Cross-validated AUC in Stata: CVAUROC

ibs.GRANADA
INSTITUTO DE INVESTIGACIÓN BIOSANITARIA

Miguel Angel Luque Fernandez
Biomedical Research Institute of Granada
Noncommunicable Disease and Cancer Epidemiology
https://maluque.netlify.com

2018 Spanish Stata Conference

24 October 2018

Contents

Cross-validation

Definition

- Cross-validation is a **model validation technique** for assessing how the results of a statistical analysis will generalize to an independent data set.
- It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice (note: performance = **model assessment**).

Cross-validation

Definition

- Cross-validation is a **model validation technique** for assessing how the results of a statistical analysis will generalize to an independent data set.
- It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice (note: performance = **model assessment**).

Cross-validation

Applications

- However, cross-validation can be used to compare the **performance** of different modeling specifications (i.e. models with and without interactions, inclusion of exclusion of polynomial terms, number of knots with restricted cubic splines, etc).
- Furthermore, cross-validation can be used in **variable selection** and select the suitable level of flexibility in the model (note: flexibility = **model selection**).

Cross-validation

Applications

- However, cross-validation can be used to compare the **performance** of different modeling specifications (i.e. models with and without interactions, inclusion of exclusion of polynomial terms, number of knots with restricted cubic splines, etc).
- Furthermore, cross-validation can be used in **variable selection** and select the suitable level of flexibility in the model (note: flexibility = **model selection**).

Cross-validation

Applications

- MODEL ASSESSMENT: To **compare** the performance of different modeling specifications.
- MODEL SELECTION: To **select** the suitable level of flexibility in the model.

Cross-validation

Applications

- MODEL ASSESSMENT: To **compare** the performance of different modeling specifications.
- MODEL SELECTION: To **select** the suitable level of flexibility in the model.

MSE

Regression Model

$$f(x) = f(x_1 + x_2 + x_3)$$

$$Y = \beta x_1 + \beta x_2 + \beta x_3 + \epsilon$$

MSE

Regression Model

$$f(x) = f(x_1 + x_2 + x_3)$$

$$Y = \beta x_1 + \beta x_2 + \beta x_3 + \epsilon$$

$$Y = f(x) + \epsilon$$

MSE

Regression Model

$$f(x) = f(x_1 + x_2 + x_3)$$

$$Y = \beta x_1 + \beta x_2 + \beta x_3 + \epsilon$$

$$Y = f(x) + \epsilon$$

MSE

Expectation

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

MSE

$$E[(Y - \hat{f}(X))^2|X = x]$$

MSE

Expectation

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

MSE

$$E[(Y - \hat{f}(X))^2|X = x]$$

Bias-Variance Trade-off

Error descomposition

$$MSE = E[(Y - \hat{f}(X))^2 | X = x] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

Trade-off

As flexibility of $\hat{f}$ increases, its variance increases, and its bias decreases.

Bias-variance trade-off

Choosing the model flexibility based on average test error

Average Test Error

$$E[(Y - \hat{f}(X))^2 | X = x]$$

And thus, this amounts to a bias-variance trade-off.

BIAS-VARIANCE-TRADE-OFF

Bias-variance trade-off

Choosing the model flexibility based on average test error

Average Test Error

$$E[(Y - \hat{f}(X))^2 | X = x]$$

And thus, this amounts to a bias-variance trade-off.

Rule

- More flexibility increases variance but decreases bias.
- Less flexibility decreases variance but increases error.

Bias-variance trade-off

Choosing the model flexibility based on average test error

Average Test Error

$$E[(Y - \hat{f}(X))^2 | X = x]$$

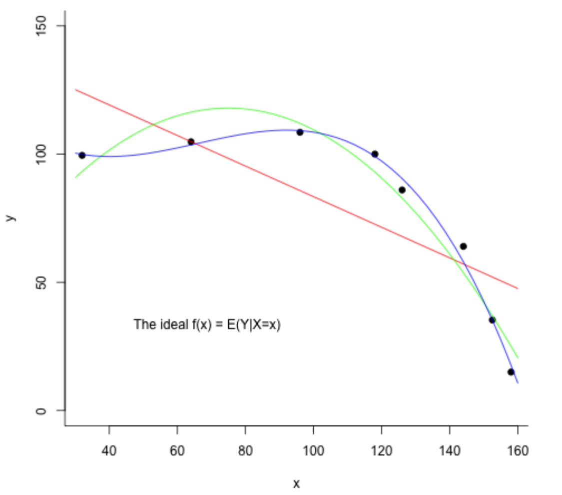
And thus, this amounts to a bias-variance trade-off.

Rule

- More flexibility increases variance but decreases bias.
- Less flexibility decreases variance but increases error.

Bias-Variance trade-off

Regression Function



Overparameterization

George E.P.Box,(1919-2013)

All models are wrong but some are useful

Quote, 1976

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration (...). Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

Overparameterization

George E.P.Box,(1919-2013)

All models are wrong but some are useful

Quote, 1976

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration (...). Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

Justification

AIC and BIC

- AIC and BIC are both maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior.

- AIC = -2*ln(likelihood) + 2*k, k = model degrees of freedom

Justification

AIC and BIC

- AIC and BIC are both maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior.
- AIC = -2*ln(likelihood) + 2*k, k = model degrees of freedom
- BIC = -2*ln(likelihood) + ln(N)*k, k = model degrees of freedom and N = number of observations.

Justification

AIC and BIC

- AIC and BIC are both maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior.
- AIC = -2*ln(likelihood) + 2*k, k = model degrees of freedom
- BIC = -2*ln(likelihood) + ln(N)*k, k = model degrees of freedom and N = number of observations.
- There is some disagreement over the use of AIC and BIC with non-nested models.

Justification

AIC and BIC

- AIC and BIC are both maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior.
- AIC = -2*ln(likelihood) + 2*k, k = model degrees of freedom
- BIC = -2*ln(likelihood) + ln(N)*k, k = model degrees of freedom and N = number of observations.
- There is some disagreement over the use of AIC and BIC with non-nested models.

Justification

Fewer assumptions

- Cross-validation compared with AIC, BIC and adjusted R^2 provides a direct estimate of the **ERROR**.
- Cross-validation makes fewer assumptions about the true underlying model.

Justification

Fewer assumptions

- Cross-validation compared with AIC, BIC and adjusted R^2 provides a direct estimate of the **ERROR**.
- Cross-validation makes fewer assumptions about the true underlying model.
- Cross-validation can be used in a wider range of model selections tasks, even in cases where it is hard to pinpoint the number of predictors in the model.

Justification

Fewer assumptions

- Cross-validation compared with AIC, BIC and adjusted R^2 provides a direct estimate of the **ERROR**.
- Cross-validation makes fewer assumptions about the true underlying model.
- Cross-validation can be used in a wider range of model selections tasks, even in cases where it is hard to pinpoint the number of predictors in the model.

Cross-validation strategies

Cross-validation options

- Leave-one-out cross-validation (LOOCV).
- k-fold cross validation.

Cross-validation strategies

Cross-validation options

- Leave-one-out cross-validation (LOOCV).
- k-fold cross validation.
- Bootstraping.

Cross-validation strategies

Cross-validation options

- Leave-one-out cross-validation (LOOCV).
- k-fold cross validation.
- Bootstraping.

K-fold Cross-validation

K-fold

- Technique widely used for estimating the test error.
- Estimates can be used to select the best model, and to give an idea of the test error of the final chosen model.

K-fold Cross-validation

K-fold

- Technique widely used for estimating the test error.
- Estimates can be used to select the best model, and to give an idea of the test error of the final chosen model.
- The idea is to randmoly divide the data into k equal-sized parts. We leave out part k, fit the model to the other k-1 parts (combined), and then obtain predictions for the left-out kth part.

K-fold Cross-validation

K-fold

- Technique widely used for estimating the test error.
- Estimates can be used to select the best model, and to give an idea of the test error of the final chosen model.
- The idea is to randmoly divide the data into k equal-sized parts. We leave out part k, fit the model to the other k-1 parts (combined), and then obtain predictions for the left-out kth part.

K-fold Cross-validation

K-fold

$$\text{CV} = \sum_{k-1}^{k} \frac{n_k}{n} MSE_k$$

$$\text{MSE}_k = \sum_{i \in C_k} (y_i - (\hat{y}_i))/n_k$$

Seeting K = n yields n-fold or leave-one-out cross-validation (LOOCV)

K-fold Cross-validation

K-fold

$$\text{CV} = \sum_{k-1}^{k} \frac{n_k}{n} MSE_k$$

$$\text{MSE}_k = \sum_{i \in C_k} (y_i - (\hat{y}_i))/n_k$$

Seeting K = n yields n-fold or leave-one-out cross-validation (LOOCV)

Model performance: Internal Validation (AUC)

AUC

- The AUC is a global summary measure of a diagnostic test **accuracy** and **discrimination.** The greater the AUC, the more able is the test to capture the trade-off between Se and Sp over a continuous range.

- An important aspect of predictive modeling is the ability of a model to generalize to new cases.

Model performance: Internal Validation (AUC)

AUC

- The AUC is a global summary measure of a diagnostic test **accuracy** and **discrimination.** The greater the AUC, the more able is the test to capture the trade-off between Se and Sp over a continuous range.
- An important aspect of predictive modeling is the ability of a model to generalize to new cases.

Model performance: Internal Validation (AUC)

AUC

- The AUC is a global summary measure of a diagnostic test **accuracy** and **discrimination.** The greater the AUC, the more able is the test to capture the trade-off between Se and Sp over a continuous range.
- An important aspect of predictive modeling is the ability of a model to generalize to new cases.

Predictive performance: internal validation

AUC

- Evaluating the predictive performance (AUC) of a set of independent variables using all cases from the original analysis sample tends to result in an overly optimistic estimate of predictive performance.
- **K-fold cross-validation** can be used to generate a more realistic estimate of predictive performance when the number of observations is not very large (Ledell, 2015).

Predictive performance: internal validation

AUC

- Evaluating the predictive performance (AUC) of a set of independent variables using all cases from the original analysis sample tends to result in an overly optimistic estimate of predictive performance.

- **K-fold cross-validation** can be used to generate a more realistic estimate of predictive performance when the number of observations is not very large (Ledell, 2015).

Predictive performance: internal validation

AUC

- Evaluating the predictive performance (AUC) of a set of independent variables using all cases from the original analysis sample tends to result in an overly optimistic estimate of predictive performance.
- **K-fold cross-validation** can be used to generate a more realistic estimate of predictive performance when the number of observations is not very large (Ledell, 2015).

CVAUROC

cvauroc

cvauroc implements k-fold cross-validation for the AUC for a binary outcome after fitting a logistic regression model and provides the cross-validated fitted probabilities for the dependent variable or outcome, contained in a new variable named **fit**.

GitHub cvauroc development version

https://github.com/migariane/cvAUROC

CVAUROC

cvauroc

cvauroc implements k-fold cross-validation for the AUC for a binary outcome after fitting a logistic regression model and provides the cross-validated fitted probabilities for the dependent variable or outcome, contained in a new variable named **fit**.

GitHub cvauroc development version

https://github.com/migariane/cvAUROC

Stata ssc

ssc install cvAUROC

CVAUROC

cvauroc

cvauroc implements k-fold cross-validation for the AUC for a binary outcome after fitting a logistic regression model and provides the cross-validated fitted probabilities for the dependent variable or outcome, contained in a new variable named **fit**.

GitHub cvauroc development version

https://github.com/migariane/cvAUROC

Stata ssc

ssc install cvAUROC

CVAUROC

cvauroc

cvauroc implements k-fold cross-validation for the AUC for a binary outcome after fitting a logistic regression model and provides the cross-validated fitted probabilities for the dependent variable or outcome, contained in a new variable named **fit**.

GitHub cvauroc development version

https://github.com/migariane/cvAUROC

Stata ssc

ssc install cvAUROC

cvauroc Stata command syntax

cvauroc Syntax

cvauroc depvar varlist [if] [pw] [Kfold] [Seed] [, Cluster(varname) Detail Graph]

Classical AUC estimation

```
. use http://www.stata-press.com/data/r14/cattaneo2.dta
. gen lbw = cond(bweight<2500,1,0.)
. logistic lbw mage medu mmarried prenatal fedu mbsmoke mrace order
Logistic regression                             Number of obs     =      4,642
---------------------------------------
         lbw | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+------------------------------
        mage |   .9959165   .0140441    -0.29   0.772     .9687674    1.023826
        medu |   .9451338   .0283732    -1.88   0.060     .8911276    1.002413
    mmarried |   .6109995   .1014788    -2.97   0.003     .4412328    .8460849
    prenatal |   .5886787    .073186    -4.26   0.000     .4613759    .7511069
        fedu |   1.040936   .0214226     1.95   0.051     .9997838    1.083782
     mbsmoke |   2.145619   .3055361     5.36   0.000     1.623086    2.836376
       mrace |   .3789501    .057913    -6.35   0.000     .2808648    .5112895
       order |    1.05529   .0605811     0.94   0.349     .9429895    1.180964
---------------------------------------
. predict fitted, pr
. roctab lbw fitted
                      ROC                    -Asymptotic Normal-
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------
         4,642     0.6939       0.0171        0.66041     0.72749
```

Crossvalidated AUC using cvauroc

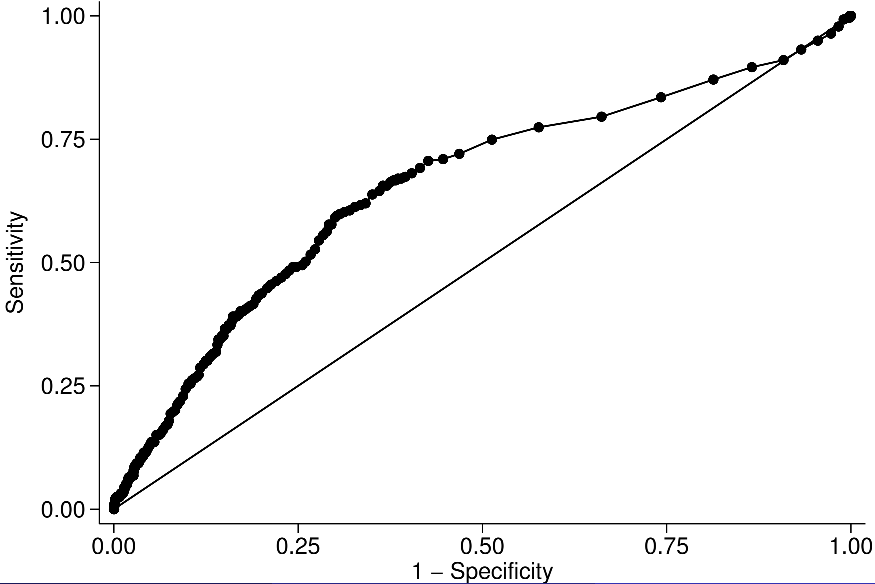
```
. cvauroc lbw mage medu mmarried prenatal fedu mbsmoke mrace order,
kfold(10) seed(12)
1-fold..............................
2-fold..............................
3-fold..............................
4-fold..............................
5-fold..............................
6-fold..............................
7-fold..............................
8-fold..............................
9-fold..............................
10-fold..............................
Random seed: 12
                      ROC                    -Asymptotic Normal-
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------
         4,642     0.6826       0.0174        0.64842      0.71668
```

cvauroc detail and graph options

```
// Using detail option to show the table of cutoff values and their respective Se,
// and likelihood ratio values.
. cvAUROC lbw mage medu mmarried prenatal1 fedu mbsmoke mrace fbaby,
kfold(10) seed(3489) detail

Detailed report of sensitivity and specificity
----------------------------------------
Correctly
Cutpoint      Sensitivity   Specificity   Classified          LR+          LR-
----------------------------------------
( >= .019 )       100.00%         0.00%        6.01%       1.0000
( >= .025 )        99.64%         0.18%        6.16%       0.9982       1.9547
( >= .026 )        99.64%         0.39%        6.36%       1.0003       0.9199
(...) Omitted results
( >= .272 )         1.08%        99.93%       93.99%      15.6389       0.9899
( >= .273 )         0.72%        99.93%       93.97%      10.4259       0.9935
( >= .300 )         0.36%        99.95%       93.97%       7.8181       0.9969
----------------------------------------
// Using the "graph" option to display the ROC curve
. cvAUROC lbw mage medu mmarried prenatal1 fedu mbsmoke mrace fbaby,
kfold(10) seed(3489) graph
. graph export "your_path/Figure1.eps", as(eps) preview(off)
```

cvauroc: Cross-validated AUC



Conclusion

cvauroc

- Evaluating the predictive performance of a set of independent variables using all cases from the original analysis sample tends to result in an overly optimistic estimate of predictive performance.

- However, **cvauroc** is user-friendly and helpful k-fold internal cross-validation technique that might be considered when reporting the AUC in observational studies.

Conclusion

cvauroc

- Evaluating the predictive performance of a set of independent variables using all cases from the original analysis sample tends to result in an overly optimistic estimate of predictive performance.

- However, **cvauroc** is user-friendly and helpful k-fold internal cross-validation technique that might be considered when reporting the AUC in observational studies.

cvauroc

- Evaluating the predictive performance of a set of independent variables using all cases from the original analysis sample tends to result in an overly optimistic estimate of predictive performance.

- However, **cvauroc** is user-friendly and helpful k-fold internal cross-validation technique that might be considered when reporting the AUC in observational studies.

References

Statistics Surveys
Vol. 4 (2010) 40–79
ISSN: 1935-7516
DOI: 10.1214/09-SS054

A survey of cross-validation procedures for model selection*

Sylvain Arlot†

CNRS; Willow Project-Team,
Laboratoire d'Informatique de l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
23 avenue d'Italie, F-75214 Paris Cedex 13, France
e-mail: sylvain.arlot@ens.fr

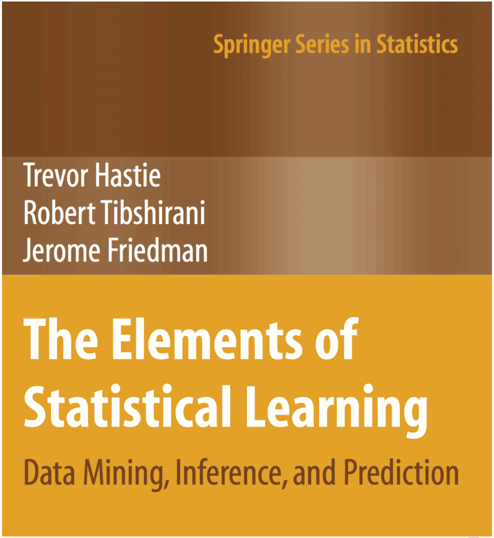
and

Alain Celisse†

Laboratoire de Mathématique Paul Painlevé
UMR 8524 CNRS - Université Lille 1,
59 655 Villeneuve d'Ascq Cedex, France
e-mail: alain.celisse@math.univ-lille1.fr

Abstract: Used to estimate the risk of an estimator or to perform model selection, cross-validation is a widespread strategy because of its simplicity and its (apparent) universality. Many results exist on model selection performances of cross-validation procedures. This survey intends to relate these results to the most recent advances of model selection theory, with a particular emphasis on distinguishing empirical statements from rigorous theoretical results. As a conclusion, guidelines are provided for choosing the best cross-validation procedure according to the particular features of the problem in hand.

References


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of
Statistical Learning
Data Mining, Inference, and Prediction

Thank you

THANK YOU FOR YOUR TIME