CS3DS19 - Data Science Algorithms and Tools

Major Coursework

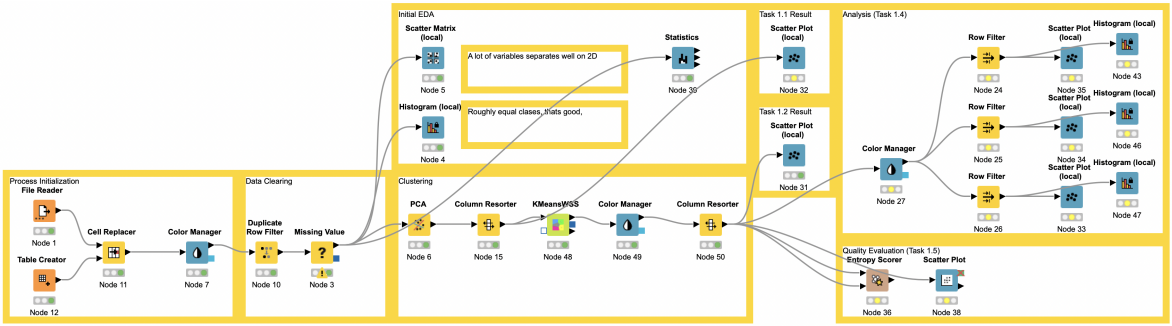
2021/2022

Student ID: 27020363

February 12, 2022

1 Task 1

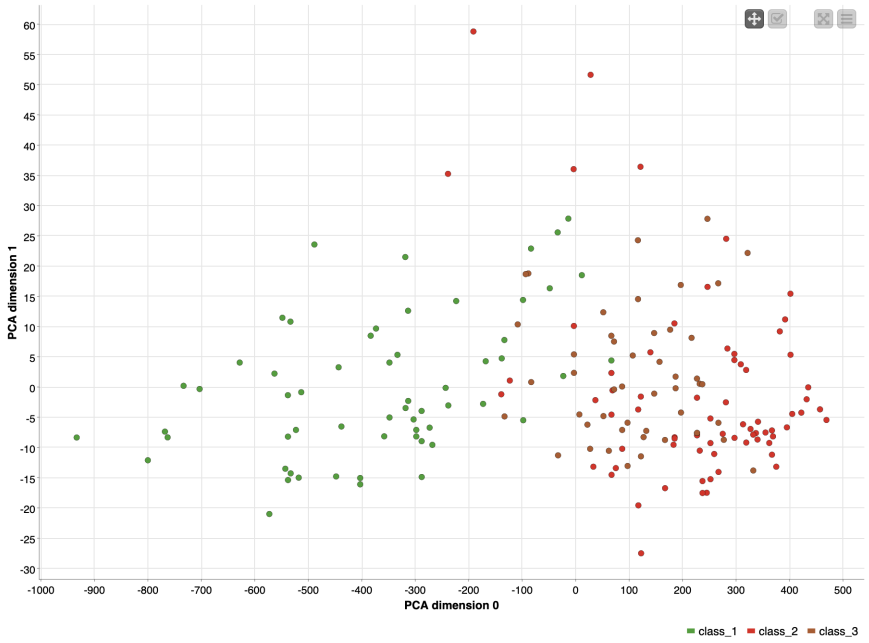
In this Section, the main parts of the workflow will be explained.



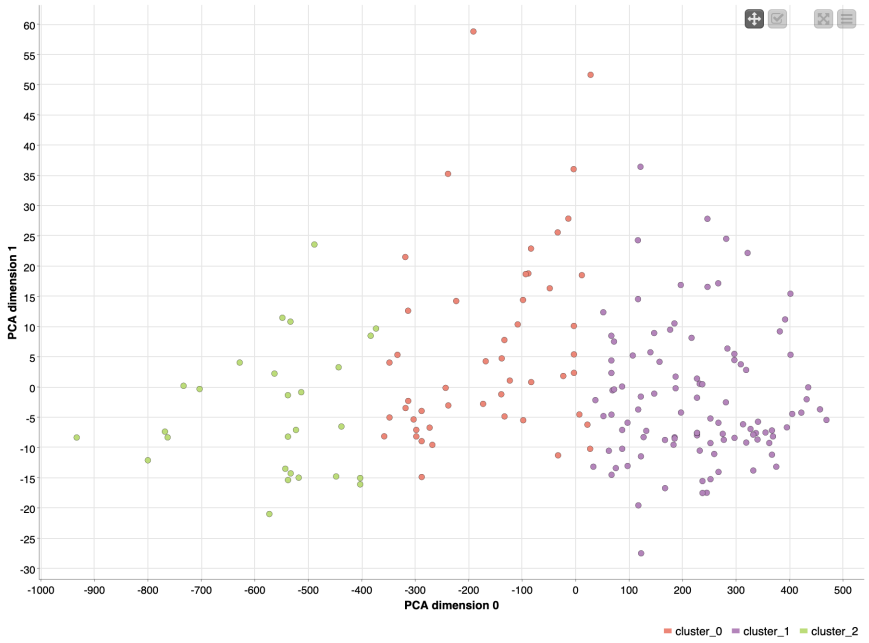
The workflow is splitted into 7 Stages. Those are

- Process Initialization : Read files, and transform the type of the class into string(was number). Finally we assign the initial colours for the subsequent plots, based on the classes field given by the dataset
- Data Clearing : Standard data clearing techniques, Drop duplicate data and missing values(by dropping all the rows involving missing data).
- Initial EDA : Some useful statistics that i used to get myself familliar with the data and their properties
- Clustering : PCA and KMeans Nodes(Futher Explaned below)
- Analysis : Evaluation of error, Classes distribution per cluster and histograms
- Quality Evaluation : Entropy scorer(Entropy/purity)

Task 1.1 : plot1



Task 1.2 : plot2

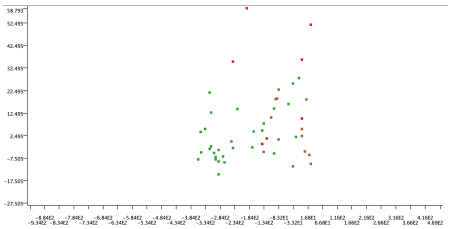


Task 1.3 : compare, discuss and explain the differences between plot1 and plot2

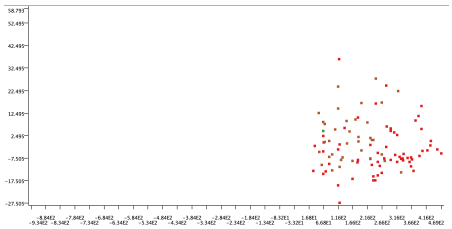
The main point of interest in the figures above, is the fact that, the clustering algorithm(K-Means, k=3) failed to appropriately partition the data, and recognise the classes provided by the dataset. We can see that there are multiple errors(Detailed analysis per-class on Task 1.4). This phenomenon can be explained due to the absence of the standardization process, something that will be explained in the next Task. The exact scale of the problem cannot be appropriately assessed with only those two plots though, we will need to examine the distribution of the classes in each cluster, something that will be done on the next task(Task 1.4)

Task 1.4 : plot3a,plot3b,plot3c

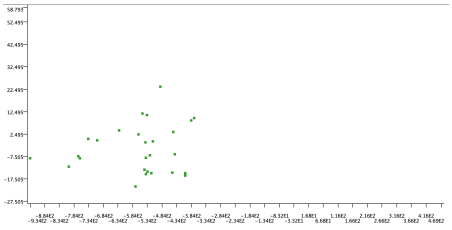
The 3 requested plots are shown below



(a) plot3a

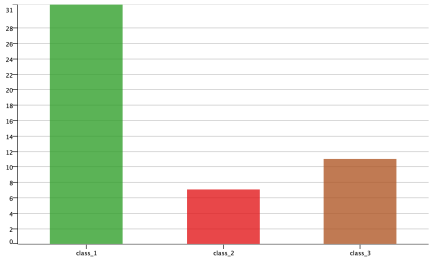


(b) plot3b

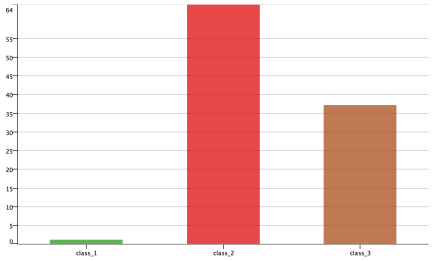


(c) plot3c

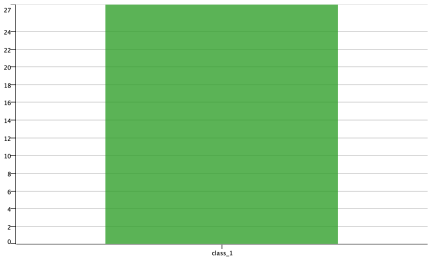
Here, we can see clearly each cluster and the classes of each datapoint that fails under those clusters. With use of histograms, we can learn the distribution of of classes in each cluster. Under ideal conditions(a good clustering), each cluster will contain the majority of the datapoints for some class, with a few missed points, unfortunately, this is not the case



(a) plot3a distribution



(b) plot3b distribution



(c) plot3c distribution

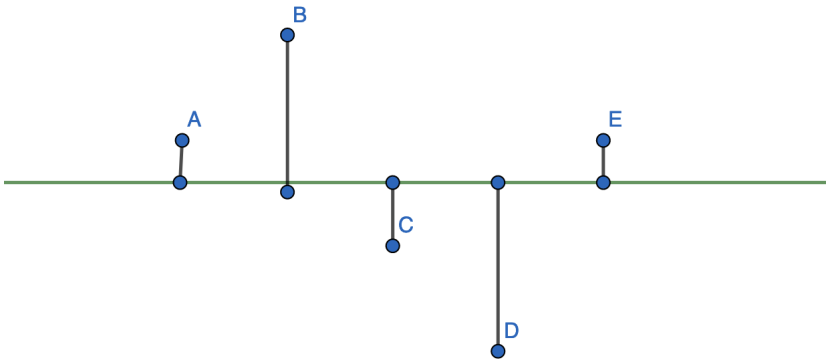
This is the result of the lack of a normalization/standarization process, something that we will explain on the next task.

Task 1.5 : Cluster Validity Measure

For our cluster validity measure, we will explain and intepret WSS (Within cluster sum of squares) and BSS (Between cluster sum of squares)

Prerequisite 1 : Sum of Square Estimate of Errors(SSE)

Before our attempt of explaining WSS and BSS, it is essential to explain SSE (Sum of Square estimate of Errors). SSE is an evaluation metric with a plethora of applications in statistics and predictive analytics[**?**]. It is used to evaluate a model's performance against a training set[**?**]. The logic behind SSE is rather simple in nature. Let X an random variable and a model $y = \bar{X}$



We can evaluate our model's performance, by calculating the Sum of Errors, where 'error' in this case, is the distance of the observed variable from the responce of our model (i.e the mean $\bar{X}$), as follows

$$SE = \sum_{i=1}^{n} X_i - \bar{X}$$

Unfortunately, our evaluation metric has a fatal flaw, it allows for error's to 'cancel out' simply because they are placed beneath our model's line. It can be proven that, for the case of $y = \bar{X}$ SE is always 0.

$$\begin{aligned} SE &= \sum_{i=1}^{n} X_i - \bar{X} \\ SE &= \sum_{i=1}^{n} X_i - \sum_{i=1} n\bar{X} \\ SE &= \sum_{i=1}^{n} X_i - n\bar{X} \\ SE &= \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i \\ SE &= 0 \end{aligned}$$

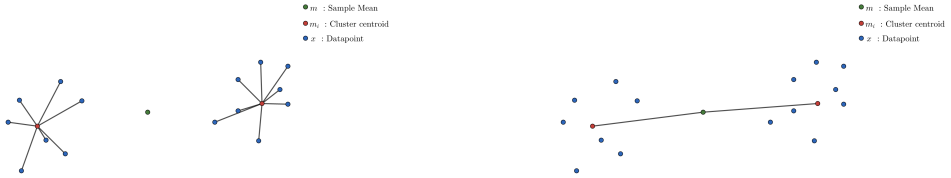
For other models(in higher dimensions, where corellation of the involved variables is not 1), it may not be zero, but the fatal flaw remains, by cancelling errors we severely underestimate the errors involved. A more appropriate approach could be to square the

distances, that would give us the Sum of Squared Errors (SSE)

$$SE = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

WSS and BSS

In the world of Clustering models performance evaluation, SSE is translated into two distinct but related ideas[**?**], WSS and BSS



(a) WSS (b) BSS

WSS

WSS or Within Cluster Sum of Squares is a clustering model performance evaluation function. Our 'model' in this case is the cluster centroid and the error is the distance for each datapoint from the cluster centroid. WSS can be evaluated with the following formula[**?**]

$$WSS = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2$$

Where k, the number of clusters involved, C_i the individual cluster and m_i, the given cluster centroid.

BSS

BSS or Between Cluster Sum of Squares is a clustering model performance evaluation function, just like WSS. Our 'model' in this case is the overall data mean and the error is the distance of each cluster's centroid to the overall mean. BSS can be evaluated with

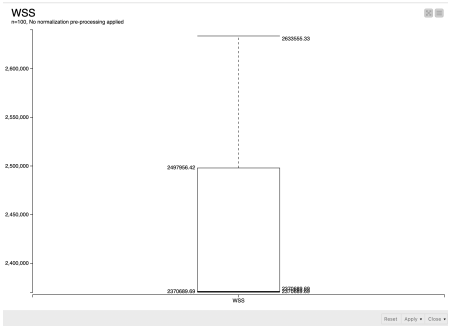
the following formula[**?**]

$$BSS = \sum_{i=1}^{k} |C_i|(m - m_i)^2$$

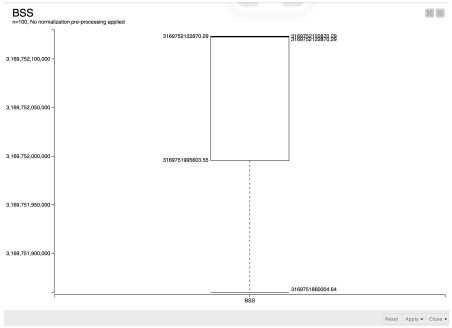
Where k, the number of clusters involved, C_i the individual cluster and m_i, the given cluster centroid.

Interpretation

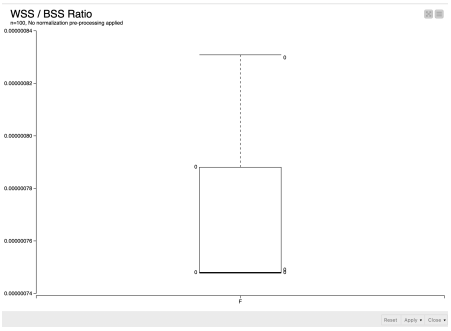
A good clustering solution, would involve well separated clusters, with low WSS to BSS ratio[**?**]. As K-Means is a heuristic based algorithm, we will not analyse WSS and BSS of a single K-Means run, but rather we will analyse their behaviour after a sequence of 100 runs.



(a) WSS



(b) BSS



(c) Ratio

As we can see, The ratio of WSS to BSS is very small, something that indicates well separated clusters(i.e a good clustering solution). However, a single cluster validity measure on its own, is insufficient to provide enough evidence for a meaningful clustering. Taking into consideration our observations from Task 1.3 and Task 1.4, this

clustering [**?**]is not meaningful, is full of errors, mainly due to the absence of the normalisation pre-processing step.

References

[1] *Shannon, C., 1948. A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), pp.379-423.*