Getting Started with the CASL Programming Language 3.1

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 Cloud Analytic Services: 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 the two data sets.
   b. Right-click on the directory and select **Create** ⇨ **Library**. Specify **movies** as the name.
3. Use the **Libraries** section to confirm that the new library is listed.
4. Start a CAS session.

2

a Enter the following line in the code editor and click to run the code.

```
cas casauto userid=sasdemo host="cloud.example.com" port=5570;
proc cas;
    session casauto;
    session.sessionstatus result=s;
    put s;
run;
```

b Click the **Code** tab. The follow note indicates that you have an active CAS session.

```
 57 cas casauto userid=sasdemo host="cloud.example.com" port=5570;
 NOTE: The session CASAUTO connected successfully to Cloud Analytic Services cloud.example.com using port 5570.
The UUID is
       0ed1cc35-c3ec-df49-be6e-68b9193eb8b8. The user is sasdemo and the active caslib is CASUSERHDFS(sasdemo).
 NOTE: The SAS option SESSREF was updated with the value CASAUTO.
 NOTE: The SAS macro _SESSREF_ was updated with the value CASAUTO.
 NOTE: The session is using 139 workers.
 58         proc cas;
 59         session TestSess;
 60         session.sessionstatus result=s;
 61         put s;
 62         run;
 NOTE: Active Session now TestSess.
 {state=Connected,number of Connections=1,Timeout=60,ActionStatus=Action is
active,Authenticated=Yes,locale=en_US}
```

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 you can access and manage your data using the tables action set. For more information on the table action set see *SAS Cloud Analytic Services: 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.

3 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 "Summary" in *SAS Cloud Analytic Services: Analytics Programming Guide*

Results: Table Information for Heart

Table Information for Caslib CASUSERHDFS									
Table Name	Number of Rows	Number of Columns	NLS encoding	Created	Last Modified	Promoted Table	Duplicated Rows	View	Compressed
HEART	5209	17	utf-8	09Sep2016:15:18:00	09Sep2016:15:18:00	No	No	No	No

Results: Column Information for Heart

Column Information for HEART in Caslib CASUSERHDFS					
Column	Label	Id	Type	Length	Formatted Length
Status		1	char	5	5
DeathCause	Cause of Death	2	char	26	26
AgeCHDdiag	Age CHD Diagnosed	3	double	8	12
Sex		4	char	6	6
AgeAtStart	Age at Start	5	double	8	12
Height		6	double	8	12
Weight		7	double	8	12
Diastolic		8	double	8	12
Systolic		9	double	8	12
MRW	Metropolitan Relative Weight	10	double	8	12
Smoking		11	double	8	12
AgeAtDeath	Age at Death	12	double	8	12
Cholesterol		13	double	8	12
Chol_Status	Cholesterol Status	14	char	10	10
BP_Status	Blood Pressure Status	15	char	7	7
Weight_Status	Weight Status	16	char	11	11
Smoking_Status	Smoking Status	17	char	17	17

Results: Simple Summary for Heart

Descriptive Statistics for HEART														
Column	Min	Max	N	Sum	Mean	Std Dev.	Std Error	Variance	Coefficient of Variation	Corrected Sum of Squares	Uncorrected Sum of Squares	t Value	Pr > |t|	Number Missing
AgeCHDdiag	32.0000	90.0000	1449	91726	63.3030	10.0182	0.2632	100.36	15.8258	145328	5951856	240.53	<.0001	3760
AgeAtStart	28.0000	62.0000	5209	229554	44.0687	8.5750	0.1188	73.5298	19.4581	382943	10499096	370.92	<.0001	0
Height	51.5000	76.5000	5203	337223	64.8132	3.5827	0.04967	12.8358	5.5277	66772	21923268	1304.91	<.0001	6
Weight	67.0000	300.00	5203	796510	153.09	28.9154	0.4009	836.10	18.8883	4349402	1.2628E8	381.89	<.0001	6
Diastolic	50.0000	160.00	5209	444633	85.3586	12.9731	0.1797	168.30	15.1983	876512	38829767	474.88	<.0001	0
Systolic	82.0000	300.00	5209	713162	136.91	23.7396	0.3289	563.57	17.3396	2935064	1.0057E8	416.23	<.0001	0
MRW	67.0000	268.00	5203	624139	119.96	19.9834	0.2770	399.34	16.6587	2077348	76947517	433.00	<.0001	6
Smoking	0	60.0000	5173	48453	9.3665	12.0315	0.1673	144.76	128.45	748677	1202513	55.99	<.0001	36
AgeAtDeath	36.0000	93.0000	1991	140438	70.5364	10.5594	0.2366	111.50	14.9701	221887	10127880	298.06	<.0001	3218
Cholesterol	96.0000	568.00	5057	1150050	227.42	44.9355	0.6319	2019.20	19.7590	10209082	2.7175E8	359.90	<.0001	152

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.

```
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

Distinct Counts for NEW_HEART_WITH_FOLD			
Column	Number of Distinct Values	Number of Missing Values	Truncated
fold	10	0	No

Results: Frequency Table

Frequency for NEW_HEART_WITH_FOLD				
Column	Numeric Value	Formatted Value	Level	Frequency
fold	0	0	1	508
fold	1	1	2	524
fold	2	2	3	538
fold	3	3	4	520
fold	4	4	5	513
fold	5	5	6	530
fold	6	6	7	550
fold	7	7	8	525
fold	8	8	9	490
fold	9	9	10	511

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 */
    xx=xx + r["columninfo"][i,1];
    i=i+1;
    j=j+1;
end;
print xx;
run;
```

1 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.

2 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}
```

Train a Decision Tree Using a Left-Out Fold

Train the decision tree by splitting the sub-sampled data, then splitting each resulting segment, and so on recursively until some constraint is met.

```
function OneFoldTree(nFold, iFold); /* #1 */
    /* generate _fold_ where */
    foldwhere1 = "_fold_ NE " || (String)iFold;
    foldwhere2 = "_fold_ EQ " || (String)iFold;
    mymodel = "gbt_" || (String)iFold; /* #2 */
    decisiontree.gbtreetrain  / /* #3 */
            table={name="new_heart_with_fold", where=foldwhere1}
            inputs=xx
            target="AgeAtDeath"
            casout={name=mymodel, replace=1}
            maxbranch=2
            maxlevel=8
            leafsize=60
            ntree=100
            binorder=1
            nbins=100
            seed=1234
            learningRate=0.1
            subsamplerate=0.7
            m=64
            ;
    decisionTree.gbtreescore result = r/ table={name='new_heart_with_fold',
            where=foldwhere2} model={name=mymodel}; /* #4 */
    print r;
    myscoredata = "gbtscore_" || (String)iFold; /* #5 */
    saveresult r  replace dataset=myscoredata; /* #6 */
    /* return prediction error; MSE or missclassification rate */
    return (r["ScoreInfo"][3,2]); /* #7 */
end func;
```

1. The FUNCTION statement creates a new function that can be called in an expression. In this example, the function is named *OneFoldTree* and has two arguments named *nFold* and *iFold*. For more information about 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 Cloud Analytic Services: 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 Cloud Analytic Services: 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.gbtreeTrain

Gradient Boosting Tree for NEW_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 Levels	8.000000
Min Number of Levels	6.000000
Max Number of Leaves	17.000000
Min Number of Leaves	9.000000
Maximum Size of Leaves	466.000000
Minimum Size of Leaves	60.000000
Random Number Seed	1234.000000

Output CAS Tables			
CAS Library	Name	Number of Rows	Number of Columns
CASUSERHDFS	gbt_0	2736	29

Run Each Model In Its Own Session

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

```
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 (*n*Fold) 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 (*n*Fold=1).

```
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.
58.744875114

NOTE:  The data set work.myscoredata  has 3 observations and 2 variables.
NOTE:  The data set work.myscoredata1  has 100 observations and 6 variables.
50.51471611

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.
57.160065856
```

```
/*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,
59.041830014,62.895729306,57.160065856}
```

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.

```
mse  = 0; /* #1 */
do i= 1 to nFold; /* #2 */
  mse = mse + ModelError[i];
end;
run;
```

1 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.

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

1 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
```