Laboratory Journal

Machine Learning Techniques

https://github.com/thmasker/ML_lab

Alberto Velasco Mata
alberto.velasco1@alu.uclm.es

Pablo Alcázar Morales
pablo.alcazar@alu.uclm.es

Diego Pedregal Hidalgo
diego.pedregal@alu.uclm.es

Beginning 10 October 2019

Contents

Thursday, 10 October 2019

1 Filter T2 dataset by weekday

We were assigned *'tuesdays'* from the given *T2 dataset*. The approach followed to extract these days was to create a *BussinessDay* matcher from the *pandas* library. A simpler approach would have been to just select those days in a hardcoded way, since there are only three tuesdays in the dataset: **2016-05-03**, **2016-05-10** and **2016-05-17**.

Thursday, 17 October 2019

1 Feature selection

After taking a look at the dataset format, we decided to choose only the features related to the 9-DOF sensors, that is, x, y and z coordinates of the accelerometer, gyroscope and magnetic field sensors.
The dataset provides this data with the FFT applied. We had to look for the way this method works, and after that we decided that the first four values of each axis were the features we wanted to work with. This lead us to a reduction from 248 features to 36.

2 PCA Estimation

After applying Principal Component Analysis to the previous 36 characteristics, we couldn't get an acceptable explained variance ratio greater than 75% with less than 5 features, so we decided to reduce the initial selected features again.

3 Only accelerometer features

We chose only the first four FFT values of the accelerometer sensor axes, getting 12 features (see Table 1)

Selected features
AccelerometerStat_x_FIRST_VAL_FFT
AccelerometerStat_x_SECOND_VAL_FFT
AccelerometerStat_x_THIRD_VAL_FFT
AccelerometerStat_x_FOURTH_VAL_FFT
AccelerometerStat_y_FIRST_VAL_FFT
AccelerometerStat_y_SECOND_VAL_FFT
AccelerometerStat_y_THIRD_VAL_FFT
AccelerometerStat_y_FOURTH_VAL_FFT
AccelerometerStat_z_FIRST_VAL_FFT
AccelerometerStat_z_SECOND_VAL_FFT
AccelerometerStat_z_THIRD_VAL_FFT
AccelerometerStat_z_FOURTH_VAL_FFT

Table 1: Features selected

We applied PCA again to these features. With two components we got a total explained variance ratio of *0.856*, so we can assume that these components are representative.
The next step we took was to see what was the relation between these two PCA components and the features (see Table 2).

- *PC-1* represents mainly the information from X and Y axes (and also a bit of Z axis).
- *PC-2* represents mainly the contributions from Z axis.

	PC-1	**PC-2**
AccelerometerStat_x_FIRST_VAL_FFT	0.297243	-0.147527
AccelerometerStat_x_SECOND_VAL_FFT	0.310501	-0.153492
AccelerometerStat_x_THIRD_VAL_FFT	0.314712	-0.148496
AccelerometerStat_x_FOURTH_VAL_FFT	0.315557	-0.148651
AccelerometerStat_y_FIRST_VAL_FFT	0.288487	-0.192882
AccelerometerStat_y_SECOND_VAL_FFT	0.306661	-0.200946
AccelerometerStat_y_THIRD_VAL_FFT	0.310738	-0.197411
AccelerometerStat_y_FOURTH_VAL_FFT	0.311343	-0.196712
AccelerometerStat_z_FIRST_VAL_FFT	0.209781	0.475163
AccelerometerStat_z_SECOND_VAL_FFT	0.245795	0.451615
AccelerometerStat_z_THIRD_VAL_FFT	0.262875	0.418231
AccelerometerStat_z_FOURTH_VAL_FFT	0.268944	0.388104

Table 2: Relation between PCA components and features

Tuesday, 29 October 2019

1 K-Means

After selecting the accelerometer features for analysis, K-Means is the following step we took in order to know a bit more about the selected data. By using k-means we expect to obtain different groups which will allow us to identify different behaviours in the data.

To decide the number of clusters to perform k-Means analysis, it is recommended to perform two different preoperations which can help us decide. That is why we computed the corresponding distortion and shilouette values for a given number of clusters (see Figure 1).

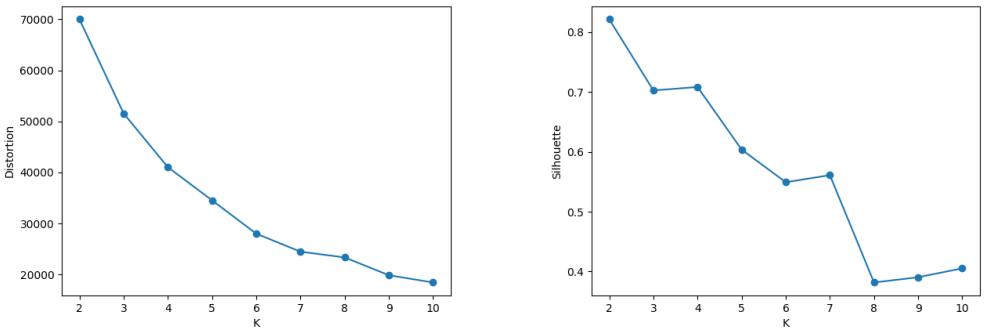

Figure 1: Distortion (left) and silhouette (right)

Clearly, there is no doubt about what the ideal number of clusters are: 4 (high shilouette and low distortion ratio). Taking 4 clusters, we proceed to perform the k-means analysis, which distributes data among this clusters, obtaining the result in Figure 2.

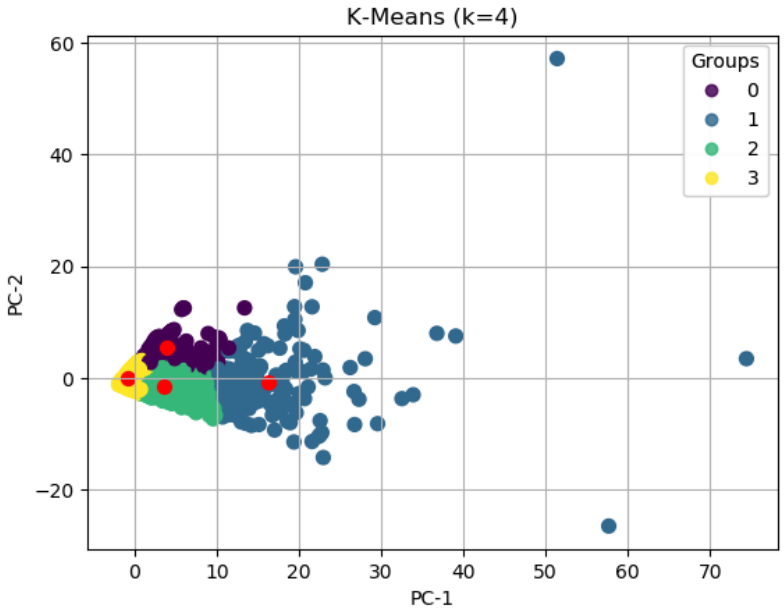

Figure 2: Data distributed in 4 clusters by k-means

As we saw after doing PCA, the second component (Y axis in Figure 2) corresponds mainly to the Z axis of the accelerometer. We interpreted the graph as follows:

- **Outliers**. There are three clear outliers, which probably are produced due to phone drops. These were extracted for later analysis (Table 1).
- **Steady phone behaviour**. Group 3 (yellow in Figure 2) is really aggregated around (0,0), so it might represent those moments were the phone is on top of a table or similar situations.
- **Main phone movement**. Group 1 (blue in Figure 2) is the most distributed group (including the aforementioned outliers). This seems like the group of interests since we want to focus on phone behaviours, movements and possible falls. It might be interesting to use clustering again on this group.

X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3	Y_4	Z_1	Z_2	Z_3	Z_4
3.07	3.07	3.08	3.27	8.80	8.89	9.12	9.15	3.12	6.74	9.67	12.30
2.71	3.07	3.28	3.49	8.41	9.38	9.57	10.16	1.46	1.53	1.66	1.73
1.93	2.27	2.31	2.64	0.98	1.03	1.29	1.38	11.90	13.42	13.58	13.72

Table 1: Outliers

To check if this interpretations are valid, we got the mean of each feature per obtained group (Table 2). It seems like we were right: **group 3** is really steady around 0 (no movement) and **group 1** is the one with the most changes in acceleration. Figure 3 represents this in a graphical way.

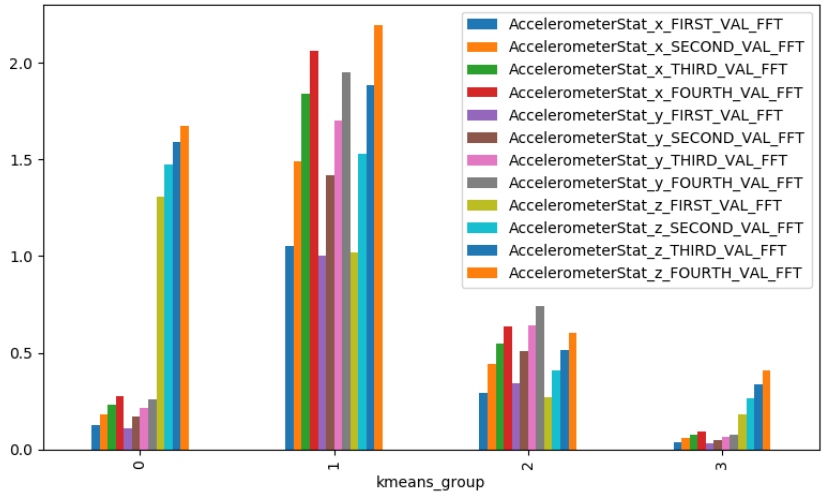

Figure 3: Mean of K-Means groups for each feature

G	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3	Y_4	Z_1	Z_2	Z_3	Z_4
0	0.12	0.18	0.23	0.27	0.11	0.16	0.21	0.25	1.30	1.47	1.58	1.67
1	1.05	1.49	1.83	2.06	1.00	1.42	1.70	1.94	1.01	1.53	1.88	2.19
2	0.29	0.43	0.54	0.63	0.34	0.51	0.63	0.74	0.27	0.41	0.51	0.60
3	0.03	0.06	0.07	0.09	0.03	0.04	0.06	0.07	0.18	0.26	0.33	0.40

Table 2: Mean of K-Means groups for each feature