Fake Note Detection Using K-Means Clustering

Sachin Mishra

Under the Guidance of

University of London & Goldsmiths, University of London

[Sachin19566@gmail.com](mailto:Sachin19566@gmail.com)

**PURPOSE OF THE PROJECT:**

Purpose of this project is simple and clear i.e to create a machine learning model which can distinguish between Fake-notes and real notes depending upon given attributes.

**DATA SETS**:

For this project I am using, data set from Open-ML. And then I am using first two column from that dataset.

This is a binary classification problem, where based on the given set of attributes, a note is labelled as a Not Fake(1)/Fake(2) note. There are a total of 1372 rows with 2 attributes and a label.

URL: <https://www.openml.org/d/1462>

**Brief Description of the dataset:**

* The dataset has 1372 rows (datapoints) and 2 columns(features) namely**V1(Variance) and V2 (skewness)**, the data is two dimensional.
* The **mean**value of the first feature is **0.43373526**. The **standard deviation** is **2.842763.**
* The **mean** value for the second feature is **1.922353.**

**METHOD:**

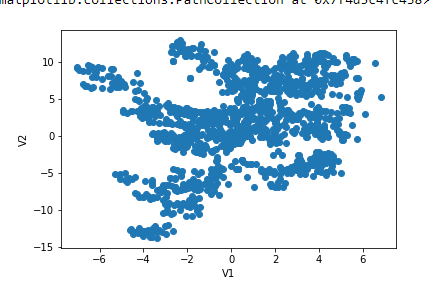
**Machine Learning Techniques:**

* I used **K-means algorithm** to implement the project.
* K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

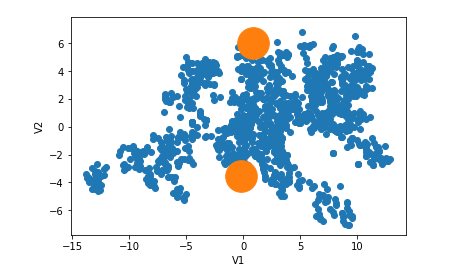
In this K-Means clustering I am making two clusters. One is for fake notes and another is for original notes.

What I am doing is basically , reading two columns from the given dataset. The first column is labelled as V1 which is nothing but Variance and second column is labelled as V2 which is nothing but skewness both derived from Wavelet Transformation of the image. And then calculating it’s mean column-wise. Then standard deviation also then I selected the number of clusters as 2. Then I visualize the data as shown below.

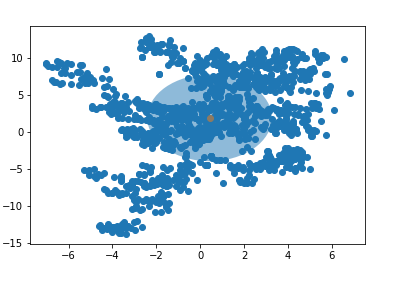
1. Scatter plot of given dataset



1. Scatter plot of the dataset with two clusters



C) Scatter plot of the data with cluster and mean as a centre of the cluster.



**SUMMARY OF THE RESULTS:**

# Yes, so by running k-means for several times I found it is stable for the given data set. And given dataset can be used to detect fake notes with low accuracy. However if we increase number of attributes then we can get more accuracy.

**Evaluation of it's suitability for K-means clustering**

* The two dimensional dataset has enough data points(1372) to train the **k-means algorithm.**
* However, the scale of the features V1 and V2 are different. So it's better to standardize (Z-score normalize) the data prior to running K-Means on it. So that none of the features dominate over each other.
* We need to find two clusters in the data dataset (and use them to do the binary classification ) using K-Means. The data seems to have 2 clusters and if the centroid are initialized properly , K-Means should be able to find the clusters.
* Finally ,as we can see inn the figure, there are many outliers in the dataset. The outliers may have negative impact on the accuracy of K-Means , since K-means relies on the distance(using Euclidean distance metric) from the centroid(mean) of the clusters ,and outliers have strong impact on the mean.

**Recommendations:**

* This machine learning model can be used to detect fake notes but since we have used only 2 features so it’s accuracy is not high.
* However if we increase the number of features then our machine learning model can deliver more accurate results thus increase in accuracy percentage.