h2omlestat cvsummary — Display cross-validation summary[+]

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

Description

`h2omlestat cvsummary` displays the cross-validation summary for each fold after performing cross-validation with `h2oml` *gbm* or `h2oml` *rf*. `h2omlestat cvsummary` reports performance metrics for each fold as well as the mean and standard deviation of each metric. The individual metrics and summary statistics are useful for evaluating the stability of the machine learning method and whether results will generalize well to new data.

Quick start

Display the 5-fold cross-validation summary after `h2oml rfregress`

```
h2oml rfregress y1 x1-x100, cv(5) h2orseed(19)
h2omlestat cvsummary
```

Specify a title for the table

```
h2omlestat cvsummary, title(5-fold CV summary)
```

Menu

Statistics > H2O machine learning

Syntax

`h2omlestat cvsummary` [, `title(`*string*`)`]

Option

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

Remarks and examples

We assume you have read *Model selection in machine learning* in [H2OML] **Intro**.

k-fold cross-validation is one of the most common model evaluation and selection techniques. Similar to the two-way holdout method, we start by splitting data into training and testing sets. However, k-fold cross-validation additionally splits the training set into k folds. In each iteration, it uses one fold for validation and the remaining $k-1$ folds as a training subset for model fitting. One way to compute a cross-validation metric is to take the average of the k validation metrics of the cross-validated models. `h2omlestat cvsummary` reports this average along with the standard deviation and the estimated metrics for each fold.

Looking at the standard deviation of cross-validated metrics over the folds can provide useful insights into the stability and reliability of a machine learning model. For example, if the standard deviation across the folds is large, it may indicate that the performance of the model is not consistent across different subsets of data and that the model will not generalize well to new data. A large standard deviation could also indicate data issues; for example, data may be insufficient for reliable training or may suffer from imbalanced classes.

Another common reason for a large standard deviation is the bias–variance tradeoff of the machine learning model. A large standard deviation can indicate overfitting, where the model is too complex and closely learns patterns in the training data. In such cases, a less complex model that provides slightly lower performance metrics but also low variance might be preferable.

Several authors have tried to find the best value of k that minimizes the bias–variance tradeoff. Based on numerous empirical analyses, Kohavi (1995) suggests $k = 10$ folds. However, cross-validation with this many folds can be computationally intensive when the dataset is large. In general, as the number of folds increases, the performance bias decreases but the variance of the performance metric and computational cost increases.

The steps for hyperparameter tuning with k-fold cross-validation are as follows:

1. Split the dataset into two sets—a training set for model fitting and selection and a testing set for the final model evaluation.
2. Perform hyperparameter tuning. For each hyperparameter configuration, apply the k-fold cross-validation method on the training set.
3. Select the best hyperparameter settings from the k-fold cross-validation, and apply them to the entire training set.
4. Use the independent testing set and the hyperparameter setting from the previous step to estimate the generalization performance.

To perform cross-validation with the `h2oml` *gbm* and `h2oml` *rf* commands, we specify the `cv()` option. After estimation, we can use `h2omlestat cvsummary` to summarize performance metrics and examine their results for each fold.

Example 1: Cross-validation summary for bias–variance tradeoff

In this example, we use gradient boosting binary classification on the auto dataset to examine the standard deviation of a cross-validated metric as an indicator for overfitting.

We start by opening auto.dta in Stata and then putting it in 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. (Because we are focused on evaluating cross-validation, we do not split the data into training and testing sets as we typically would in practice.) 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
  (output omitted)
. _h2oframe put, into(auto)
Progress (%): 0 100
. _h2oframe change auto
```

We perform gradient boosting binary classification with 3-fold cross-validation and use 5,000 trees.

```
. h2oml gbbinclass foreign price mpg weight length, cv(3, modulo) h2orseed(19)
> ntrees(5000)
Progress (%): 0 0.3 1.2 2.4 3.6 11.4 15.6 19.9 26.4 31.9 32.7 33.3 33.8 34.6 38.
> 0 41.0 45.6 52.7 100
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      = 5,000            Learning rate       =     .1
              actual = 5,000            Learning rate decay =      1
Tree depth:                             Pred. sampling rate =      1
           Input max =     5            Sampling rate       =      1
                 min =     1            No. of bins cat.    =  1,024
                 avg =   2.7            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	1.80e-17	2.487799
Mean class error	0	.1197552
AUC	1	.8902972
AUCPR	1	.7719202
Gini coefficient	1	.7805944
MSE	4.00e-33	.1135748
RMSE	6.32e-17	.3370087

Next we report the cross-validated metrics for each fold, together with the mean and standard deviation.

```
. h2omlestat cvsummary
Cross-validation summary using H2O
```

Metric	Mean	Std. dev.	Fold 1	Fold 2	Fold 3
Log loss	2.467125	2.757786	.8134241	5.650739	.9372107
F1	.8586183	.0740218	.9230769	.7777778	.875
F2	.8872107	.0564633	.882353	.8333333	.9459459
F0.5	.8369541	.1209393	.9677419	.7291667	.8139535
Accuracy	.9055555	.0607667	.96	.84	.9166667
Precision	.825926	.1556878	1	.7	.7777778
Recall	.9107143	.0778375	.8571429	.875	1
Specificity	.9019608	.0898544	1	.8235294	.882353
Misclassification	.0944444	.0607667	.04	.16	.0833333
Mean class error	.0936625	.0498267	.0714286	.1507353	.0588235
Max. class error	.1456583	.0295116	.1428571	.1764706	.1176471
Mean class accuracy	.9063376	.0498267	.9285714	.8492647	.9411765
Misclassification count	2.333333	1.527525	1	4	2
AUC	.919779	.0744504	.984127	.8382353	.9369748
AUCPR	.7621639	.180335	.9663477	.624682	.6954619
MSE	.1134442	.0786849	.0400411	.196517	.1037744
RMSE	.3218485	.1216001	.2001026	.4433024	.3221404

For illustration purposes, we focus on the log-loss metric; for details, see [H2OML] ***metric_option***. In the first row of the output, the mean is 2.47 and the standard deviation is 2.76. Further analysis reveals that fold 2 has a large log-loss metric. One possible explanation is that, given the simplicity of this dataset, fitting a model with a large number of trees might lead to overfitting, which is why the model does not generalize well for data in fold 2. To investigate, we fit a less complex model with the default 50 trees and report the cross-validation results.

```
. h2oml rfbinclass foreign price mpg 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 =   3              No. of bins cat.     =  1,024
                 avg = 5.6              No. of bins root     =  1,024
                 max =   8              No. of bins cont.    =     20
Min. obs. leaf split =   1              Min. split thresh.   = .00001
Metric summary
```

Metric	Training	Cross-validation
Log loss	.3097282	.8764794
Mean class error	.1284965	.2036713
AUC	.9278846	.8435315
AUCPR	.8502403	.6751862
Gini coefficient	.8557692	.6870629
MSE	.1088474	.1504919
RMSE	.3299203	.3879328

```
. h2omlestat cvsummary
Cross-validation summary using H2O
```

Metric	Mean	Std. dev.	Fold 1	Fold 2	Fold 3
Log loss	.8879563	.7421946	.3638948	.5627286	1.737245
F1	.7857143	.0795395	.8571429	.7	.8
F2	.8286436	.0311104	.8571429	.7954546	.8333333
F0.5	.7504579	.1172045	.8571429	.625	.7692308
Accuracy	.8516667	.0825126	.92	.76	.875
Precision	.7301587	.1379789	.8571429	.5833333	.75
Recall	.8630952	.0103098	.8571429	.875	.8571429
Specificity	.8442266	.1237666	.9444444	.7058824	.882353
Misclassification	.1483333	.0825126	.08	.24	.125
Mean class error	.1463391	.0569079	.0992063	.2095588	.1302521
Max. class error	.1932773	.0873303	.1428571	.2941177	.1428571
Mean class accuracy	.8536609	.0569079	.9007937	.7904412	.8697479
Misclassification count	3.666667	2.081666	2	6	3
AUC	.843643	.067583	.9206349	.8161765	.7941176
AUCPR	.663395	.0049219	.6678722	.6581247	.6641881
MSE	.150353	.0437331	.112672	.1983087	.1400785
RMSE	.3850852	.0556203	.3356665	.4453186	.3742706

We can see that the mean and standard deviation of the log loss are now much smaller.

◁

Stored results

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

Matrix

`r(cvsummary)` summary of cross-validation metrics and metrics for each fold

Reference

Kohavi, R. 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, August 20–25*, vol. 2: 1137–1143. San Francisco: Morgan Kaufman.

Also see

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