Lab Topic 08 - Classification using K-Means Clustering

CNM Peralta

Background

The K-Means clustering algorithm is an **unsupervised machine learning algorithm** that attempts to find the **innate clustering behavior** of its training data. It does so by attempting to partition *n* feature vectors into *k* clusters by minimizing the **within-cluster sum of squares**, that is, finding a clustering where the distance of each feature vector to its cluster's center is minimized.

The Algorithm

The K-Means algorithm is rather simple in nature:

1. Initialize *k* centroids randomly (choose *k* random observations from the training data). These centroids will represent each cluster.
2. Until the centroids no longer change:
    a. Correspond data points to the nearest cluster (compute distance to each cluster's centroid, then choose closest cluster). Remember:

$$d = \sqrt{\sum_{i=0}^{m-1} (x_i - c_i)^2}$$

where *x* is the current feature vector being classified, *c* is the current centroid to which *x*'s distance is being computed, and *m* is the feature vector length.

    b. Update centroids by averaging corresponding coordinates of the feature vectors. For example, the average of corresponding coordinates if the feature vector length is 2 would be:

$$c_i = \left(\frac{x_0 + x_1 + x_2 + \ldots}{count(class\ i)}, \frac{y_0 + y_1 + y_2 + \ldots}{count(class\ i)}\right)$$

If the feature vector length is 3, then it would be:

$$c_i = \left(\frac{x_0 + x_1 + x_2 + \ldots}{count(class\ i)}, \frac{y_0 + y_1 + y_2 + \ldots}{count(class\ i)}, \frac{z_0 + z_1 + z_2 + \ldots}{count(class\ i)}\right)$$

And so on.

Exercise

Create a program that classifies feature vectors using K-Means clustering.

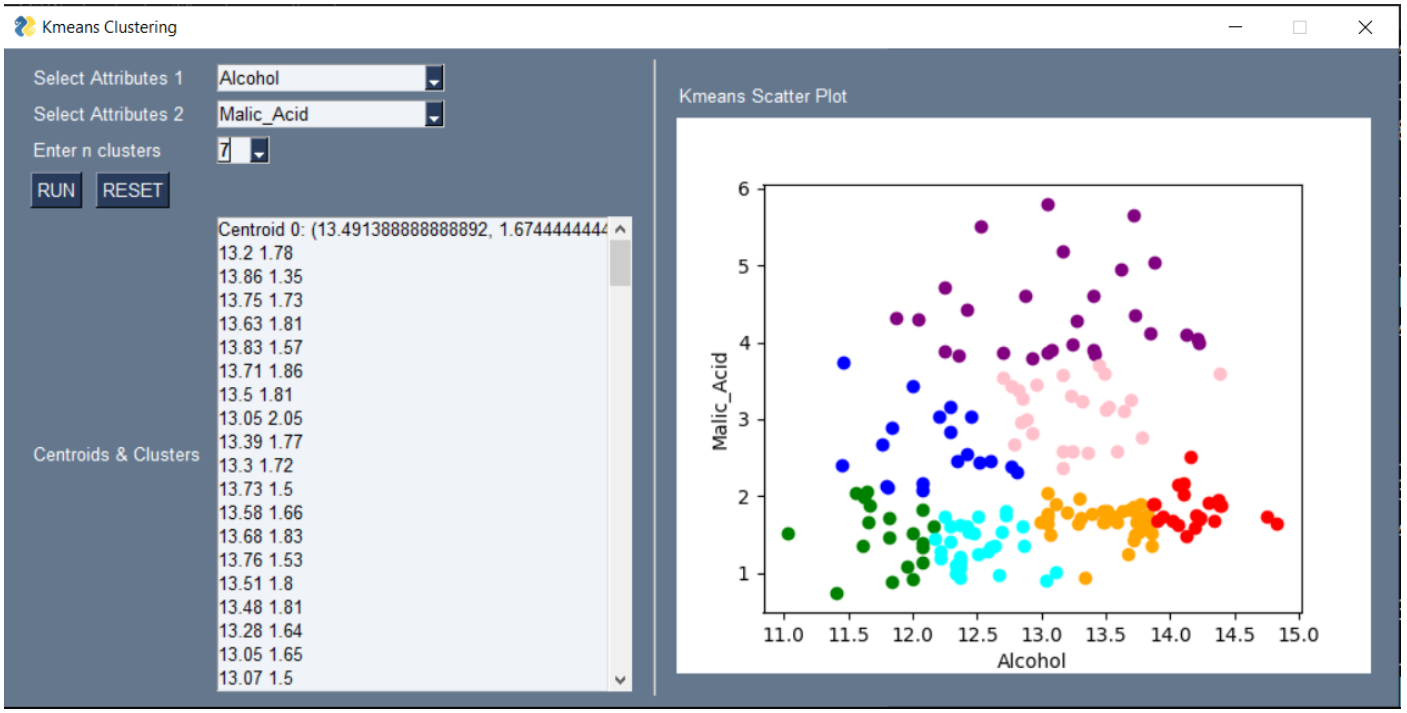
A dataset containing information regarding different chemical content of wine will be used. The task is to classify the points in wine.csv. You need to ask the user two columns of data which will be used to classify the points into k clusters. The number of clusters must also be asked to the user.

The output of the program is a file named output.csv and a scatterplot graph. Output.csv contains the centroids and the points under each centroid. For the scatterplot, each color signifies a cluster and maximum of 10 clusters. It must also output the centroids and the points under it in a scrollable list box.

Example Output (**output.csv**):

```
Centroid: 0 (16.98586956521739, 1.9304347826086963) [15.6, 1.71]
[11.2, 1.78]
[18.6, 2.36]
[16.8, 1.95]
[15.2, 1.76]
[14.6, 1.87]
[17.6, 2.15]
[14.0, 1.64]
[16.0, 1.35]

Centroid: 1 (22.179069767441863, 2.770581395348837)
[21.0, 2.59]
[20.0, 1.92]
[20.0, 1.57]
[20.0, 1.81]
[25.0, 2.05]
[22.5, 1.5]
[20.5, 1.81]
[20.5, 1.73]
[20.4, 1.61]
```



Scoring

- The criteria for this exercise is as follows:

Criteria	**Points**
Read input dataset correctly	1
Properly working UI for getting 2 attributes and n clusters	2
Compute distances correctly	3
Classify points correctly	2
Write output.csv correctly	2
Show scatterplot	2
Total	12

Reference

Stuart Russell and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall Press, Upper Saddle River, NJ, USA.