h2omlestat hitratio — Display hit-ratio table

Description

`h2omlestat hitratio` reports hit ratios after multiclass classification performed by `h2oml gbmulticlass` or `h2oml rfmulticlass`. A hit ratio measures how often the correct class is within the top-k predicted classes. The top-k hit ratio is the proportion of observations for which the correct class has one of the k highest predicted probabilities.

Quick start

Display the top-k hit ratios

```
h2omlestat hitratio
```

Same as above, but report results for the validation frame

```
h2omlestat hitratio, valid
```

Menu

Statistics > H2O machine learning

Syntax

```
h2omlestat hitratio [ , options ]
```

options	Description
`title(`*string*`)`	specify title to be displayed above the table
`train`	specify that hit ratios be reported using training results
`valid`	specify that hit ratios be reported using validation results
`cv`	specify that hit ratios be reported using cross-validation results
`test`	specify that hit ratios be computed using the testing frame
`test(`*framename*`)`	specify that hit ratios be computed using data in testing frame *framename*
`frame(`*framename*`)`	specify that hit ratios 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

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

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

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

`train` specifies that hit ratios 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 hit ratios 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 hit ratios 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 hit ratios be computed on the testing frame specified with `h2omlpostestframe`. 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 hit ratios 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 hit ratios 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.

Remarks and examples

For multiclass classification, the hit ratio measures how often the correct class is in one of the top-k predicted classes, where the top-k predicted classes are ranked by predicted probabilities. For example, when computing the top-2 hit ratio, if the true class for an observation has one of the two highest predicted probabilities, then it is considered a "hit"; it is considered a "miss" otherwise. The top-2 hit ratio is the proportion of observations having such a hit. `h2omlestat hitratio` provides a table of top-k hit ratios. If there are more than 10 classes, H2O limits the computation to a maximum of top-10 hit ratios.

In practice, the hit ratio is useful in situations where multiple predictions are made and the true class does not need to have the highest predicted probability but does need to be within the top few. For example, in recommendation systems or search engines, the output is presented as a ranked list of results. The correct result needs to be somewhere near the top of that list, but it does not necessarily need to be the first one.

▷ Example 1: Hit ratios

We use a well-known `iris` dataset, where the goal is to predict a class of iris plant. This dataset was used in Fisher (1936) and originally collected by Anderson (1935). We start by initializing a cluster, opening the dataset in Stata, and importing the dataset as 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. We also use the `_h2oframe split` command to split the dataset, specifying 70% of observations in the training frame and 30% in the validation 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/r19/iris
(Iris data)
. h2o init
  (output omitted)
. _h2oframe put, into(iris)
Progress (%): 0 100
. _h2oframe split iris, into(train valid) split(0.7 0.3) rseed(19)
. _h2oframe change train
```

We define the global macro `predictors` to store the names of the predictors, and we use the `h2oml rfmulticlass` command to perform random forest multiclass classification. We use default settings for all hyperparameters, and we specify an H2O random-number seed for reproducibility. We also specify the name of our validation frame in the `validframe()` option.

```
. global predictors seplen sepwid petlen petwid
. h2oml rfmulticlass iris $predictors, validframe(valid) h2orseed(19)
Progress (%): 0 100
Random forest multiclass classification using H2O
Response: iris                          Number of classes    =      3
Frame:                                  Number of observations:
  Training:   train                                Training =    113
  Validation: valid                             Validation =     37
Model parameters
Number of trees      =  50
             actual =  50
Tree depth:                             Pred. sampling value =     -1
           Input max =  20              Sampling rate        =   .632
                 min =   1              No. of bins cat.     =  1,024
                 avg = 3.2              No. of bins root     =  1,024
                 max =   6              No. of bins cont.    =     20
Min. obs. leaf split =   1              Min. split thresh.   = .00001
Metric summary
```

Metric	Training	Validation
Log loss	.0821639	.1523995
Mean class error	.0456654	.0747475
MSE	.0269054	.0555373
RMSE	.1640287	.2356636

The top-1 hit ratio is closely related to the misclassification error, which we will report first by using the `h2omlestat confmatrix` command.

```
. h2omlestat confmatrix
Confusion matrix using H2O
Validation frame: valid
```

	Predicted					
iris	Setosa	Versico~r	Virginica	Total	Error	Rate
Setosa	11	0	0	11	0	0
Versicolor	0	10	1	11	1	.091
Virginica	0	2	13	15	2	.133
Total	11	12	14	37	3	.081

This confusion matrix based on validation results shows that the highest predicted probabilities from the model misclassified three observations, resulting in a misclassification error of 0.08. This means that the top-1 hit ratio is 0.92 (1 − 0.08). In other words, the true class has the highest predicted probability for 92% of observations.

To determine the top-2 hit ratio, we need to know whether the true class for each of the three misclassified observations has the second highest predicted probability. To check, we predict the class and corresponding probabilities using the validation frame. By default, `h2omlpredict` generates predictions in the current working frame. (We can use `_h2oframe pwf` to check which is the current

frame.) To make predictions in the validation frame, we set it as our postestimation frame by using the h2omlpostestframe command. We use h2omlpredict to obtain the predicted class, the default prediction. We then specify the pr option to obtain the predicted probabilities of each class.

```
. h2omlpostestframe _valid
(validation frame valid is now active for h2oml postestimation)
. h2omlpredict pr_class
(option class assumed; predicted class)
Progress (%): 0 100
. h2omlpredict pr_setosa pr_versicolor pr_virginica, pr
Progress (%): 0 100
```

Because the h2omlpostestframe command does not physically change the current frame, we use the _h2oframe change command to change the working frame before listing the misclassified observations.

```
. _h2oframe change valid
. _h2oframe list iris pr_class pr_setosa pr_versicolor pr_virginica
> if pr_class != iris, abbreviate(14)
        iris    pr_class  pr_setosa  pr_versicolor  pr_virginica
1 Versicolor   Virginica          0       .2038981      .7961019
2  Virginica  Versicolor          0       .8080754      .1919246
3  Virginica  Versicolor          0       .8631397      .1368603
[3 rows x 5 columns]
```

In the first row, we see that the model misclassified true class Versicolor as Virginica with the probability 0.8. For this observation, the probability of predicting Versicolor, the true class, is the second highest probability of 0.2. Similarly, for the next two observations, the second highest predicted probability corresponds to the true class. Consequently, for all misclassified observations, the top-2 predicted classes contain the true class; thus, the top-2 hit ratio is 1.

The h2omlestat hitratio command provides an easy way to obtain the hit ratios we computed manually.

```
. h2omlestat hitratio
Hit-ratio table using H2O
Validation frame: valid
```

Top	Hit ratio
1	.9189189
2	1
3	1

From this table, we confirm that the true class has the highest predicted probability for 92% of observations in the validation data. The true class has one of the two highest predicted probabilities for 100% of the observations.

In this example, we see top-1, top-2, and top-3 hit ratios. For classification problems in which the response has many classes, h2omlestat hitratio will report all top-k hit ratios up to the top-10 hit ratio.

◁

Stored results

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

Matrices
`r(hitratio)` hit ratios

References

Anderson, E. 1935. The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society* 59: 2–5.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

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 confmatrix** — Display confusion matrix