K-means based Clustering

Objective

Find the optimal value of number of clusters (K) for K-means algorithm for a given data set through scripting in SPSS Modeler.

Scripting in SPSS Modeler

Like any other scripting language, scripting in SPSS helps in automating tasks that would be highly repetitive or time consuming to perform manually. Here, a script could be a Stream script or a Standalone script or a SuperNode script.

Creating & Saving streams

```
create stream <stream-name>
save stream as <stream-name with full path>
```

Local Scripting Variables

- Declaration
  ```
  var <variable-name>
  ```

- Initialization
  ```
  set <varibale-name> = <value>
  ```

- Usage
  ```
  ^<variable-name>
  ```

Variable names, such as ^newVar, are preceded with a caret (^) symbol when referencing an existing variable whose value has already been set. The caret is not used when declaring or setting the value of the variable.

Creating Nodes

```
create NODE create NODE at X Y
create NODE between NODE1 and NODE2
create NODE connected between NODE1 and NODE2
```

NODE can be any valid node like *variablefilenode*, *typenode*, *tablenode*, etc. Nodes can be created using variables also. For example -

```
var x set x = create variablefilenode
rename ^x as "Input"  position ^x at 200 200
```

Referencing Nodes and its properties

Nodes can be specified by name or by type or by its unique ID. In case of multiple nodes of same type, specification by type will give error. For example, all three commands below will set *space* as delimiter for the variable file node.

```
set ^x.delimit_space = true set
Input.delimit_space = true set
:variablefilenode.delimit_space = true
```

Properties of any node can be accessed using dot (.) operator. Like in above cases, the property `delimit_space` of `variablefilenode` has been set to true. For complete list of properties of individual nodes, refer SPSS Properties reference manual.

Connecting Nodes

```
connect NODE1 to NODE2
connect NODE1 between NODE2 and NODE3
```

Conditional and Iterative command execution

- if...then...else

```
if EXPR then       STATEMENTS
1   else
STATEMENTS 2 endif
```

- for...endfor

```
for PARAMETER in LIST   STATEMENTS
endfor

for PARAMETER from M to      STATEMENTS
N endfor
```

K-means using Scripting

Given a input dataset, upper and lower limit of K, write a script to do following tasks -

1. Read given input dataset properly.
2. For each value of K from lower limit to upper limit

2.1. Connect source node to a K-means node. Specify value of K and run the model.

2.2. Connect source node to model nugget

2.3. Connect model nugget to a derive node and derive a new field as square of distance of the point from its cluster

2.4. Connect derive node to a "Set Globals" node and set it to set sum of squared errors(SSE) as a global variable

3. Connect a report node to source node. Configure it to write a *.CSV file as output where each line will contain value of K and corresponding SSE for all the values of K in the user specified range. Run the node.

4. Read above generated file into another source node.

5. Connect this source node to a plot node and configure it to draw a plot of number of cluster Vs. SSE. Value of K having abrupt change in slope represents an optimal value of number of clusters.

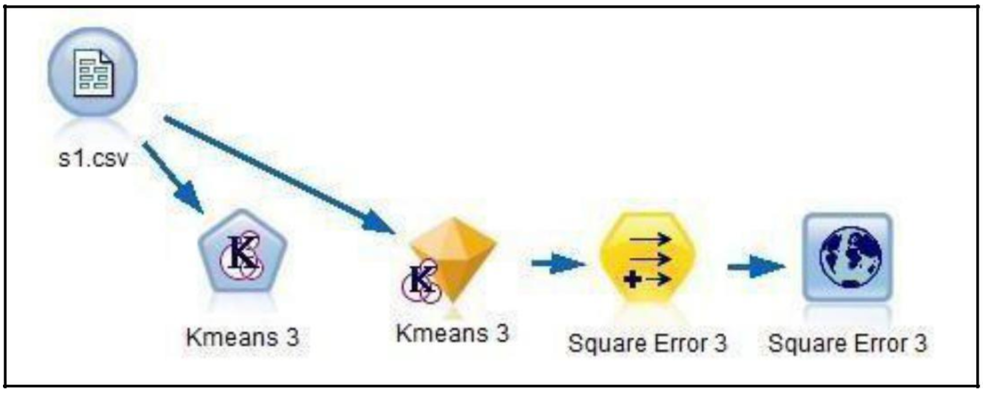

Fig 1 - Stream after one iteration of for loop in the above script

Sample Script

Script for reading a dataset and printing it to a table node

```
create stream "sample script"
var vname
set vname = create variablefilenode
set ^vname.full_filename = 'C:\ProgramFiles\IBM\SPSS\s1_modified.csv'
set ^vname.default_value_mode=Read
create tablenode at 250 250
connect ^vname to :tablenode
execute :tablenode
```

Assignment

Data for Q1-Q4: s1 data

1) Visualize s1 data and observe the distribution of data points. What kind of distribution exists? Is it a clustered data?

2) Apply k-means and observe clusters.

3) Insert some outliers in s1 data manually. Now apply k-means and compare the clusters with the clusters obtained in Q2. What difference do you get? Do you find variation in SSE’s?

4) Apply k-means multiple times and observe the clusters and SSE? What difference do you get?

5) Apply k-means on Iris Dataset.