Requirements

To use SAS Cloud Analytic Services, the client machine must meet the following requirement:

- Use 64-bit Linux

There are additional requirements that are common with other programming languages. For other requirements, see “Programming Basics” in SAS Viya: System Programming Guide.

Connect and Start a Session

How To

1. If you have not already done so, download the data sets that are used in examples. You can download the data at http://support.sas.com/documentation/onlinedoc/viya/examples.htm. Put them in a directory that is accessible to SAS.

2. Open SAS Studio from the URL in the form of http://hostname:port. Sign in using your user ID and password for your operating system account.
   a. Use the Server Files and Folders section to navigate to the directory that has data sets.
   b. Right-click on the directory and select Create ➔ Library. Specify heart as the name.

3. Use the Libraries section to confirm that the new library is listed.

4. Check that you have a CAS session.
   a. Enter the following line in the code editor and click 🚀 to run the code.
      ```
      cas casauto list;
      ```
      If you get the error: "ERROR: Request failed. Session CASAUTO not recognized", see Tip: Automatically Connect and Generate Librefs.
   b. Click the Code tab. The following note indicates that you have an active CAS session.
NOTE: The session CASAUTO is ACTIVE using port 5570 and host cloud.example.com for user sasdemo. The session UUID is 0ed1cc35-c1ec-df49-be6e-68b9193eb8b8.

About Your Connection and Server

After you connect, a session is started for you. As a documentation convention, a variable that is named S is used to represent the session.

S
  The name of the variable that is in CASL. It represents the session that is started for you in SAS Cloud Analytic Services.

session
  The software process that is started on the same hosts as SAS Cloud Analytic Services. When you reference your session through the S variable, statistical computations and actions are run on the server.

As soon as you connect, a good practice is to print information about the connection and session:

print s;

{state=Connected,number of Connections=1,Timeout=60, ActionStatus=Action is active,Authenticated= Yes,locale=en_US}

Your results will show different values. In the event that you have a network interruption between CAS and the server, the UUID for the session can be used to reconnect to a session.

To learn the most basic information about the server, you can run the serverStatus action that is part of the builtins action set:

proc cas;
  builtins.serverStatus;
run;

{About={CAS=Cloud Analytic Services,Version=3.01, VersionLong=V.03.01,Copyright=Copyright © 2014-2016 SAS Institute Inc. All Rights Reserved., System={Hostname=cloud.example.com,OS Name=Linux,OS Family=LIN X64, OS Release=2.6.32-573.el6.x86_64,OS Version=#1 SMP Wed Jul 1 18:23:37 EDT 2015,Model Number=x86_64},license={site=SAS Institute Inc., siteNum=x,expires=x,gracePeriod=62,warningPeriod=31}},}
Example: Train Gradient Boosted Trees with k-fold Cross Validation

About the Example

The purpose of this example is to describe how to train gradient boosted trees with k-fold cross validation. A k-fold cross validation process finds the average estimated validation error (misclassification error for nominal targets or average square error for interval targets) for the trained model. During cross validation, all data are divided into k subsets (folds). For each fold, a new model is trained, and then validated using the selected fold. The validation error estimates are then averaged over each set of training and scoring executions to obtain a single value.

Load the Data

Create a CAS session and set up a CAS engine libref that you can use to connect to the CAS session. It assumes that you have a CAS server already available; contact your system administrator if you need help starting and terminating a server.

libname mycas mysess; /* #1 */
data mycas.heart; /* #2 */
  set sashelp.heart;
run;

1 Define your own session and CAS engine librefs that connect to the CAS server. The CAS statement creates the CAS session named mysess, and the LIBNAME statement creates the mycas CAS engine libref that you use to connect to this session. It is not necessary to explicitly name the CASHOST and CASPORT of the CAS server in the CAS statement, because these values are retrieved from the corresponding SAS option values.

2 The DATA step creates the data table mycas.heart, which consists of 5209 observations that have 17 variables. This DATA step assumes that your CAS engine libref is named mycas, but you can substitute any appropriately defined CAS engine libref.

Explore Data Using CAS Actions

This section describes how to access and manage your data using the tables action set. For more information on the table action set see SAS Viya: System Programming Guide.

proc cas;
  tableinfo /table='heart'; /* #1 */
run;

  columninfo result=r /table='heart'; /* #2 */
  print r;
run;

  simple.summary/table='heart'; /* #3 */
run;

1 The table action set executes the tableInfo action, which shows information about a table.

2 The table action set executes the columnInfo action, which shows column information.
The simple action set executes the summary action, which generates descriptive statistics of numeric variables such as the sample mean, sample variance, sample size, sum of squares, and so on. For more information about the simple action set, see “Simple Analytics Action Set” in SAS Visual Analytics Programming Guide.

Results: Table Information for Heart

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Number of Rows</th>
<th>Number of Columns</th>
<th>NLS encoding</th>
<th>Created</th>
<th>Last Modified</th>
<th>Promoted Table</th>
<th>Duplicated Rows</th>
<th>View</th>
<th>Compressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEART</td>
<td>5209</td>
<td>17</td>
<td>utf-8</td>
<td>09Sep2016:15:18:00</td>
<td>08Sep2016:15:18:00</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Results: Column Information for Heart

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Number of Rows</th>
<th>Number of Columns</th>
<th>NLS encoding</th>
<th>Created</th>
<th>Last Modified</th>
<th>Promoted Table</th>
<th>Duplicated Rows</th>
<th>View</th>
<th>Compressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEART</td>
<td>5209</td>
<td>17</td>
<td>utf-8</td>
<td>09Sep2016:15:18:00</td>
<td>08Sep2016:15:18:00</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Results: Simple Summary for Heart

<table>
<thead>
<tr>
<th>Column</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
<th>Sum</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Error</th>
<th>Variance</th>
<th>Coefficient of Variation</th>
<th>Corrected Sum of Squares</th>
<th>Uncorrected Sum of Squares</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th>Number Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeAtDisq</td>
<td>32.000</td>
<td>90.000</td>
<td>1449</td>
<td>91720</td>
<td>63.3030</td>
<td>10.0162</td>
<td>0.2932</td>
<td>100.36</td>
<td>15.0250</td>
<td>145320</td>
<td>5951026</td>
<td>240.53</td>
<td>&lt;.0001</td>
<td>3790</td>
<td></td>
</tr>
<tr>
<td>AgeAtStart</td>
<td>20.000</td>
<td>90.000</td>
<td>5209</td>
<td>220554</td>
<td>44.6697</td>
<td>6.7570</td>
<td>0.1180</td>
<td>73.5290</td>
<td>19.4581</td>
<td>302940</td>
<td>10499096</td>
<td>370.92</td>
<td>&lt;.0001</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>51.5000</td>
<td>76.5000</td>
<td>5209</td>
<td>397223</td>
<td>64.6812</td>
<td>3.0527</td>
<td>0.0967</td>
<td>12.9356</td>
<td>5.5277</td>
<td>00772</td>
<td>21922389</td>
<td>1340.91</td>
<td>&lt;.0001</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>67.0000</td>
<td>90.0000</td>
<td>5209</td>
<td>796510</td>
<td>153.80</td>
<td>26.9154</td>
<td>0.4809</td>
<td>636.80</td>
<td>16.8683</td>
<td>4344902</td>
<td>12093828</td>
<td>391.99</td>
<td>&lt;.0001</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Diastolic</td>
<td>50.0000</td>
<td>90.0000</td>
<td>5209</td>
<td>444033</td>
<td>66.5506</td>
<td>16.7527</td>
<td>0.7076</td>
<td>168.30</td>
<td>15.9893</td>
<td>670512</td>
<td>39029767</td>
<td>474.66</td>
<td>&lt;.0001</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Systolic</td>
<td>52.0000</td>
<td>90.0000</td>
<td>5209</td>
<td>713462</td>
<td>83.91</td>
<td>23.7046</td>
<td>0.3280</td>
<td>658.57</td>
<td>17.3968</td>
<td>2005064</td>
<td>1007785</td>
<td>418.23</td>
<td>&lt;.0001</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>MBG</td>
<td>57.0000</td>
<td>90.0000</td>
<td>5209</td>
<td>824110</td>
<td>109.56</td>
<td>19.0834</td>
<td>0.2776</td>
<td>369.36</td>
<td>16.6557</td>
<td>2077348</td>
<td>7640176</td>
<td>433.00</td>
<td>&lt;.0001</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>0.00000</td>
<td>60.0000</td>
<td>5173</td>
<td>48453</td>
<td>5.9465</td>
<td>0.5215</td>
<td>0.0673</td>
<td>144.78</td>
<td>15.84</td>
<td>474887</td>
<td>1202563</td>
<td>56.50</td>
<td>&lt;.0001</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>AgeAtDeath</td>
<td>50.0000</td>
<td>90.0000</td>
<td>1551</td>
<td>142458</td>
<td>70.5364</td>
<td>10.5554</td>
<td>0.3266</td>
<td>111.50</td>
<td>14.5701</td>
<td>221867</td>
<td>10127580</td>
<td>258.08</td>
<td>&lt;.0001</td>
<td>3218</td>
<td></td>
</tr>
<tr>
<td>Cholesterol</td>
<td>96.0000</td>
<td>568.0000</td>
<td>5057</td>
<td>1150500</td>
<td>227.42</td>
<td>44.9354</td>
<td>0.6319</td>
<td>2019.20</td>
<td>19.7590</td>
<td>10209082</td>
<td>2.717568</td>
<td>359.90</td>
<td>&lt;.0001</td>
<td>152</td>
<td></td>
</tr>
</tbody>
</table>

Generate Folds for Cross Validation

Use CASL to partition the table and specify the number of partition folds for cross validation. For this example, we are generating ten folds. A column named _fold_ is produced, and the only way to get repeatable folds is to copy the new table and add the _fold_ column.

```plaintext
partition / table={name='heart',
  compvars={'_fold_'},
  comppgm='call streaminit(__rankid*1000);_fold_=floor(rand("UNIFORM")*10);'}
outtable={name='new_heart_with_fold', replace=True}
run;
```

Log Output: Partitioned Table Log Output

```
{caslib=CASUSERHDFS(casuser),tableName=NEW_HEART_WITH_FOLD,rowsTransferred=27,
 shuffleWaitTime=0.0000758171,minShuffleWaitTime=0,maxShuffleWaitTime=1.9073486E-6,
 averageShuffleWaitTime=4.5672957E-7}
```
Verify the _Fold_ Column

In order to verify the new _fold_ column, perform these simple statistics.

```
summary/ table='new_heart_with_fold' inputs={'_fold'};
run;
simple.distinct / table={name='new_heart_with_fold', inputs='_fold_'};
run;
freq/ table={name='new_heart_with_fold', inputs={'_fold_'});
run;
columninfo result=r /table={name='new_heart_with_fold'};
run;
```

**Results: Simple Distinct Results**

<table>
<thead>
<tr>
<th>Column</th>
<th>Number of Distinct Values</th>
<th>Number of Missing Values</th>
<th>Truncated</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>fold</em></td>
<td>10</td>
<td>0</td>
<td>No</td>
</tr>
</tbody>
</table>

**Results: Frequency Table**

<table>
<thead>
<tr>
<th>Column</th>
<th>Numeric Value</th>
<th>Formatted Value</th>
<th>Level</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>fold</em></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>508</td>
</tr>
<tr>
<td><em>fold</em></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>524</td>
</tr>
<tr>
<td><em>fold</em></td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>938</td>
</tr>
<tr>
<td><em>fold</em></td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>920</td>
</tr>
<tr>
<td><em>fold</em></td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>513</td>
</tr>
<tr>
<td><em>fold</em></td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>530</td>
</tr>
<tr>
<td><em>fold</em></td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>550</td>
</tr>
<tr>
<td><em>fold</em></td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>525</td>
</tr>
<tr>
<td><em>fold</em></td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>480</td>
</tr>
<tr>
<td><em>fold</em></td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>511</td>
</tr>
</tbody>
</table>

Remove the _Fold_ Column

Remove the _fold_ column since it is not our analysis variable.

```
nVars=dim(r['columninfo'])-1;
```

Create an Input Variable List

Create a variable list where you define each variable. Each variable is defined by an expression, character, or numeric value. For this example, we use CASL to create our variable list. The syntax for the variable list follows the Assignment statement syntax. For more information see "ASSIGNMENT Statement" in SAS Cloud Analytic Services: CAS Procedure Programming Guide and Reference.

```
i=4;
j=2;
xx= {r['columninfo*'][1,1]}; /*#1*/
do while (i<=nVars); /*#2*/
The target \( xx \) uses the result variable from the previously run columninfo, and \( r \) as its expression in addition to searching for the value in \( 1, 1 \).

The DO WHILE statement executes statements in a DO loop repetitively as long as the condition is true. For more information see "Using a DO WHILE Statement" in SAS Cloud Analytic Services: CAS Procedure Programming Guide and Reference.

Log Output: Variable List Concatenated

```
{Status, Sex, AgeAtStart, Height, Weight, Diastolic, Systolic, MRW, Smoking, AgeAtDeath, Cholesterol, Chol_Status, BP_Status, Weight_Status, Smoking_Status}
```
the FUNCTION statement syntax, see “FUNCTION Statement” in SAS Cloud Analytic Services: CAS Procedure Programming Guide and Reference.

2 Generate a model name.

3 Create a model without iFold, and score on the holdout iFold using the train model. The action set decision tree executes the action gbtreetrain that trains a gradient boosting tree. This function can be easily expanded to train other models such as a neural network with cross validation. For more information about gbtreetrain action syntax, see “Train gradient boosting tree” in SAS Visual Analytics Programming Guide.

4 Score iFold-th data with a trained model. The action set decision tree executes the action gbtreescore that scores a table using a gradient boosting tree model. For more information about gbtreescore action syntax, see “Score a table using gradient boosting tree” in SAS Visual Analytics Programming Guide.

5 Generate a score data set name. In this example the score data set name is myscoredata. The data set name is evaluated as "gbtscore_" || (String)iFold. For more information about the Assignment statement syntax, see “ASSIGNMENT Statement” in SAS Cloud Analytic Services: CAS Procedure Programming Guide and Reference.

6 The SAVERESULT statement creates a SAS data set from the results of an ACTION. For more information about the SAVERESULT statement syntax, see “SAVERESULT Statement” in SAS Cloud Analytic Services: CAS Procedure Programming Guide and Reference.

7 The RETURN statement returns a value from the current function. For more information about the RETURN statement syntax, see “RETURN Statement” in SAS Cloud Analytic Services: CAS Procedure Programming Guide and Reference.

Partial Results: Decision Tree Action

The SAS System
Results from decisionTree.gbtreescore

Gradient Boosting Tree for NEP_HEART_WITH_FOLD

| Number of Trees       | 100.000000 |
| Distribution          | 1.000000   |
| Learning Rate         | 0.100000   |
| Subsampling Rate      | 0.700000   |
| Number of Selected Variables (M) | 14.000000 |
| Number of Bins        | 100.000000 |
| Number of Variables   | 14.000000  |
| Max Number of Tree Nodes | 33.000000 |
| Min Number of Tree Nodes | 17.000000 |
| Max Number of Branches | 2.000000   |
| Min Number of Branches | 2.000000   |
| Max Number of Leaves  | 9.000000   |
| Min Number of Leaves  | 8.000000   |
| Max Number of Leaves  | 9.000000   |
| Min Number of Leaves  | 8.000000   |
| Maximum Size of Leaves| 400.000000 |
| Minimum Size of Leaves| 60.000000  |
| Random Number Seed    | 1234.000000|

Output CAS Tables

<table>
<thead>
<tr>
<th>CAS Library</th>
<th>Name</th>
<th>Number of Rows</th>
<th>Number of Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASUSERHDFS</td>
<td>gbt_0</td>
<td>2730</td>
<td>29</td>
</tr>
</tbody>
</table>
Run Each Model In Its Own Session

You can run each model in its own session by creating a function for each model.

```sas
function KFoldCV(nFold); /*#1*/
   do i=1 to nFold; /*#2*/
      myerror[i] = OneFoldTree(nFold, i);
   end;
   return (myerror); /*#3*/
end func;
```

1. The FUNCTION statement creates a new function that can be called in an expression. In this example, the function is named `KFoldCV` and has one argument named `nFold`.

2. The DO statement, Iterative iterates over the list that starts at the value of 1 to the value of `nFold`. For more information about the DO Iterative syntax, see “DO Statement, Iterative Statement” in SAS Cloud Analytic Services: CAS Procedure Programming Guide and Reference.

3. The RETURN statement returns a value from the current function. For more information about the RETURN statement syntax, see “RETURN Statement” in SAS Cloud Analytic Services: CAS Procedure Programming Guide and Reference.

Train k (nFold) Models

During cross validation, all data are divided into k subsets (folds). For each fold, a new model is trained then validated using a selected fold. In this example, we have ten folds that we are going to train against the selected fold (`nFold=1`).

```sas
nFold = 10;
ModelError = KFoldCV(nFold);
run;
```
Log Output: Trained Models

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
85.74875114

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
58.51471611

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
71.307144715

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
59.846180435

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
63.013744439

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
59.198893492

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
60.686242454

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
59.041830014

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
62.895729306

NOTE: The data set work.myscoredata has 3 observations and 2 variables.
NOTE: The data set work.myscoredata1 has 100 observations and 6 variables.
57.16065856

/*print error into log*/
print ModelError;
run;

Output Log: Model Error for Each Fold

{58.744875114, 50.51471611, 71.307144715, 59.846180435, 63.013744439, 59.198893492, 60.686242454, 59.041830014, 62.895729306, 57.16065856}

Compute Average Error Rate from Cross Validation

This section describes how to compute the average error rate or the misclassification error or average square of k-fold cross validation.

\[
\text{mse} = 0; /*1*/
\]
\[
\text{do } i = 1 \text{ to nFold}; /*2*/
\]
\[
\text{mse} = \text{mse} + \text{ModelError}[i];
\]
end;
Compute the average error rate from cross validation. The mean squared error (MSE) of an estimator measures the average of the squares of the deviations.

The DO statement, Iterative iterates over the list that starts at the value of 1 to the value of nFold. For more information about the DO statement, Iterative syntax see “DO Statement, Iterative Statement” in SAS Cloud Analytic Services: CAS Procedure Programming Guide and Reference.

Log Output: Conversion of String to Number

NOTE: String '58.744875114' convert to number.
NOTE: String '50.51471611' convert to number.
NOTE: String '71.307144715' convert to number.
NOTE: String '59.846180435' convert to number.
NOTE: String '63.013744439' convert to number.
NOTE: String '59.198893492' convert to number.
NOTE: String '60.686242454' convert to number.
NOTE: String '59.041830014' convert to number.
NOTE: String '62.895729306' convert to number.
NOTE: String '57.160065856' convert to number.

```
mse = mae/nFold; /*#1*/
rmse = sqrt(mse);
print "mse=" mse "rmse=" rmse;
run;
```

Print the average mean squared error (MSE) and root mean squared error (RMSE) to the log.

Log Output: Average MSE and RMSE

```
mse=60.240942194 rmse=7.7615038616
```

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2017, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. September 2017 3.2-P2:casl