Project Instruction

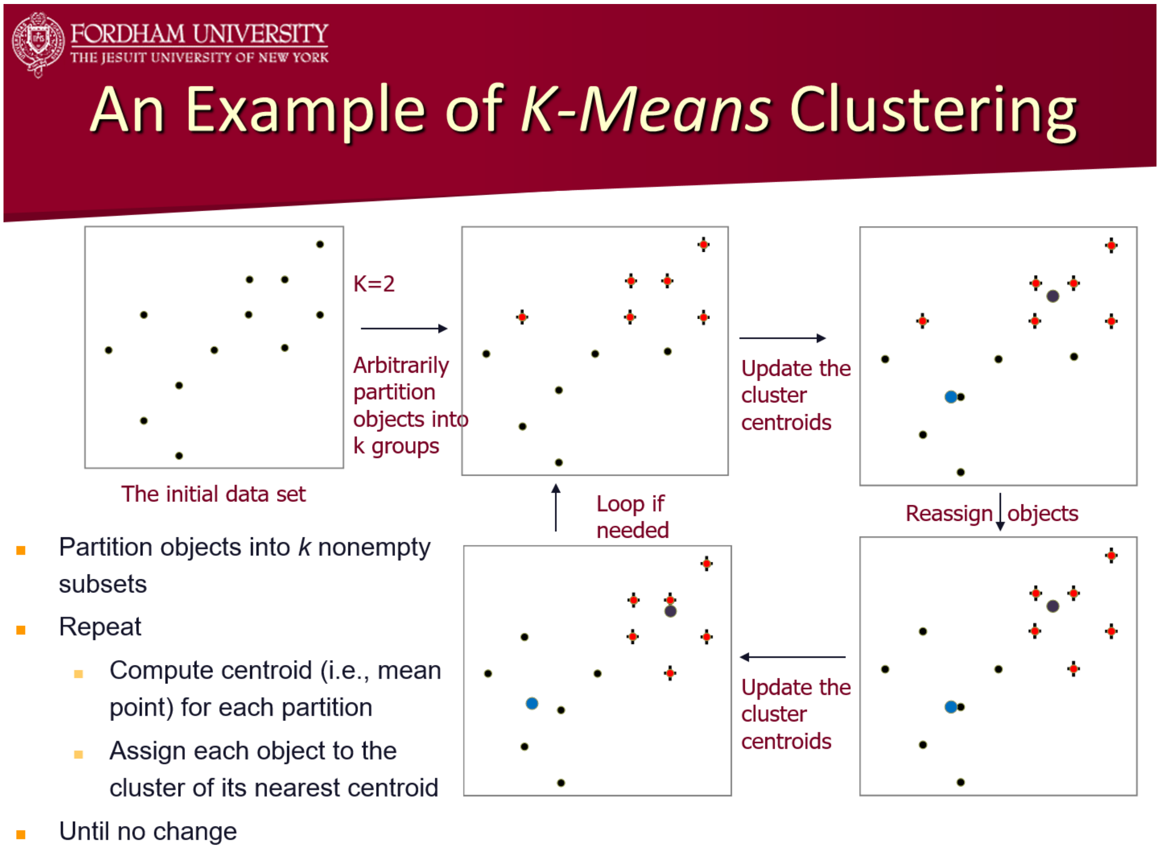
5 Dec 2017
Yuanyuan John Pei

The K-means clustering is a clustering method that could be applied in many business situations, such as differentiating products and discriminating customers.

This python script is based on a data set recording 1000 customers' weekly spending on a series of food categories, or *supermarket_weekly_customer.csv*. Each row records a customer. Column attributes include ID, Email, Gender, Age, and Spending on Categories, where Age and Spending are made input attribute to the model. The data set is generated by Mockaroo . The model divides the whole data set into a given number of clusters, by adding a new column to this csv: 'cluster'. Visualization is not supported by this python script; it has to be done in other tools such as Tableau.

Please note that this python file is NOT guaranteed to be compatible to other input csv files. The original csv file is uploaded as well. Please test the python script using that file.

At the very beginning the user should be asked to set 3 parameters: the number of clusters (1-10), maximum iteration (5-100), and convergence percentage threshold (50%-99%). As the model begins, a cluster column is arbitrarily assigned. In every iteration, a centroid is calculated for each cluster, then each node is reassigned to the cluster with the closest centroid. Convergence is tested by how much the cluster column changes in this iteration. The iteration stops until the threshold is met or the maximum iteration limit is reached.



FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

An Example of *K-Means* Clustering

K=2

Arbitrarily partition objects into k groups

The initial data set

Update the cluster centroids

Reassign objects

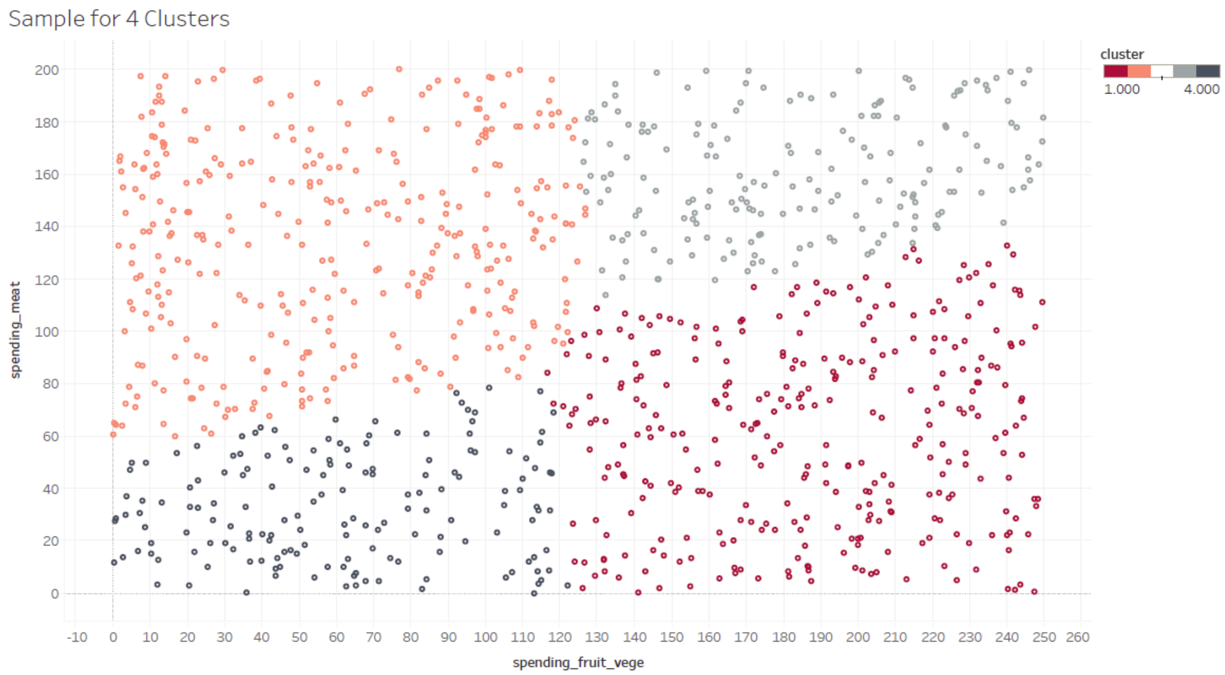
Loop if needed

Update the cluster centroids

- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

Lecture Slides by Professor Francis MacCrory

One sample result is obtained with 4 clusters, maximum iteration equal to 15, and convergence threshold equal to 60%. A new column 'cluster' is added to the csv file, and using Tableau I was able to generate this diagram. The 2 most important predictors are Spending on Fruit and Vege, and Spending on Meat.



Spending_fruit_vege vs. spending_meat. Color shows details about cluster.

This python script is 100% genuine, with only several lines of code using references from Stack Overflow. Please see note in the script.