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

We get the columns that ends with "MEAN"because that represents the median values. In that way the dataset is more representative. After that we work on the visualization.
We do an attemt and the PCA fails because we haven't 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 so low, that indicate data are not representative.
Next, we start clustering with k-means. We analice different groups in order to interpret the plots. We part the job: Ivan analice the magnetic field, Benjamin the gyroscope and Dídimo the accelerometer.
If we do k-means with the complete dataset, the computer are out of memory, that is the reason why we only considerate the values of only 5 days.
We do k-means in a range of 2 to 20 clusters to know what the ideal number of that. In order to get it we use the silhouette method and we analice which is the best coefficient. When we already know the ideal number of cluster we do the plot showing the different groups.