Machine Learning - Assigment 2 Report

Mário Gomes, 54896, Analysis and Engineering of Big Data
Miguel Calado, 54491, Analysis and Engineering of Big Data

Abstract

In this second assignment of the Machine Learning course we were asked to analyze three clustering algorithms (K-means, Density-based spatial clustering of applications with noise and Gaussian mixture models) in order to better understand the structure of the dataset composed by seismic events. We begin by explaining the pre processement done on the data and then we center our analysis on the three algorithms, focusing on how they differ between each other. Finally, we end this report with a suggestion of the best algorithm and the range of values where its optimal parameter might be.

1. Data Preprocessing

At the beginning of this work it was provided a extensive dataset containing information on 3881 earthquake events, the first task was to create a new dataset from the extraction and transformation of the useful information of the original one, that is the earth fixed and centered coordinates of each seismic event (note that each earthquake event present on the original dataset could have happened due to either a seismic or nuclear explosion events), adding posteriorly a feature containing the fault information.

Given the nature of the data there's no need to standardize nor normalize the dataset in question, this is due to the form of calculation of the coordinates.

$$\begin{aligned} x &= 6371 * cos(latitude * \pi/180) * cos(longitude * \pi/180) \\ y &= 6371 * cos(latitude * \pi/180) * sin(longitude * \pi/180) \\ z &= 6371 * sin(latitude * \pi/180) \end{aligned}$$

Both $cos(.)$ and $sin(.)$ are limited functions who take values on the interval $[-1, 1]$, due to this mathematical property and the fact that each of the three features that result in the earth coordinates are multiplied by 6371, those features will take values on the same interval.

The shuffling of the dataset might be in order due to the nature of some of the unsupervised learning algorithms that we are going to apply. For instances, in the initialization of the K-means clustering algorithm the assignment of k data points as centroids might have to with its position in the data set.

In addition, because of the goal of this project, the rows with nuclear explosion were removed.

2. Metrics Adequability

In this work we were asked to use both internal and external indexes for the clusters' validation. As internal index it was used the silhouette score and as external index the rand, adjusted rand, precision and recall measures. In our opinion the silhouette score was the most important validation metric and the one that makes more sense in using given the nature of our data, for instances, high values of this metric are associated with more reliable cluster, since the points are more condensed inside of the cluster.

In the other hand we have the external indexes, however given the nature of this clustering problem and the knowledge of the faults beforehand this class of indexes becomes of less important compared with the internal ones. Let us consider for instances the adjusted rand score, a improvement of the rand score, given this one to be highly sensitive to the number of clusters, to the number of groups (associated with each fault) within each cluster and the variability of the metric. High values of the adjusted rand score are associated with the building of clusters associated with the same fault, which isn't of much interest given the prior knowledge of the faults. In the other hand low values of this metric are associated with clusters having a high variability of faults inside them, which can be a more interesting situation, since it can point to a high density fault region. But again, all of the information that external indexes can tell is already known, considering a earth map discriminating the tectonic faults.

3. K-Means

3.1. ***Clustering Procedure***

The K-means is a prototype based, complete and exclusive clustering algorithm, since all data points present in the dataset are allocated to exactly one cluster based in the distance between itself and each of the clusters' prototype (that is usually its centroid). This algorithm starts by randomly picking k data points as centroids of the k clusters and then each observation will be allocated to the cluster whose centroid is closest based on some dissimilarity measure, after the allocation phase the centroid of the cluster is recomputed.

3.2. ***Advantages and Disadvantages***

Given the way K-means performs the assignment of data points to a cluster this algorithm gives reliable results when there's groups of data points arranged in convex shapes and well separated between each other. The algorithm might be suitable for this particular problem since there can coexist a set of data points arranged in a spherical way along some faults.

Despite having a simple idea and easy to implement this unsupervised algorithm hides some several drawbacks like the lack of capacity to deal with noisy data and the presence of some outliers, which means the algorithm tends to group all observation sufficiently close to a clusters' prototype without checking whether they can be discordant from the other elements. Also the number of clusters has to be predefined *á priori* and the fact that the clusters tend to show similar shapes does not make sense in this dataset. Given these reasons we think that this algorithm does not fit the problem.

3.3. ***Metrics Discussion and Best Parameter Selection***

We will start this subsection by discussing the relevancy of each metric for this particular clustering algorithm. Our aim in the building of the k clusters and given the properties of the clustering algorithm is that they become as far from each other as possible, given no importance whatsoever to the fact that the algorithm might group points associated with different faults. Having this in mind, our metric based in which we will choose a optimal parameter is the silhouette score.

We started our study, analyzing the behaviour of the different metrics as a function of the number of clusters k, considering a range from 3 to 29.

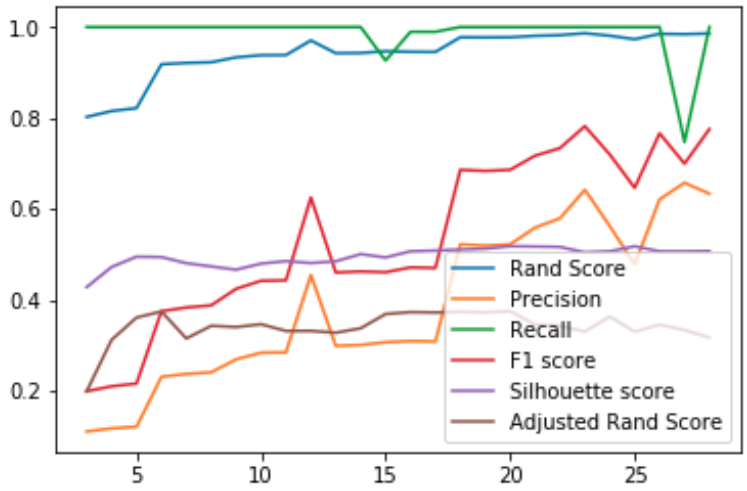

Figure 1: Metrics across a range of K values

Being 29 the total number of faults and 3, the minimal value from which we considered interesting to evaluate the metrics.

From the analysis on the plot associated with the internal index, we can see that the curve first reaches maximal values in the interval between 20 and 25 clusters. In which we considered to our best parameter k. In the finalization of this subsection, we would like to emphasize that given the k-means attempts to minimize the within-cluster sum of squares, that all cluster tend to have comparable spatial extent and that all data points must belong to some cluster, consequently having a greater distance between each other is related to the fact that in average the data points are more close to each other, thus the variability within each cluster is lower.

4. Gaussian Mixture Model

4.1. ***Clustering Procedure***

A Gaussian mixture model is a example of a complete probabilistic clustering algorithm, since one of the assumptions about the data is that it's generated from a linear combination of k Gaussian distributions, having each one their own localization parameter μ_k and dispersion parameter Σ_k.

The basic idea of this generative algorithm is given k Gaussian components, initialized when random pairs of parameters as well as the linear combination coefficients π_k, to find the maximum likelihood estimates of the

parameters based on the Expectation - Maximization algorithm that best explains our data.

The E-M algorithm, as the name suggests, is composed of a estimation phase where we compute the posterior probabilities $\gamma(z_{nk})$ of the n-th data point to belong in the cluster k, and a maximization phase where the parameters μ_k, Σ_k and π_k associated with each cluster will be recomputed according with the posterior probabilities, the process will repeat as many time as needed until it converges.

4.2. ***K-Means and Gaussian Mixture Model***

The E-M algorithm is one of the similarities between the Gaussian mixture model and the K-means, noting that in the last the $\gamma(z_{nk})$ are all equal and the co-variance matrix Σ_k equals the $\mathcal{I}_{p\times p}$, the μ_k is the prototype and the linear coefficients π_k are not calculated since the data points are assumed to belong to the closest cluster. This differences on the E-M algorithm lead to some other differences on this two clustering techniques, for example, the Gaussian model is allowed to form different shape clusters.

4.3. ***Advantages and Disadvantages***

The application of this clustering technique may be justified due to the possibility of a normal distribution of seismic events considering the Central Limit Theorem. Besides this, the fact that the algorithm does not tend to form clusters with similar dimensions and shapes makes sense given the nature of the data, for example, certain clusters can exhibit different shapes given the correlation between the gaussian distributions associated (given their proximity). As this model is a kind of probabilistic clustering, it can implicitly deal with outliers, in a way that the distant the data points are from the center of the gaussian distribution, the less likely the data point is to belong to its cluster.

Although the algorithm can show a good performance if the data shows a Gaussian behaviour, it may not perform well if this assumption is not met neither if the data displays some kind of dependency. Given the resemblance with the K-means algorithm, the user needs to know $'a\ priori$ the number of components/clusters, with which will fit the model in the data.

4.4. ***Metrics Discussion and Best Parameter Selection***

Considering Gaussian Mixture Model to be a probabilistic clustering technique our approach on the metrics discussion and parameter selection will be

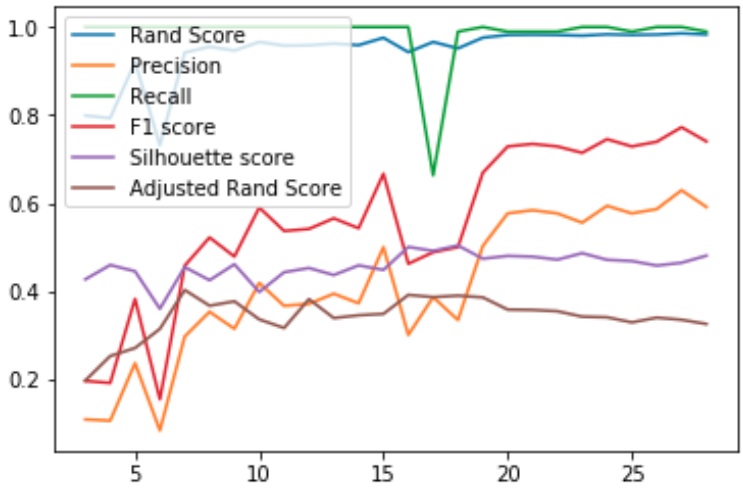

Figure 2: Metrics across a range of components

based on the identification of possible outliers. These outliers, given these work seismic context, may indicate the occurrence of intraplate earthquakes. Briefly this categoric of seismic events take place within the interior of tectonic as opposed to the interplate events that took place at the boundary of tectonic plates. This kind of occurrence might be useful to detect due to the heavy damage they can inflict.

Given this situation, the parameter selection will be based on the silhouette score and on the adjusted rand index. The adjusted rand score will be used in order to verify that each cluster contains mainly observations associated with some principal fault and the silhouette score to check if the cluster are properly separated between each other. The range for the number of components considered follows the same justification as in the K-means algorithm.

From the justifications shown above and the interest of applying this probabilistic clustering algorithm, we considered that the optimal number of components will be around the 18 clusters, value at which both scores considered take high values compared with the others.

5. DBSCAN

5.1. ***Clustering Procedure***

Density-Based Spatial Clustering Of Applications with Noise or DBSCAN, uses a different technique than prototype-based clustering achieved with K-Means, something that helps solve some of the problems that this type of algorithm has. DBSCAN considers clusters as areas of high density separated by areas of low density, that is, it finds core samples of high density and expands clusters from them. A cluster is a set of core samples, each close to each other and a set of non-core samples that are close to a core sample. Clusters found by DBSCAN can have any shape as opposed to K-Means, that assumes that clusters are convex shaped.

5.2. ***Advantages and Disadvantages***

DBSCAN is deterministic, always generating the same clusters when given the same data in the same order. Even though the core points will always be assigned to the same clusters, the labels of those clusters will depend on the order in which those points are placed in the dataset and the clusters to which border points are assigned can differ depending on the data order. Also, we dont need to specify the number of clusters in the data. DBSCAN can find arbitrarily shaped clusters. It identifies noise and is robust to .

One of its big disadvantages is that it cannot cluster data well with large differences in densities, since the combination of different parameters (which will be referenced in the next section) cannot be chosen appropriately for all clusters.

5.3. ***Selecting the parameter*** ϵ

To select the best value for parameter ϵ, we can follow the same approach used in the recommended paper. The methodology consists of plotting the sorted distance of each point to its fourth-nearest neighbor, this is the number that the authors recommend because through experiences made, the k-dist plots for $k > 4$ do not significantly differ from the 4-dist one and, in consequence, they need considerably more computation.

We started by first instantiating an object of the KNeightborsClassifier with 4 neighbors and fitted the model with the coordinate features of our dataset and with array of 0s with the size of the dataset. We only needed a dummy that simulates the target values to be considered by the classifier.

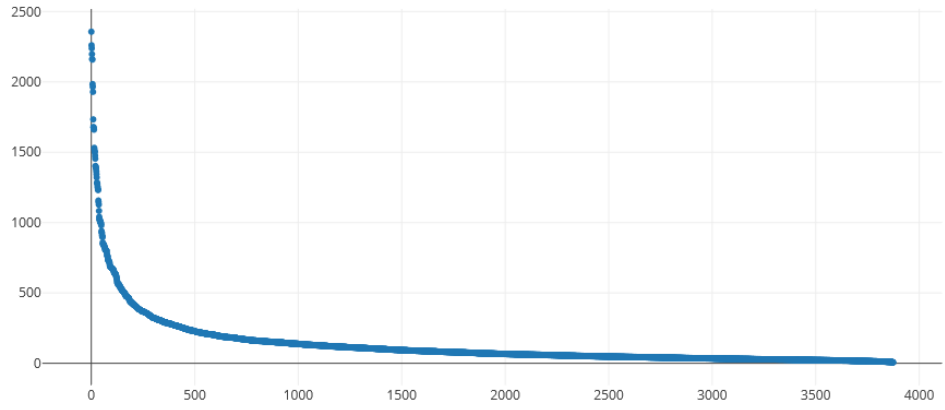

Figure 3: Elbow plot to select the best ϵ

Having done this, we can access the matrix to the k-nearest neighbors using the method kneighbors. After that sorted the distances to the 4th neighbor in descending order and plotted this on the y axis. The x axis has the values between 1 and the length of our dataset. The achieved plot is represented below.

To choose an interval of points that represent the ϵ we need to select an area near the plots elbow.

By observation of the plot, we can see that there are very large distances before the elbow, while after this area the distances are quite smaller. This allows us to conclude that the big distances, in the context of our dataset, correspond to very isolated points from the rest, while the smaller distances correspond to the points that are close together to each other.

5.4. ***Metrics Discussion and Best Parameter Selection***

The parameter to be optimized is ϵ since it allows to distinguish the distances that will characterize the points as noise or clusters candidates. Based on the method described previously, we can select a small range of values that are in the elbow area. We considered an interval of values between 200 and 600. The plot with the according metrics is plotted below.

By observation, we can note differences in the different metrics. Starting with the silhouette score, given its low values that derive from the fact that all points count for the calculation of this core, even those that were not assigned to any cluster. We can conclude that this metric is not appropriated for this

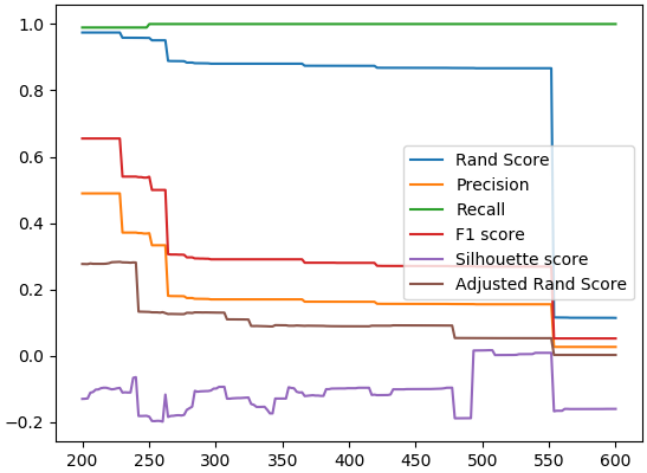

Figure 4: Metrics across a range of ϵ

kind of algorithms because it treats noise just like a cluster.

In addition, its also interesting to note that recall wouldnt be a good metric because it stabilizes at around 1.0. The more secure option would be to use external index metrics, with the sense that if an earthquake occurs, they most belong to the same fault, and if they are close points, they must belong to the same cluster. The metric chosen was adjusted Rand Score because of its adjustments.

In summary, looking at the Adjust Rand Score line, we can conclude that a interesting range of values for ϵ would be between 200 and 250.

6. Conclusions

In the conclusion of this work we would like to emphasize the idea that the best clustering algorithm is dependent of the information we may be interested to obtain from the dataset, however, the algorithm that shown more consistent scores regardless of the class of indexes used was the DBSCAN. Given the geographical nature of data this information comes as no surprise because the clusters generated from this technique will tend to aggregate data points close to each other and in the other hand the majority of the points will be close to the fault lines.

One of the possible application of this clustering algorithm based on this data, is the characterization of geological activity in that area, caused by the earth's internal heat.