UCLM - Escuela Superior de Informática - Ciudad Real

Machine Learning
Lab Book

Made by:
Benjamín Cádiz de Gracia
Iván García Herrera
Dídimo Javier Negro Castellanos

Date: October, 2018.

Índice

1. Monday, 15 October 2018

1.1. Steps

- First, we have created a *cookiecutter* based directory structure.
- Then, we have revised the dataframe.
- We have deleted non numerical columns, it means, the columns that contain string values.
- We have deleted the columns which had all `NaNs`. If we don't have care enough and we try to delete rows with `NaNs` directly, we will delete the complete dataset, because there are some columns with all null values.
- We have removed duplicate rows, that is, two or more rows with exactly the same values. We have seen that in this dataset there are not duplicate rows.
- We have deleted the rows containing not a number (`NaNs`) values.
- We have tried to identify some outliers in some columns but we haven't have success.

2. Thursday, 18 October 2018

2.1. Steps

- We get the columns that ends with "MEAN" because that columns represent the average values. In that way the dataset is more representative. After that, we work on the visualization.
- We have done a `PCA` attempt but the `PCA` fails because we haven't deleted the version column and there are rows where the version number is not a float, it's a string like "x.x.x", so we eliminate the version column. The covariance is pretty high, about 0,85.

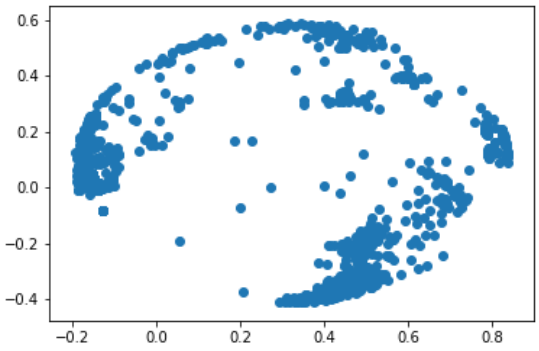

Figura 1: Principal Component Analysis

- Next, we start clustering with k-means. We analyse different groups in order to interpret the plots. We distribute the job: Ivan analyse the magnetic field, Benjamin the gyroscope and Dídimo the accelerometer.

- If we do k-means with the complete dataset, the computer is out of memory, that is the reason why we only consider the values of 5 days.
- We do k-means in a range of 2 to 50 clusters to know what is the ideal number of clusters. In order to get it, we use the *silhouette* method and we analyse which is the best coefficient. When we already know the ideal number of clusters, we plot the graphic, showing the different groups.

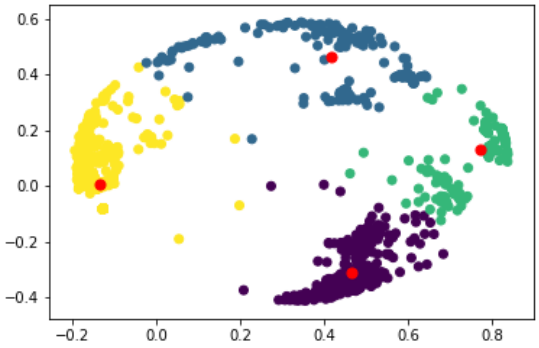

Figura 2: K-means clustering algorithm with best number of clusters

3. Friday, 19 October 2018

3.1. Steps

- We have cleaned the source code because we have seen that the implemented functionality was correct, so we have decided to organice the code in classes to avoid the repetition of code. In this way, the structure is clearer and it's easier to understand the code.

4. Tuesday, 23 October 2018

4.1. Steps

- We have studied the operation of clustering algorithm to identify outliers, which is DBSCAN
- We have based in the moodle example and we have tried to do `DBSCAN` with a symetric matrix but we didn't understand the results.
- Then we decided to use a conectivity matrix as input for a hierarchical clustering algorithm (`AgglomerativeClustering`). This conectivity matrix was obtained applying the *"k nearest neighbors"* algorithm. This algorithm returns a matrix that indicates what k neighbors is connected an element from the matrix `AgglomerativeClustering` returns a clusters hierarchy.
- We have applied the `DBSCAN` algorithm to know what points from the dataset are core points, what points are border points and which ones are noise. The noise points will be considered outliers.

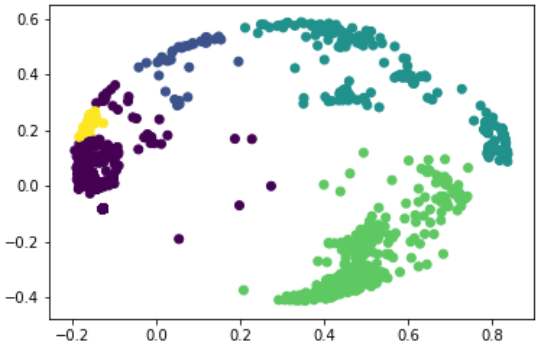

Figura 3: Agglomerative Clustering algorithm using a knn conectivity matrix

- We have plotted the `DBSCAN` graphics to see the outliers.

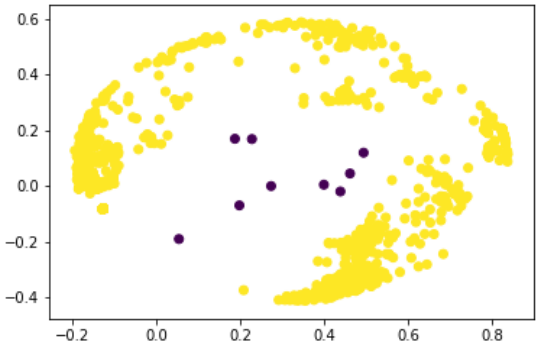

Figura 4: DBSCAN algorithm, used to identify outliers

5. Monday, 29 October 2018

5.1. Steps

- We have analysed and interpreted the K-means groups and the outliers.
- To achieve the previous point, we have grouped data by "*labels*" column and the columns with `X, Y, Z` contains the average of all data of distinct groups.
  To know what rows from the original dataset are outliers, we have added "*labels*" column to the dataset. This column contains a "-1" value if the row is an outlier. We have filtered the rows which had a "-1" value in the "*labels*" column and we have obtained the outliers.
- He have plotted both interpreted results, the k-means groups and the outliers in a 3D graphic, using the original dataset. In this graphic (see Figure 5), we can see where are the outlier points with regard to the other points of the dataset.

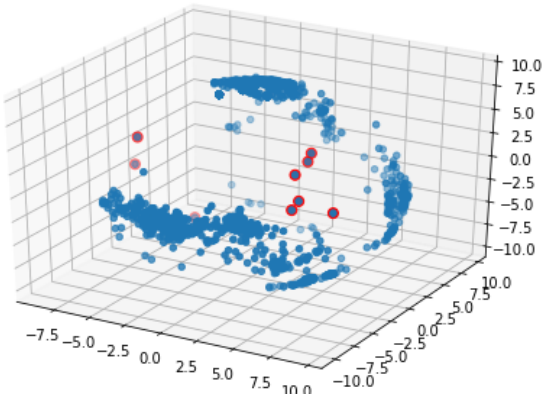

Figura 5: The outlier points are shown in red color. The blue points are all the other dataset points

6. Sunday, 4 November 2018

6.1. Interpreting results

6.1.1. Accelerometer

- In the first group we can conclude supposing that the person is stand up because the axis Y has got the maximum value.
- The second and the third group we can suppose that the person is walking because the values isn't so clearly.
- The fourth group has got values near to 0 and because that we can suppose the mobile phone is in repose.

Index	lerometerStat_x_M	lerometerStat_y_M	lerometerStat_z_M
0	0.248045	-9.46863	-0.520377
1	6.38539	4.4016	-2.16908
2	6.07473	-3.87995	-6.87392
3	-0.343393	-0.158427	8.7404

6.1.2. Gyroscope

- In the first and third group with the values we can't see anything clear. We can suppose that the person is walking.
- In the second group we can see how the mobile is in repose.

Index	'oscopeStat_x_ME	'oscopeStat_y_ME	'oscopeStat_z_ME
0	0.0848001	-0.350585	-0.0511788
1	-0.00948414	-0.005377	-0.00776391
2	-0.084057	0.349342	-0.0117479

6.1.3. Outliers

In the accelerometer outliers hasn't relationship with each other but we can see that this values are near with other outliers, so it's possible that the person had done something stranger.

On the other hand, in the gyroscope outliers is the same case that accelerometer, but we can see the all outliers values are near to 0. It's all we can say because we haven't got more data

7. Repository Link

`https://github.com/ivangarrera/MachineLearning`

8. Video Link

`https://www.youtube.com/watch?v=iqD1iOzJJwU`