K-means Clustering Assignment in KNIME:

Create a KNIME workflow that utilizes the K-means clustering learning method to train a model on the Iris training data set. This data set can be found at the UCI machine learning repository.

http://archive.ics.uci.edu/ml/datasets/Iris

k-Means

The “k-Means” node groups input patterns into k clusters on the basis of a distance criterion and calculates their prototypes. The prototypes are built as the mean value of the cluster patterns. This node takes the training data on the input port and presents the model at the blue squared output port and the training data with cluster assignment on the data output port (white triangle).

The “k-Means” node can be found in the “Node Repository” in the `“Mining” -> “Clustering”` category.

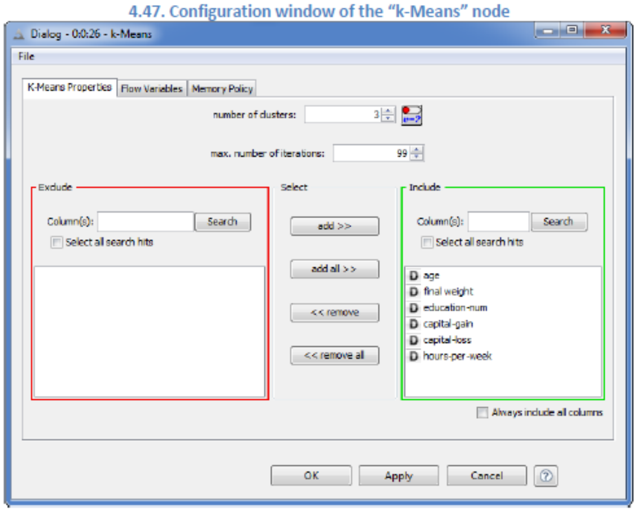
In the configuration window you need to specify:

- The final *number of clusters k*
- The *maximum number of iterations* to ensure that the learning operation converges within a reasonable time
- The *columns* to be used to calculate the distance and the prototypes
- Flag “Always include all columns” is alternative to the column selection frame.

Column selection is performed by means of an “Exclude”/”Include” frame.

- The columns to be used for the distance calculation are listed in frame “Include”. All other columns are listed in frame “Exclude”.
- To move from frame “Include” to frame “Exclude” and vice versa, use buttons “add” and “remove”. To move all columns to one frame or the other use buttons “add all” and “remove all”.

4.47. Configuration window of the “k-Means” node



number of clusters: 3

max. number of iterations: 99

Include:

- age
- final weight
- education-num
- capital-gain
- capital-loss
- hours-per-week

□ Always include all columns

Hint: Clustering algorithms are based on distance and therefore normalization of the attribute values is typically required before clustering.

Cluster Assigner can be utilized on the Iris data set to assign the test dataset to the associated cluster.

Task #1:

Create a KNIME workflow that utilizes the K-means clustering learning method to train a model on the Iris training data set. Set the initial value of K to default k=3. This corresponds well to the 3 existing classes of Iris categories as provided in the training data set.

Did every instance of the Iris flower assigned to the correct cluster? If not which one?

Task #2:

Download the bank training data set attached here and create your own K-means clustering model workflow.

The marketing department of a financial firm keeps records on customers, including demographic information and the type of accounts. When launching a new product, such as a "Personal Equity Plan" (PEP), a direct mail piece, advertising the product, is sent to existing customers, and a record kept as to whether that customer responded and bought the product. The data contains the following fields

id	a unique identification number
age	age of customer in years (numeric)
sex	MALE / FEMALE
region	inner_city/rural/suburban/town
income	income of customer (numeric)
married	is the customer married (YES/NO)
children	number of children (numeric)
car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
mortgage	does the customer have a mortgage (YES/NO)
pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)

Experiment with several different values for K. Analyze the cluster assignment for the ID12101. For k=3 and K=5 was the ID12101 assigned to the same of different cluster?

How about k=6 or k=7?

Task #3:

For K=6 what is the largest cluster?