ECLT5810 E-Commerce Data Mining Technique
SAS Enterprise Miner -- Clustering

I. Clustering Node

Some background:
- cluster similar data into groups

In EM, K-means Clustering

1. Initialization
   • Set no. of cluster (K) in the entire data set
   • Set K initial points to be the means of initial clusters (initial seeds)

2. Refine Clusters (iterative)
   • For each data in the data set, calculate its distance between the current clusters
   • Assign each data to its nearest cluster (some methods use other criterion to assign data to cluster).
   • Recalculate the cluster info. (mean, distribution, …)

Repeat Process 2 until the K clusters become stable.

N.B. Parameters need to adjusted:
1. no. of clusters
2. when evaluating the quality of cluster setting, you usually need to calculate distances between clusters ➔ distance measure for clusters
3. distance measure between data and cluster ➔ in EM, Euclidean distance is used:

\[
\text{dist}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

4. clustering criterion to assign data to cluster in the iterative process

Variables Tab

<table>
<thead>
<tr>
<th>Standardization:</th>
<th>None (default)</th>
<th>not standardized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>divided by the range</td>
<td></td>
</tr>
<tr>
<td>Std Dev.</td>
<td>divided by the standard deviation</td>
<td></td>
</tr>
</tbody>
</table>

Clusters Tab
- specify options for the Segment Identifier and to specify the Number of Clusters. The term segment (profile segment) refers to a cluster of observations
Number of Clusters
- User specify
- Automatic; press [Selection Criterion] button to change the way the tool automatically determines the number of clusters:

![Selection Criterion Image]

**Clustering method:** method to define distance between two clusters

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
</tr>
</thead>
</table>
| Average  | \[
| Centroid | \[
| Ward (default) | \[

**Clustering cubic criterion cutoff** (default is 3):

1. clustering cubic criterion (CCC): method to evaluate the quality of cluster
2. cluster setting with CCC > cutoff will be chosen

Minimum number of clusters (default is 2)
Maximum number of clusters (default is 40)

**Minimum Cluster Size**

1. Proportional to sample size:
   e.g. if 1. no. of training data = 10000
   2. Minimum Cluster Size in Final subtab = 500
   3. Proportional to sample size = 2000
   then
   Minimum Cluster Size = (2000*[(500/10000)]) = 100

2. Training value:
   Minimum Cluster Size = 500

3. User specify

**Seeds Tab**

**General Subtab**

![Seeds Tab Image]
**Clustering criterion:**
- given a data set, you have criterion used to construct clusters

<table>
<thead>
<tr>
<th>Clustering criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Absolute Deviation (Median) Deviation</td>
<td>Deviation</td>
</tr>
<tr>
<td>Modified Ekblom-Newton Deviation</td>
<td>Minimizes the sum of squared distance of data to the cluster mean, only one iteration is performed</td>
</tr>
<tr>
<td>Least Squares (Fast)</td>
<td>Same as above with more than one iteration</td>
</tr>
<tr>
<td>Newton</td>
<td>Deviation</td>
</tr>
<tr>
<td>Midrange</td>
<td>Minimizes the midrange distance of the data points from the cluster means</td>
</tr>
</tbody>
</table>

**Initial Subtab**
- specify how the cluster seeds are to be initialized

**Use incremental training for one pass:**
- the seeds are allowed to drift as the algorithm selects initial seeds

**Seed replacement:**
- specify initial seeds (initial seeds must be complete - contains no missing values)

<table>
<thead>
<tr>
<th>Seed replacement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>Seeds are chosen to be as far apart as possible</td>
</tr>
<tr>
<td>Partial</td>
<td>Only seeds not meet minimum distance requirement will be replaced</td>
</tr>
<tr>
<td>None</td>
<td>Choose first n complete data in the data set</td>
</tr>
<tr>
<td>Random</td>
<td>Randomly select complete cases</td>
</tr>
</tbody>
</table>

**Final Subtab**

**Convergence Criterion:**
- non-negative
iterations terminate when the maximum distance by which any seed has changed is less than or equal to the minimum distance between initial seeds times this value

Missing Values Tab
- specify how data observations containing some missing values are to be handled

An Example of Clustering

Data set : **SAMPSIO.DMABASE** (no need to set target variable)
Data Partition : Random; Training: 70%; Validation: 30%
Clustering : default

Running the Clustering Node → View results

Results Browser

**Partition** Tab
- provides a graphical representation of key characteristics of the clusters

Left: three-dimensional pie chart with the following default settings:
1. Std deviation (SD of instances in cluster)
2. Frequency (no. of instances in cluster)
3. Radius (size of cluster → distance of instance farthest from the cluster mean)

Right: A grid plot of the input means for the entire training data set

**Right-Click Pop-up Menu of Left Pie Chart**

*Select clusters to view…*
- enables you to select clusters, a sorting variable, and a sorting order.
- In the first run of the Clustering node, the pie chart may be difficult to interpret because there are too many segments. The pie chart may be much easier to interpret when a subset of the segments are selected for display

*Summary statistics…*
- shows summary statistics for interval and class variables of data in the selected cluster

*Cluster Profile…*
- displays a decision tree (path) for one selected cluster
- enables you to identify influential inputs (variables)

*View data…*
- displays a data table for the selected clusters

Save as data set...
- enables you to save the observations for the selected cluster(s) to a data set from Cluster Data Set window
- you can use a successor node to read and process the cluster subset data set

Delete subset data sets...
- delete the subset data sets that you created in the Cluster Data Set window

Variables Tab
- lists all of the input variables that are used in the Clustering analysis
- Importance value is computed to represent the relative importance of the variable to the formation of the clusters (1: most important)

Distances Tab
- provides a graphical representation of the size of each cluster and the relationship among clusters
- axes are determined from multidimensional scaling analysis
- asterisks : cluster centers
- circles : cluster radii

- To display a table of cluster distances select:
  * Pull down menu: Format → Distance Table*

Profiles Tab
- provides a graphical representation of the Categorical Variables and the Interval Variables for each cluster
- to swap viewing Categorical and Interval variables:
  * pull down menu View → Categorical / Interval Variables*

Categorical Variables
- pull down menu Format → Set Grid dimensions...(row: 2; column: 9)
- The row variable displays one categorical variable
- The slice variable displays a second categorical variable
- The height variable displays an interval variable

**Interval Variables**

**Statistics Tab**
- displays information about each cluster in a tabular format

<table>
<thead>
<tr>
<th>Frequency of the Cluster</th>
<th>number of cases in each cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root-Mean Square Standard Deviation</td>
<td>the root mean square error across variables of the cluster standard deviations</td>
</tr>
<tr>
<td>Maximum Distance from Cluster Seed</td>
<td>the maximum distance from the cluster seed to any case in the cluster</td>
</tr>
<tr>
<td>Nearest Cluster</td>
<td>the number of the cluster with a mean closest to the mean of the current cluster</td>
</tr>
<tr>
<td>Distance to Nearest Cluster</td>
<td>the distance between the centroids (means) of the current cluster and the nearest other cluster</td>
</tr>
<tr>
<td>(Remaining columns)</td>
<td>Mean of each Input Variable</td>
</tr>
</tbody>
</table>

- View Statistics Plot
  Pull down menu View ➔ Statistics Plot

**CCC Plot Tab**
- displays a plot of the Cubic Clustering Criterion plotted against the number of clusters that the Clustering node automatically selects.
- activate only if Automatic in the Number of Clusters check box in the Clusters tab is selected