INFS 5102 – Unsupervised Methods in Analytics

Practical #4: Cluster Analysis (2)

Objectives:

Continue to learn how to use the Cluster and Segment profile tools of SAS EM

Submission:

- *What to submit*:
  1. A PDF report generated by the Reporter node/tool of SAS EM, containing the information (diagram, results etc.) about the exercise done by you for this practical.
  2. Your answer to the questions, included in a word or PDF file. Remember to include the question/step numbers (d, …) and the questions in your answer file too.
- *Deadline of the submission*: 11:59PM (Adelaide Time), Tuesday of Week 6.
- *Submission link*: "**Submission Link of Prac #4"** in **Week 5 section** on Learnonline course site.
- *Marks:* Prac#4 (part of the ongoing assessment of the course) is worth 2% of the total marks of the course.

Instructions:

The **DUNGAREE** data set gives the number of pairs of four different types of dungarees sold at stores over a specific time period. Each row represents an individual store. There are six columns in the data set. One column is the store identification number, and the remaining columns contain the number of pairs of each type of jeans sold.

Name	Model Role	Measurement Level	Description
STOREID	ID	Nominal	Identification number of the store
FASHION	Input	Interval	Number of pairs of fashion jeans sold at the store
LEISURE	Input	Interval	Number of pairs of leisure jeans sold at the store
STRETCH	Input	Interval	Number of pairs of stretch jeans sold at the store
ORIGINAL	Input	Interval	Number of pairs of original jeans sold at the store
SALESTOT	Rejected	Interval	Total number of pairs of jeans sold (the sum of **FASHION**, **LEISURE**, **STRETCH**, and **ORIGINAL**)

a. Create a project named **Prac#4**, and in the project, create a diagram named **myPrac4**.

b. Create a data source with the **DUNGAREE** data set. (you can find the DUNGAREE dataset in Metadata Repository→Shared Data→Libraries→AAEM)

c. Determine whether the model roles and measurement levels assigned to the variables are appropriate.

d. Examine the distribution of the variables.

Are there any unusual data values? ____

Are there missing values that should be replaced? ______________________________

e. Assign the variable **STOREID** the model role ID and the variable **SALESTOT** the model role Rejected. Make sure that the remaining variables have the Input model role and the Interval measurement level.

Why should the variable **SALESTOT** be rejected? ______________________________

f. Drag the **DUNGAREE** data source to the diagram workspace to create an Input Data node for a process flow.

g. Add a **Cluster** node to the diagram workspace and connect it to the **Input Data** node.

h. Select the **Cluster** node. Leave the default setting as **Internal Standardization → Standardization**. What would happen if inputs were not standardized? ____________________________________

____

i. Run the diagram from the Cluster node and examine the results.
Does the number of clusters created seem reasonable? ____________________________________

j. Specify a maximum of six clusters and rerun the Cluster node. How does the number and quality of clusters compare to that obtained in part **h**? ____

k. Use the Segment Profile node to summarize the nature of the clusters.

What are the nature/characteristics of each of the 6 clusters, respectively? ____________________

(**Hint**: Based on the histograms showing the distributions of variables in each cluster, i.e. the histograms in the top-right "Profile _SEGMENT_" sub-window, use 1 to 2 sentences to describe the nature of each of the clusters with respect to the distributions of variables)

l. Use the Reporter node to produce a report as part of Practical #4 submission.

Save the PDF report, and include the report as part of your submission by following the instruction given on page 1 of the practical document.