements of `Best.Multi` are the same as `Best.Single`.

`Ranked.Rslts` The results for the `nummod` best combinations, as ranked by the AUC for $D = K$ vs. $D < K$ (after correcting for optimism due to resubstitution bias). The columns are the same as the elements of `Best.Single` and `Best.Multi`.

References

Meisner, A., Parikh, C.R., and Kerr, K.F. (2017). Using multilevel outcomes to construct and select biomarker combinations for single-level prediction. UW Biostatistics Working Paper Series, Working Paper 423.

Examples

```
library(MASS)
## example takes ~1 minute to run

set.seed(15)
p = 16 ## number of predictors
matX <- matrix(rep(0.3,p*p), nrow=p, ncol=p) ## covariance matrix for the predictors
diag(matX) <- rep(1,p)

simD <- apply(rmultinom(400, 1, c(0.6,0.335,0.065)),2,which.max)
simDord <- simD[order(simD)]
numobs <- table(simDord)

simX1 <- mvrnorm(numobs[1], rep(0,p), 2*matX)
simX2 <- mvrnorm(numobs[2], c(1.5, 1, rep(0.5,(p-2)/2), rep(0.1,(p-2)/2)), 2*matX)
simX3 <- mvrnorm(numobs[3], c(rep(2,2), rep(0.8,(p-2)/2), rep(0.1,(p-2)/2)), 2*matX)
simX <- rbind(simX1, simX2, simX3)

exdata <- data.frame("D"=simDord, simX)

multiselect(data=exdata, size=2, Breps=20, nummod=10)
```

Index