Machine Learning - Milestone 1

Rafael Barón González - Ripoll
UCLM
Ciudad Real, Spain
Rafael.Baron@alu.uclm.es

Samuel González Linde
UCLM
Ciudad Real, Spain
samuel.gonzalez3@alu.uclm.es

Author Keywords
Machine Learning; Dimensionality Reduction; Principal Component Analysis; Clustering by Density; Hierarchical Clustering

INTRODUCTION

The purpose of this work is to explore the environmental data collected by various U.S. Federal Government Agencies from two cities (San Juan, Puerto Rico and Iquitos, Peru) to gain a better understanding of the Denge Spread Phenomena. These data are from a competition of the site DrivenData.

The overall objective is to use unsupervised learning techniques to make a preliminary exploration of the data and to extract conclusions from discarded elements, etc..

DATASET

The Dataset is a sample of the Dengue of two cities: San Juan and Iquitos, based on environmental data. Our sample contains 363 rows of data and it's only from San Juan. This sample contains information by weeks from the year 1992 to 1998 and it shows the humidity, maximum a minimum temperatures, precipitations, etc. by columns.

TRANSFORM

We begin removing columns that we don't need. As the data represent time by weeks, the column *'week_start_date'* is redundant, so it gets eliminated. The city is represented in the dataset as the original dataset has two different cities, but our sample is only from one, so the column *'city'* gets removed too. Then we check for the null values and found that 93 rows contains at least one null value. As 93 represent about a third of the total rows we can't just simply remove them, so we choose to change the null values to the mean of the others values.

DIMENSIONALITY REDUCTION

The number of features of a data set has an impact on the performance of the algorithms. So goal is to build a new dataset that preserves the information that the original has but with less attributes so we can reduce that impact to the performance. So we are going to reduce the number of features by checking the correlation that they have between them.

Correlation and feature reduction

Perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them. If we study the correlation matrix (corresponding to the Figure 1) we can observe two columns with the features. The color blue represents the correlation between them. A lighter color means less correlated and a darker color means higher correlation.

Excluding the diagonal of the matrix, which are the correlation that a feature has with itself, several features have a very high correlation or even a perfect correlation with other ones.

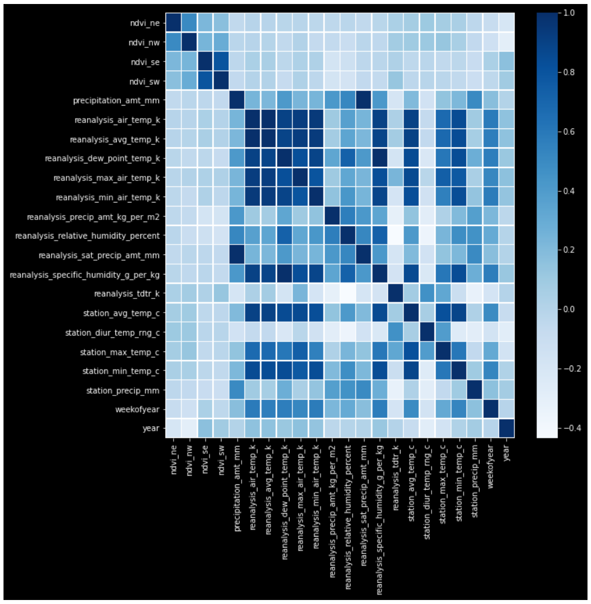

Figure 1. Correlation matrix

Therefore those values will add redundant information to our data, so we have reduce them. We are going to take into account that a value superior or equal to 0.9 is high enough to consider a feature very correlated to other one so we can get rid of them. The resultant correlation matrix is quite better than the previous one and has less features (Figure 2).

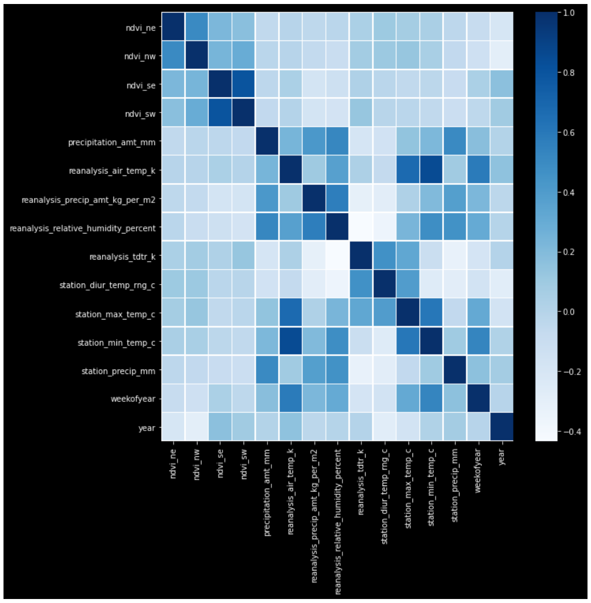

Figure 2. Correlation matrix

Principal Component Analysis

Principal Component Analysis is an orthogonal linear transformation that turns a set of possibly correlated variables into a new set of variables that are as uncorrelated as possible. First of all we have to normalize the data in order to make PCA work properly. For that reason we are going to use the StandardScaler normalization which is going to standardize the features by removing the mean and scaling to unit variance.

PCA reveals the internal structure of the data by taking our dataset that has a lot of dimensions (features) and flattening it to only 2 dimensions which gives us a better way of understanding the data (Figure 3).

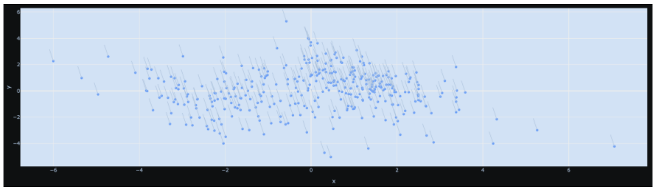

Figure 3. Scatter plot of the normalized data

OUTLIER IDENTIFICATION

In order to find outliers in our data we are going to use DBSCAN that uses the distance and a minimum number of points per cluster to classify points as outliers. Once we have detected those outliers we analyze them and decide if we are going to take them into account or not.

DBSCAN

The maximum distance between two samples for them to be considered in the same neighbourhood or *eps* is calculated from a range from 3 to 9, on 0.2 steps. With that information and the visualized data we see one big cluster with only some outliers, and with a *eps* value of 4.1 we get 1 cluster and 12 outliers, which we think makes sense according to the scatter plot of the normalized data (Figure 3).

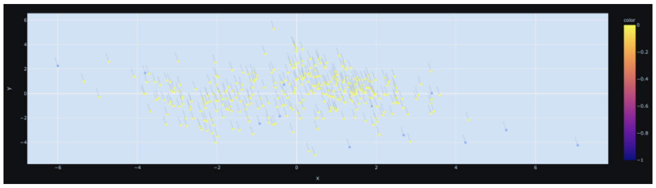

Figure 4. Outlier detection through DBSCAN

As seen in Figure 4 the algorithm finds two outliers right in the middle of the big cluster, we don't know why, but because it's only two values and the others are really outliers we leave it as it is.

Then we just simply drop the rows of the outliers detected and clean the dataset from outliers, leaving the dataset with 351 rows.

HIRARCHICAL CLUSTERING

Hierarchical clustering is a typical clustering analysis approach that creates nested partitions of the data layer by layer, grouping the objects into a tree of clusters.

As we did not take into account the outliers encountered by the dbscan, and we have dropped them from the dataset, we have to normalize again the data. We are going to do that by using the StandardScaler again. Once we have the normalize data we are going to apply the algorithm to compute a structure of groups organized in a form of a hierarchical tree, this is called dendrogram.

A dendrogram is a tree that shows how the clusters are organized. Each node is a cluster and each leaf node is a singleton cluster. The root of the tree is all the collection of elements and the leafs represents the groups of elements with more similarity.

What we see in the endogram (Figure 5) is five different groups *(threshold = 50)*. We think that means that there is some variance in the environment through the years but it's mostly the same *(purple group)*, only some times has change considerably *(green group)*.

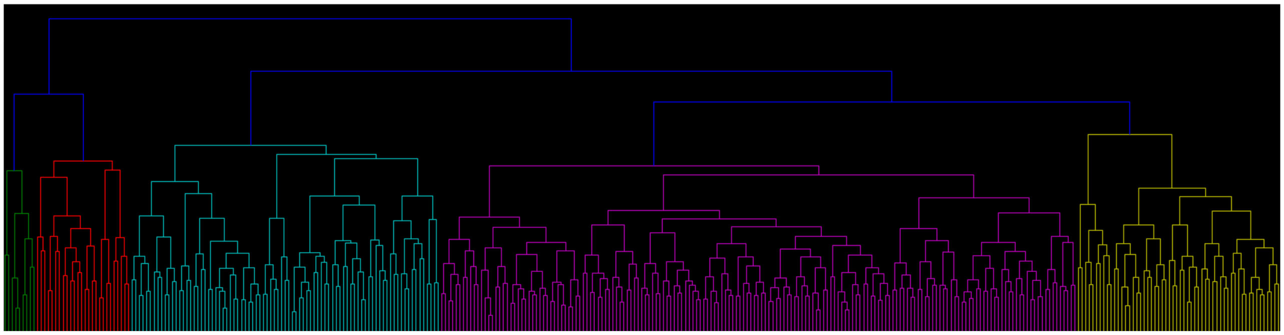

Figure 5. Dendogram