HA-Lab 3: Clustering

Learning Objectives

This lab assignment aims to familiarize you with some of the clustering techniques that you have learnt in the course.

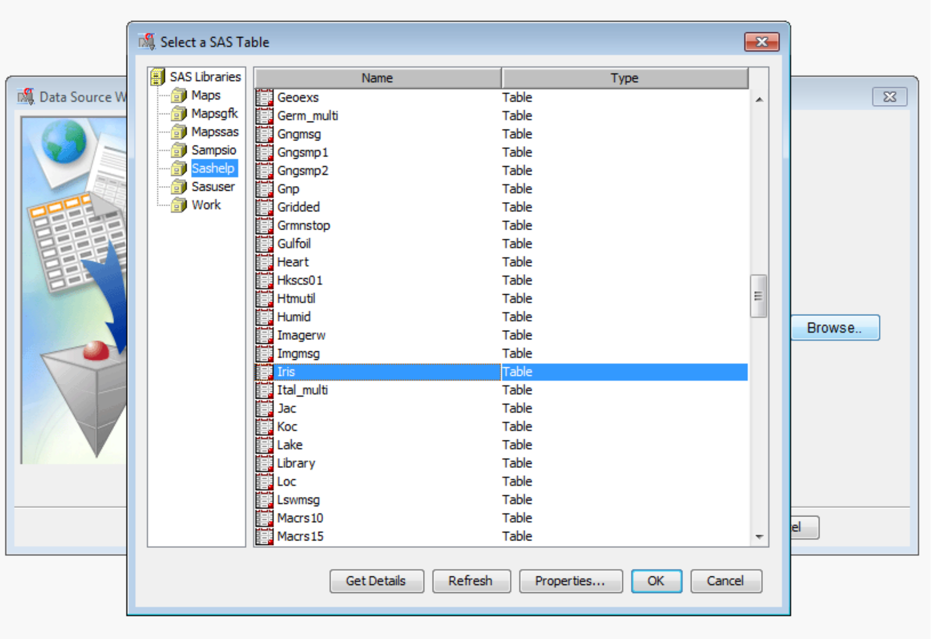
By the end of this lab, students will be familiar with the following concepts:

1. K-Means Clustering using SAS EM.
2. Effect of Standardization on clustering
3. Visualizing different number of clusters for the same dataset

Instructions

1. **Get the data:** In this part, we will first import the relevant data on which we will perform k-means clustering.

   a. Create a new Data Source: From **Sashelp,** choose the **Iris** dataset.

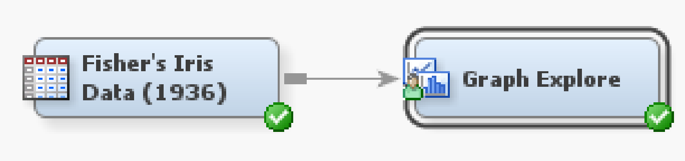


Name	Type
Geoexs	Table
Germ_multi	Table
Gngmsg	Table
Gngsmp1	Table
Gngsmp2	Table
Gnp	Table
Gridded	Table
Grmnstop	Table
Gulfoil	Table
Heart	Table
Hkscs01	Table
Htmutil	Table
Humid	Table
Imagerw	Table
Imgmsg	Table
Iris	Table
Ital_multi	Table
Jac	Table
Koc	Table
Lake	Table
Library	Table
Loc	Table
Lswmsg	Table
Macrs10	Table
Macrs15	Table

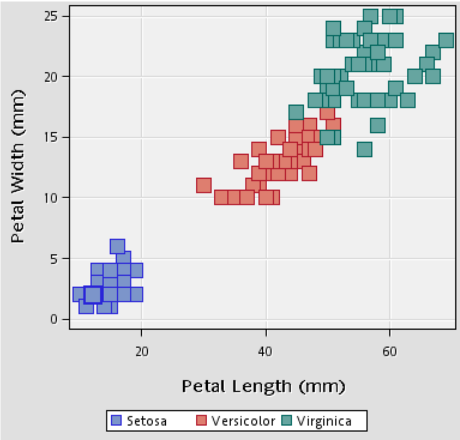
   b. Keep clicking next, and finally finish. Drag and drop the Fisher's Iris Data Source node onto a new Diagram Workspace.

2. **Clustering:** In this part, we will use the clustering node to perform k-means clustering on the Iris dataset

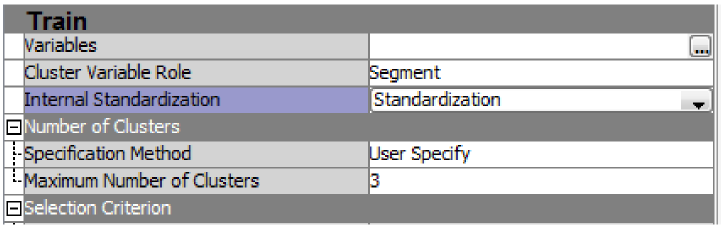
   a. Drag the dataset node and **Graph Explore** node to the diagram workspace. Connect, and run the two nodes.



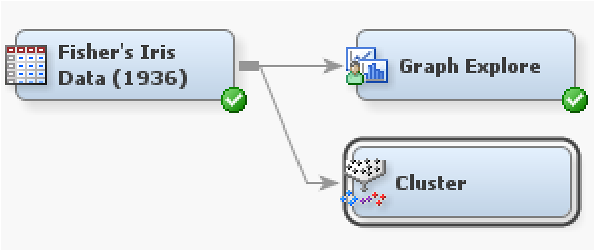
b. View the results of **Scatter Plot** in **GraphExplore.** For the scatter plot Choose roles for the variables – **Petal Length** as **X, Petal Width** as **Y, Species** as **Group.** You should see the following plot.



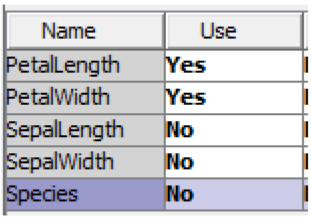
c. In this scenario, we know the labels to the datasets, and we can see a clear distinction in the values of the variables for each of the three classes. Now drag the **Cluster** node from **Explore** tab, and place it on the workspace. Connect it to the Data node, and change its properties. In number of clusters, change **Specification Method** to **User Specify,** and **Maximum number of Clusters** to 3. Note that an internal **Standardization** approach is being used by SAS EM (here all the values are divided by the standard deviation).



Train	
Variables	
Cluster Variable Role	Segment
Internal Standardization	Standardization
⊟ Number of Clusters	
Specification Method	User Specify
Maximum Number of Clusters	3
⊟ Selection Criterion	

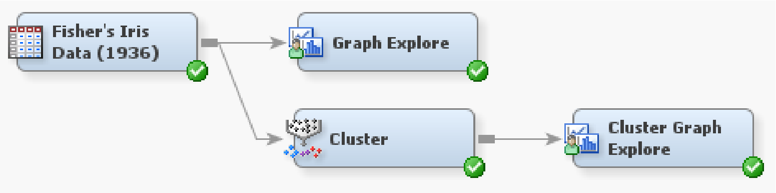


d. In the **Variable** properties of **Cluster** node, set the properties such that only **Petal Length** and **Petal Width** will be used for clustering.

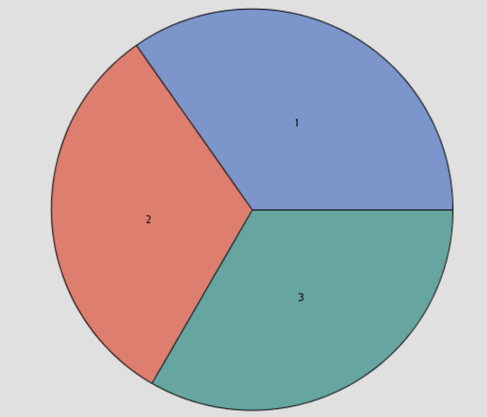


Name	Use
PetalLength	Yes
PetalWidth	Yes
SepalLength	No
SepalWidth	No
Species	No

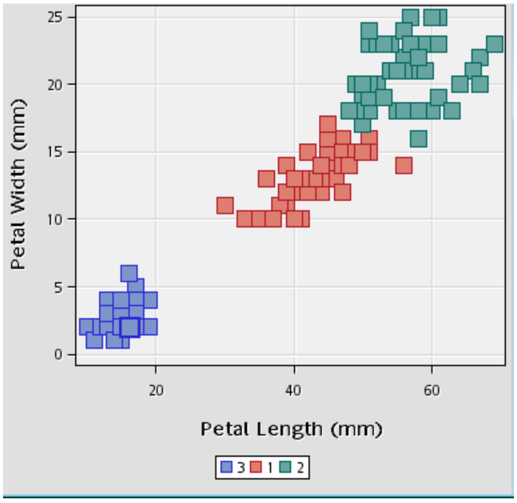
e. Drag another **Graph Explore** node to the workspace. Connect it to the **Cluster** node. Appropriately rename it. Run the nodes.



f. First, let us view the results of the **Cluster** node. The segment size tells us how many different. The **Segment Size,** tells us the number of elements for each of the 3 clusters. From our original data, we know each class should have exactly 50 instances, but here we see that each segment or cluster contains slightly different number of items. You can see the frequencies in the **Mean Statistics** as well.



g. Now we will try to visualize this clustering from the **Cluster Graph Explore** node. View this node's results, and make a scatter plot, but this time, the variable **_SEGMENT_** is assigned the role of **Group.**



h. Compare this cluster result to the scatter plot obtained with the class labels in part (b), and you can observe a few instances have been assigned differently.

Exercise

You have been asked to do cluster analysis and find out the important variables which will help in clustering the Iris Dataset. The dataset is the same that has been used in the instructions.

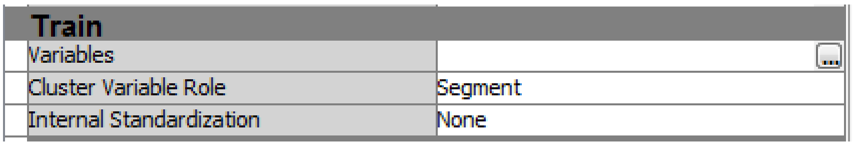
Please conduct appropriate analysis and answer the following questions in your report.

1. There are 4 variables in the dataset. Plot the scatter plot of the data using the following pairs of variables (using **Species** as **Group** role):
    a. Sepal Width and Sepal Length
    b. Petal Width and Sepal Length
    c. Petal Width and Sepal Width

2. Setting the number of clusters to 3, and the variables selected for clustering to (like in part (d) of instructions:
    a. Sepal Width and Sepal Length
    b. Petal Width and Sepal Length
    c. Petal Width and Sepal Width

    For each case, produce the scatter plot, with **_SEGMENT_** is assigned the role of **Group**

3. Repeat Question 2, but this time without data normalization. This means, for each clustering task, set the **Internal Standardization** to **None.** Comment on why results of this clustering analysis are different from those obtained in Question 2.



Train	
Variables	
Cluster Variable Role	Segment
Internal Standardization	None

4. For this dataset, we happened to know a priori the right number of clusters. In a real world scenario, we will not know the number of clusters. Additionally, the clusters may not correspond the "ideal" cluster that we want. Using **Petal Width** and **Sepal Length** as the variables**,** perform cluster analysis with number of clusters = 2, 4, and 6. Report the number of items in each segment for all 3 cluster analysis, and provide the appropriate scatter plots. (Remember to change the Internal Standardization to **Standardization**)

*Please attempt all the questions completely. All questions require you to produce 3 scatter plots each – and to obtain each scatter plot, you have to perform the entire cycle of cluster analysis (except for Question 1). Question 3 requires you to make comments on your results, and Question 4 requires you to report some information about each clustering analysis.

Report Submission

The length of the report should not be more than **4 pages**.
Your report should be titled "Lab 3-Clustering", followed by your name, and SMU email id.
This should be followed by your answers to the above questions.
Submit your lab report on E-learn Dropbox (Lab3 submission) by the deadline**.**