K- Means Clustering

Clustering for data analysis

Robin Andersson 2016-01-17

Introduction

This assignment was focused on the subject of K-means clustering, a way of analyzing data and discovering patterns between all given objects. The data in this case is a set of 178 different wines, categorized into 3 different cultivars. Each wine has some statistical information, e.g. Alcohol, color intensity, hue etc. These will be used to determine if the wines in for example cultivar 1 will all be clustered together, or if there will be deviations of the wines within each cluster. This could as an example show that a manufacturer is doing something odd with it's grapes, since the wine is more like that of a different cultivar than the one it's supposed to belong to.

Data structures.

Looking at the problem, the first obvious problem to solve would be how to store the information about the wines in a structured way. As stated in the introduction, each wine has a series of attributes that will determine which cluster it will belong to in the end. As it is simple to see each wine as an object with similar attributes, the Class structure was used.

The next problem is to store all the wineobjects in a simple way. Since all wineobjects are individuals, we don't want to store them as groups of some special kind straight away, as is the goal of the assignment, but rather just store them together all in one single place. Therefore the List structure was used. This list will not be changed after all the wines has been added.

The clusters will have to keep track of which wines belong to them. This would most easily be made with a list. Also, this list is probably going to change a bit during the course of the algorithm working, so some functionality to keep this simple would be preferred. A cluster could also keep track of one of the given centroids, and add functionality for it too. With all this in mind, the Class structure is the given choice.

The wines are then to be stored within one of 3 clusters. For this the dictionary is used, where the cluster number (e.g 1, 2, 3) is used as keys, and the value is the cluster itself.

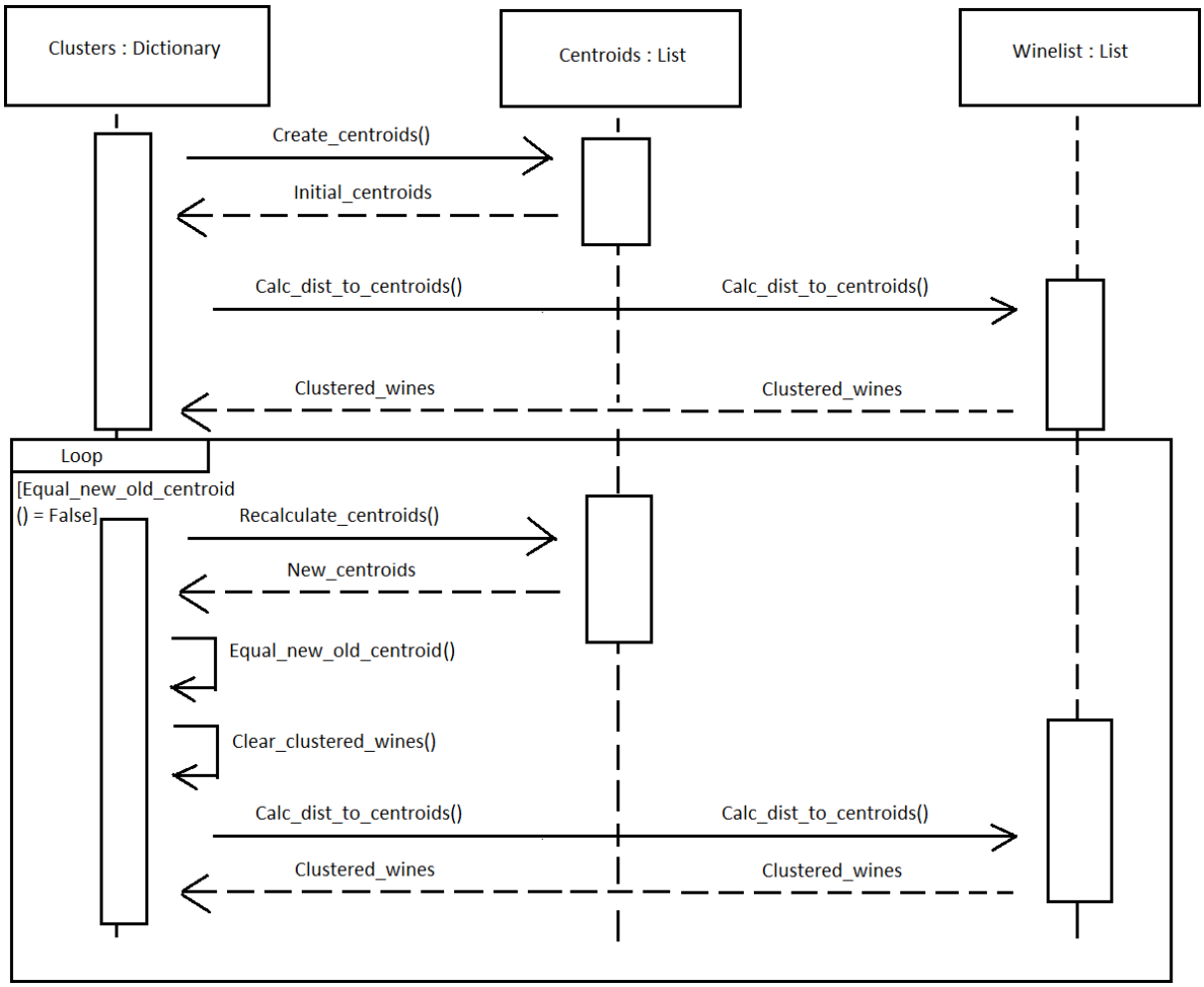
The centroids that will be used will also be stored within a list. These will then be used together with the wineslist in order to calculate the distance between each wine with every centroid. The closest centroid

The cluster class

The functionality of the cluster class will have to be a bit more clarified. As stated in above, the cluster class will take care of a number of things. First of all the wines that at the moment of the current iteration have the given centroid as it's closest centroid will be saved within a cluster object. These wines will also be used when calculating the clusters future centroid position, which is also a function within the cluster class. This wineslist will, after the newly calculated centroid, then be cleared of all entries and assigned wines once more.

Since there is a need to see what's inside a cluster there are also a couple of print functions inside which will help tell the result of the algorithms work.

UML Sequence Diagram



In the diagram above the overall process of the algorithm is presented.

At first the initial centroids are created, which in turn are used by the wines in the wineslist to calculate which centroid each wine is closest to. This function then returns a dictionary where the wines which are closest to the first centroid are all stored in the dictionary with key = 1, and so on.

After this the algorithm enters a loop stage, where the above process continues, with some differences:

- The centroids are recalculated based on the wines that are paired with it.
- After this calculation a test is performed to see whether the centroids has moved or not. If not, the algorithm has finished it's job.
- If the Equal function is false, the clustered wines are cleared from the clusters.
- Recalculate distance from wines to centroids (as before)

After the algorithm finishes the clusters are printed in the format of:

Cluster 1
Cultivar1: # of wines
Cultivar 2: # of wines
Cultivar 3: # of wines
Cluster 2
…