h2omlestat confmatrix — Display confusion matrix[+]

[+]This command includes features that are part of StataNow.

Description

`h2omlestat confmatrix` displays a confusion matrix after binary or multiclass classification performed by `h2oml gbbinclass`, `h2oml rfbinclass`, `h2oml gbmulticlass`, or `h2oml rfmulticlass`. A confusion matrix is a summary table for the prediction performance of a machine learning classification model. It displays how different observations are classified based on correct and incorrect predictions. It provides a more informative breakdown of a model's performance than a single metric.

Quick start

Display the confusion matrix after classification

```
h2omlestat confmatrix
```

As above, but report confusion matrix based on a validation set

```
h2omlestat confmatrix, valid
```

As above, but use a threshold value of 0.5 to determine negative versus positive predicted classes

```
h2omlestat confmatrix, valid threshold(0.5)
```

Menu

Statistics > H2O machine learning

Syntax

```
h2omlestat confmatrix [ , options ]
```

options	Description
Main	
metric(*metric*)	specify the metric to be used to select the optimal threshold after binary classification
threshold(#)	specify the threshold value for the predicted probabilities after binary classification
Reporting	
title(*string*)	specify the title to be displayed above the table
labels(*lnames*)	specify label names for rows and columns
nototals	suppress row and column totals
norowtotals	suppress row totals
nocoltotals	suppress column totals
noerrors	suppress the error column
norate	suppress the rate column
train	specify that the confusion matrix be reported using training results
valid	specify that the confusion matrix be reported using validation results
cv	specify that the confusion matrix be reported using cross-validation results
test	specify that the confusion matrix be computed using the testing frame
test(*framename*)	specify that the confusion matrix be computed using data in testing frame *framename*
frame(*framename*)	specify that the confusion matrix be computed using data in H2O frame *framename*
framelabel(*string*)	label frame as *string* in the output

collect is allowed; see **[U] 11.1.10 Prefix commands**.

train, valid, cv, test, test(), frame(), and framelabel() do not appear in the dialog box.

Options

Main

metric(*metric*) specifies the classification metric to be used for selecting a threshold value. This option is valid only after binary classification. *metric* can be one of f1 (the default), f2, fhalf, accuracy, precision, recall, specificity, minclassaccuracy, meanclassaccuracy, tn, fn, tp, fp, tnr, fnr, tpr, fpr, or mcc. For definitions, see [H2OML] ***metric_option***. Only one of metric() or threshold() is allowed.

threshold(#) specifies the cutpoint for the predicted probabilities after binary classification. The specified # must be a value between 0 and 1. Observations with a predicted probability greater than the specified threshold() will be classified as "positive", and the remaining observations will be classified as "negative". By default, the selected threshold value maximizes the F1 score. The list of threshold values for which threshold-based metrics are computed corresponds to the predicted probabilities of the positive class (the positive class is the largest numeric value, such as 1 in a 0/1 coded

variable, or the second label in lexicographical order). If the specified `threshold(#)` is not in the list of predicted probabilities, a result based on the closest threshold value is reported. Only one of `threshold()` or `metric()` is allowed.

Reporting

`title(`*string*`)` specifies the title to be displayed above the table.

`labels(`*lnames*`)` specifies the label names for rows and columns. By default, label names show the class names of the categorical response variable. The specified number of labels must be equal to the number of classes of the categorical response variable. The specified labels should be separated by spaces. If the label itself contains spaces, it must be enclosed with double quotes.

`nototals` suppresses the totals for rows and columns. `nototals` is not allowed with `norowtotals` or `nocoltotals`.

`norowtotals` suppresses the totals for rows. `norowtotals` is not allowed with `nototals`.

`nocoltotals` suppresses the totals for columns. `nocoltotals` is not allowed with `nototals`.

`noerrors` suppresses the error column.

`norate` suppresses the rate column.

The following options are available with `h2omlestat confmatrix` but are not shown in the dialog box:

`train`, `valid`, `cv`, `test`, `test()`, and `frame()` specify the H2O frame for which the confusion matrix is reported. Only one of `train`, `valid`, `cv`, `test`, `test()`, or `frame()` is allowed.

`train` specifies that the confusion matrix be reported using training results. This is the default when neither validation nor cross-validation is performed during estimation and when a postestimation frame has not been set with `h2omlpostestframe`.

`valid` specifies that the confusion matrix be reported using validation results. This is the default when validation is performed during estimation and when a postestimation frame has not been set with `h2omlpostestframe`. `valid` may be specified only when the `validframe()` option is specified with `h2oml` *gbm* or `h2oml` *rf*.

`cv` specifies that the confusion matrix be reported using cross-validation results. This is the default when cross-validation is performed during estimation and when a postestimation frame has not been set with `h2omlpostestframe`. `cv` may be specified only when the `cv` or `cv()` option is specified with `h2oml` *gbm* or `h2oml` *rf*.

`test` specifies that the confusion matrix be computed on the testing frame specified with `h2oml postestframe`. This is the default when a testing frame is specified with `h2omlpostestframe`. `test` may be specified only after a testing frame is set with `h2omlpostestframe`. `test` is necessary only when a subsequent `h2omlpostestframe` command is used to set a default postestimation frame other than the testing frame.

`test(`*framename*`)` specifies that the confusion matrix be computed using data in testing frame *framename* and is rarely used. This option is most useful when running a single postestimation command on the named frame. If multiple postestimation commands are to be run on the same test frame, `h2omlpostestframe` provides a more convenient and computationally efficient process for doing this.

`frame(`*framename*`)` specifies that the confusion matrix be computed using the data in H2O frame *framename*.

framelabel(*string*) specifies the label to be used for the frame in the output. This option is not allowed with the cv option.

stata.com

Remarks and examples

A confusion matrix is a popular tool for assessing model performance for classification. It consists of a simple grid that contains information about the model's performance in terms of correct and incorrect predictions. A confusion matrix summarizes the types of errors the model makes and allows you to determine areas in which the model predictions can be improved.

Below is an example of a confusion matrix where we predict the origin of a car to be either Domestic or Foreign. Rows of the confusion matrix correspond to the actual classes, and columns correspond to predicted classes. In H2O, a "positive" class corresponds to a class that contains 1, True, or the second label in lexicographical order. In our case, the positive class corresponds to the car origin being Foreign.

```
. h2omlestat confmatrix
Confusion matrix using H2O
Training frame: train
```

	Predicted				
foreign	Domestic	Foreign	Total	Error	Rate
Domestic	37	8	45	8	.178
Foreign	0	18	18	0	0
Total	37	26	63	8	.127

```
Note: Probability threshold .2083 that maximizes F1
      metric used for classification.
```

In this example, the 37 in the upper left cell indicates that there are 37 observations for which the actual class is Domestic and the model correctly predicts this class. Because Domestic is treated as a "negative" class in this example, the result in this cell is also known as the number of true negatives. On the other hand, 8 is the number of observations belonging to the Domestic class that were misclassified by the model as Foreign, that is, 8 is the number of false positives. Similarly, 0 and 18 are the numbers of false negatives and true positives, respectively. The predicted class for each observation is determined based on a threshold value of 0.208, which is reported above the table. A predicted probability greater than 0.208 will classify the car as Foreign, while a probability below this threshold will classify the car as Domestic. By default, h2omlestat confmatrix uses the threshold that maximizes the F1 score. However, you can select a threshold value or specify that a threshold be selected that maximizes another metric.

The Error column in the output reports the number of misclassified observations for each class, and the Rate column reports the misclassification error rate.

When there are more than two classes, the number of rows and columns in the confusion matrix corresponds to the number of classes. The examples below demonstrate h2omlestat confmatrix after binary classification. For an example with more than two classes, see example 1.

Example 1: Model comparison

In this example, we use the confusion matrix obtained from 3-fold cross-validation to compare two machine learning methods, random forest and gradient boosting machine (GBM), at their default values.

We start by opening the 1978 automobile data (`auto.dta`) in Stata and then putting the data into an H2O frame. Recall that `h2o init` initiates an H2O cluster, `_h2oframe put` loads the current Stata dataset into an H2O frame, and `_h2oframe change` makes the specified frame the current H2O frame. For details, see *Prepare your data for H2O machine learning in Stata* in [H2OML] **h2oml** and see [H2OML] **H2O setup**.

```
. use https://www.stata-press.com/data/r18/auto
(1978 automobile data)
. h2o init
. _h2oframe put, into(auto)
. _h2oframe change auto
```

We run random forest binary classification with 3-fold cross-validation. We store the estimation results by using the `h2omlest store` command so that we can use the results in example 2.

```
. h2oml rfbinclass foreign price mpg trunk weight length, cv(3, modulo)
> h2orseed(19)
Progress (%): 0 100
Random forest binary classification using H2O
Response: foreign
Frame:                                  Number of observations:
  Training: auto                                    Training =     74
                                                Cross-validation =     74
Cross-validation: Modulo                Number of folds      =      3
Model parameters
Number of trees      =  50
              actual =  50
Tree depth:                             Pred. sampling value =     -1
           Input max =  20              Sampling rate        =   .632
                 min =   4              No. of bins cat.     =  1,024
                 avg = 5.8              No. of bins root     =  1,024
                 max =   9              No. of bins cont.    =     20
Min. obs. leaf split =   1              Min. split thresh.   = .00001
Metric summary
```

Metric	Training	Cross-validation
Log loss	.7514549	.4192503
Mean class error	.1127622	.1809441
AUC	.9200175	.8706294
AUCPR	.7622589	.624291
Gini coefficient	.840035	.7412587
MSE	.1081766	.1406502
RMSE	.3289021	.3750336

```
. h2omlest store myrf
```

We report the confusion matrix by using the `h2omlestat confmatrix` command.

```
. h2omlestat confmatrix
Cross-validation confusion matrix using H2O
```

	Predicted				
foreign	Domestic	Foreign	Total	Error	Rate
Domestic	45	7	52	7	.135
Foreign	5	17	22	5	.227
Total	50	24	74	12	.162

```
Note: Probability threshold .38 that maximizes F1 metric
      used for classification.
```

Because cross-validation was implemented during estimation, by default, `h2omlestat confmatrix` reports results that correspond to cross-validation.

Next we implement GBM and report the confusion matrix.

```
. h2oml gbbinclass foreign price mpg trunk weight length, cv(3, modulo)
> h2orseed(19)
Progress (%): 0 94.9 100
Gradient boosting binary classification using H2O
Response: foreign
Loss:     Bernoulli
Frame:                                  Number of observations:
  Training: auto                                  Training =     74
                                     Cross-validation =     74
Cross-validation: Modulo                Number of folds     =      3
Model parameters
Number of trees      =  50              Learning rate       =     .1
             actual =  50               Learning rate decay =      1
Tree depth:                             Pred. sampling rate =      1
           Input max =   5              Sampling rate       =      1
                 min =   2              No. of bins cat.    =  1,024
                 avg = 3.9              No. of bins root    =  1,024
                 max =   5              No. of bins cont.   =     20
Min. obs. leaf split =  10              Min. split thresh.  = .00001
Metric summary
```

Metric	Training	Cross-validation
Log loss	.0796245	.3856675
Mean class error	0	.1284965
AUC	1	.9125874
AUCPR	1	.8214532
Gini coefficient	1	.8251748
MSE	.017155	.1286581
RMSE	.1309771	.3586894

```
. h2omlestat confmatrix
Cross-validation confusion matrix using H2O
```

foreign	Predicted Domestic	Foreign	Total	Error	Rate
Domestic	41	11	52	11	.212
Foreign	1	21	22	1	.045
Total	42	32	74	12	.162

```
Note: Probability threshold .1228 that maximizes F1
      metric used for classification.
```

We can see that random forest is better in predicting `Domestic` cars (45 true negatives versus 41). However, it is not straightforward to quantify how much better because random forest also has more false negatives than does GBM (5 false negatives versus 1). In such cases, we recommend comparing the recall and precision metrics of the two models, which can be obtained from the `h2omlestat threshmetric` command.

In general, when you are interested in quantifying how well a method predicts positives, then the recall metric is recommended.

◁

▷ Example 2: Threshold and metric selection

In example 1, the entries of the confusion matrix were computed using the threshold value that maximizes the F1 score. However, we can instead select a different threshold by using the `threshold()` option or request that `h2omlestat confmatrix` select a threshold value based on optimizing a different metric. Recall that the threshold is a cutoff above which observations are predicted to belong to the positive class and below which observations are predicted to belong to the negative class. Thus, if we change the threshold, the entries of the confusion matrix will also change. Below, we show two confusion matrices with threshold values equal to 0.5 and 0.25 for the random forest.

When we specify the threshold value, `h2omlestat confmatrix` may not report the confusion matrix for the exact value specified. In H2O, the list of possible threshold values for which threshold-based metrics have been computed is limited to the predicted probabilities of the positive class. Therefore, `h2omlestat confmatrix` reports a confusion matrix using the closest available predicted probability of a positive class as the threshold value.

We first restore the random forest estimation results from example 1 with the `h2omlest restore` command and then specify the threshold value in `h2omlestat confmatrix` by using the `threshold(0.25)` option.

```
. h2omlest restore myrf
(results myrf are active now)
. h2omlestat confmatrix, threshold(0.25)
Cross-validation confusion matrix using H2O
```

foreign	Predicted Domestic	Foreign	Total	Error	Rate
Domestic	38	14	52	14	.269
Foreign	3	19	22	3	.136
Total	41	33	74	17	.23

```
Note: Probability threshold .244 that is closest to the
      specified .25 used for classification.
```

Next we obtain the confusion matrix for a threshold value of 0.5.

```
. h2omlestat confmatrix, threshold(0.5)
Cross-validation confusion matrix using H2O
```

	Predicted				
foreign	Domestic	Foreign	Total	Error	Rate
Domestic	46	6	52	6	.115
Foreign	9	13	22	9	.409
Total	55	19	74	15	.203

```
Note: Probability threshold .5 used for classification.
```

We can see that different threshold values substantially change the reported results. The selection of the threshold value depends on the problem that the data scientist is trying to answer. For example, if it is important to classify all `Foreign` cars correctly, then we could choose the threshold that maximizes the true-positive rate by specifying the `metric(tpr)` option.

```
. h2omlestat confmatrix, metric(tpr)
Cross-validation confusion matrix using H2O
```

	Predicted				
foreign	Domestic	Foreign	Total	Error	Rate
Domestic	32	20	52	20	.385
Foreign	0	22	22	0	0
Total	32	42	74	20	.27

```
Note: Probability threshold .0885 that maximizes
      true-positive rate metric used for classification.
```

Stored results

`h2omlestat confmatrix` stores the following in `r()`:

Scalars

`r(threshold)`	specified threshold (with option `threshold()`)
`r(threshold_a)`	actual threshold

Macro

`r(metric)`	metric for threshold selection

Matrix

`r(confmatrix)`	confusion matrix

Also see

[H2OML] **h2oml** — Introduction to commands for Stata integration with H2O machine learning+

[H2OML] **h2omlestat aucmulticlass** — Display AUC and AUCPR after multiclass classification+

[H2OML] **h2omlestat threshmetric** — Display threshold-based metrics for binary classification+