**Implementation of K-Means Clustering Algorithm from Scratch:**

**Bank Notes Authentication**

**Abstrac**t

This project is centered on differentiating genuine and forged banknotes using K-Means, an unsupervised machine learning task. Data were collected from images that were taken from genuine and forged banknote-like specimens. For feature extraction, mathematical tools, *Wavelets* were used. The final images have 400 x 400 pixels. Wavelet Transform tools were used to transform these features into numeric values (*variance* and *skewness* of the images). These two independent features were used to train the machine learning model to automatically detect forged banknotes. [Source](https://www.openml.org/d/1462)

**K-Means Clustering**

Machine Learning is all about building a model using data, statistical techniques, and algorithms to help the computer learn and make predictions. There are two types of machine learning based on whether it requires human supervision, Supervised and Unsupervised Learning. The former requires human help to learn the pattern from data, while the latter needs little or no human effort.

K-Means is an unsupervised machine learning algorithm that performs the division of data into clusters. Data in each cluster have some measure of similarity. There are two types of clustering (Hierarchical and Partitional or Centroid). K-Means is a partitional clustering that finds application in diagnostic systems, search engine, wireless sensor networks, and academic performance. K-Means is an iterative algorithm with the following steps

* Data standardization (Scaling the data)
* Cluster number selection (k = n) and select points in the dataset (centroids).Where n is a positive integer greater or equal to 2
* The distance from each data point from each of the centroids is calculated
* Assign each data point to the closest centroids
* Update the centroids by calculating the average value
* Recursively run steps three to five until the centroids stop changing.

Convergence is said to have occurred when the centroids stop changing. The value of k with least Within Sum of Square (WSS) was selected.

WSS = ∑ (x*i* – c*i*) 2

x*i* = data point, c*i* = centroids

**The Dataset**

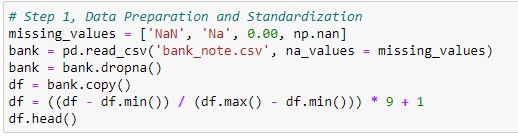
Data were extracted from images that were taken from genuine and forged banknote-like specimens. There are 1372 instances with two numeric continuous variables, V1 (variance of Wavelet Transformed image) and V2 (skewness of Wavelet Transformed image)

**Table 1: Statistical Description of the Data**

|  |  |
| --- | --- |
|  |  |

**Algorithm**

* Data normalization, is a way of scaling the data to ensure uniformity, the minimum and maximum values are 1 and 10 for each column.



* Random selection of centroids, these are randomly selected points in the dataset.

|  |
| --- |
|  |

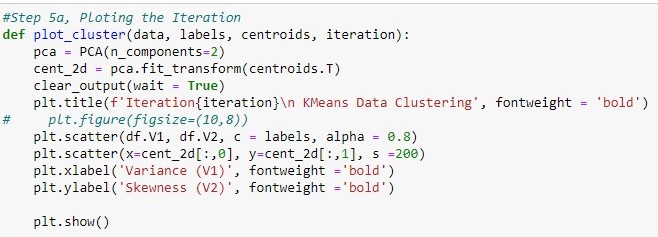
* Calculating the Euclidean distance distant from each of the data points from each of the centroids. The data points are assigned to the centroids with the minimum distant.

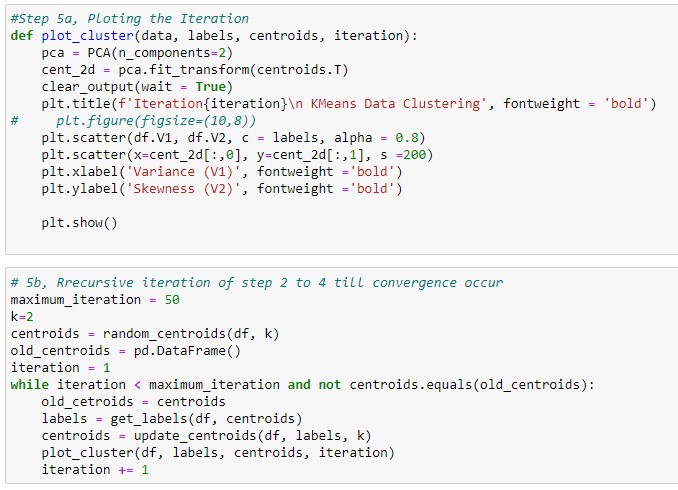
|  |
| --- |
|  |

* Updating the centroids, the average value over the members of each cluster is then set as the new centroids.

|  |
| --- |
|  |

* Recursive iteration of step two to step four until the centroids do not change, at this point, convergence is said to occur.





**Discussion**

The centroids appeared to fluctuate in the first few iterations because they were randomly selected. Convergence occurs under the first fifty iterations fig 4a.

There are 793 records in cluster zero which represent 62.10% as against 37.90% in cluster one totaling 579 records. The minimum and maximum variance of wavelet for data in cluster one are -7.0421 and 6.8248 and the skewness of wavelet are 0.5773 and 12.9516 respectively. For data in cluster zero, the minimum and maximum values of the variance of wavelet -5.2943 and 5.7403, whilst the skewness ranges from -13.7731 and 1.8127. The output of the algorithm was validated with the standard Scikit learn model using a classification report and confusion matrix (see appendix, fig 1a, 1b, 2a, and 2b). About thirty data points represent a 2.86% difference in cluster one and twenty represent 3.86% in cluster zero. This difference may be due to the measure of central tendency (mean) used in updating the centroids (Harmonic mean < Geometric mean < Arithmetic mean).

Forged banknotes are a global phenomenon and of great importance to currency management and monetary policy execution. This study shows that K-Means clustering has the potential to help cluster and segregate forged banknotes.

|  |  |
| --- | --- |
|  |  |

*Fig. 1a and 1b*

|  |  |
| --- | --- |
|  |  |

*Fig. 2a and 2b*

Appendix

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |

Reference

1. Banknotes Datasets, <https://www.openml.org/d/1462>
2. Machine Learning Algorithms, <https://learning.oreilly.com/videos/machine-learning-algorithms/9781789800289/>
3. Kaggle Machine Learning Codes algorithms-from-scratch

<https://www.kaggle.com/code/milan400/machine-learning->algorithms-from-scratch