Lab Book

Machine Learning

Córdoba Romero, Javier
Corroto Martín, Juan José
Guerrero del Pozo, Álvaro

Beginning 10 October 2018

Contents

Wednesday, 10 October 2018

1 First steps

Done by Juan José Corroto, Javier Córdoba and Álvaro Guerrero

We've initialized the repo with `coookiecutter` directory structure and learnt the purpose of each directory. We've also learnt how to work with `pandas`, read and write a csv and how to work with spyder.

2 Data cleaning

We've picked the data of only the first day, we've done this by looking at the raw data and looking at the column *"TimeStemp"*.
Then we've deleted the *UUID, Version* and *TimeStemp* columns because they were non-numerical values.
Then we've tried deleting the rows that had *NaN* values, and we ended up deleting the whole dataframe. So, after scanning the dataframe, we have realized that the columns

1. *RotationVector_cosThetaOver2_MEAN*
2. *RotationVector_cosThetaOver2_MEDIAN*
3. *RotationVector_cosThetaOver2_MIDDLE_SAMPLE*

Had all their values as *Nan*.
After eliminating these 3 columns, we proceeded as before: we eliminated those rows that had *NaN* values and deleted 2 rows.
Then we saved this new data as processed data in its corresponding folder: data/interim/.

Wednesday, 17 October 2018

1 Feature Selection

Done by Juan José Corroto, Javier Córdoba and Álvaro Guerrero

First, we have modified how we pick the first day data: instead of getting the 2038 first rows, we convert the *'Timestemp'* to days and choosing the day 28 (the first one).

In addition to the columns dropped the last week, we have also dropped *UUID*. Then every *FFT* and *Middle Sample* rows have been dropped too.

We have chosen just two sensors to do an initial exploration: **Accelerometer** and **Linear acceleration**. The purpose is to try and explore the data related to the movement of the phone.
So, we create a new dataframe with only *Acelerometer* and *Linear Aceleration*. Finally, we drop *Covariances*, as we don't think they will be useful cause they are calculated as the relation *2-by-2* of the axis.

2 Visualization - Clustering

We then calculate the *PCA* with an *explained_variance_ratio* of **0.78** and **0.14**, which is a very good representation of the original 18 features we were studying.
With this results, we plot a *scatterplot* to visualize it(Figure 1).
After that we run the *k-means* algorithm with different number of centroids and print the *silhouette* and *distortion*. We initialize the centroids using the *K-means++*algorithm as well. You can see the measurements of distortion and silhouette of KMeans with different number of centroids in figures 2 and 3 respectively.

By looking at the plots, we have decided to choose **4** as the final number of centroids.

Finally, we run the *k-means* algorithm with said number of centroids and plot the results, each cluster having a different color, and the centroids being colored in red.

Before trying to get meaning from the clusters, we have decided to repeat this experiment by doing some exploration on the features.

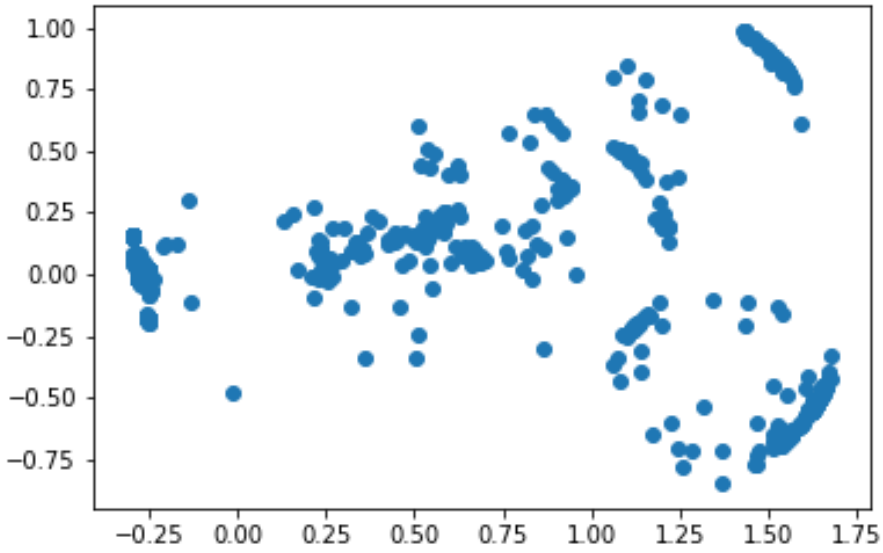

Figure 1: Plot of the data after PCA

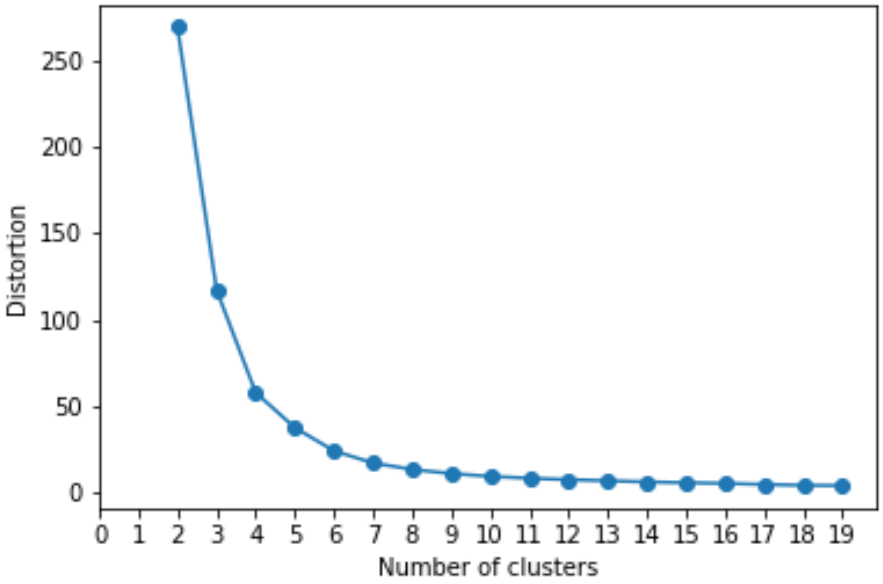

Figure 2: Distortion of K-means with different number of centroids

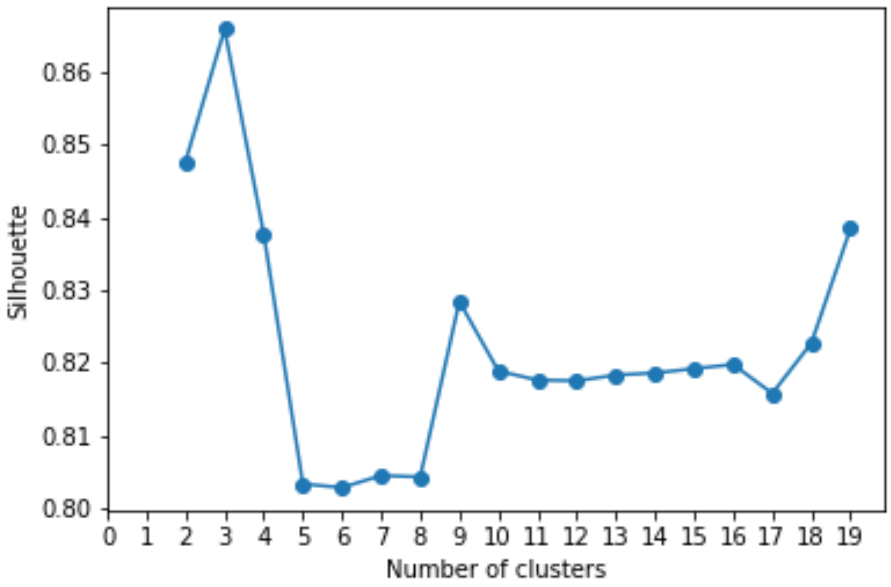

Figure 3: Silhouette of K-means with different number of centroids

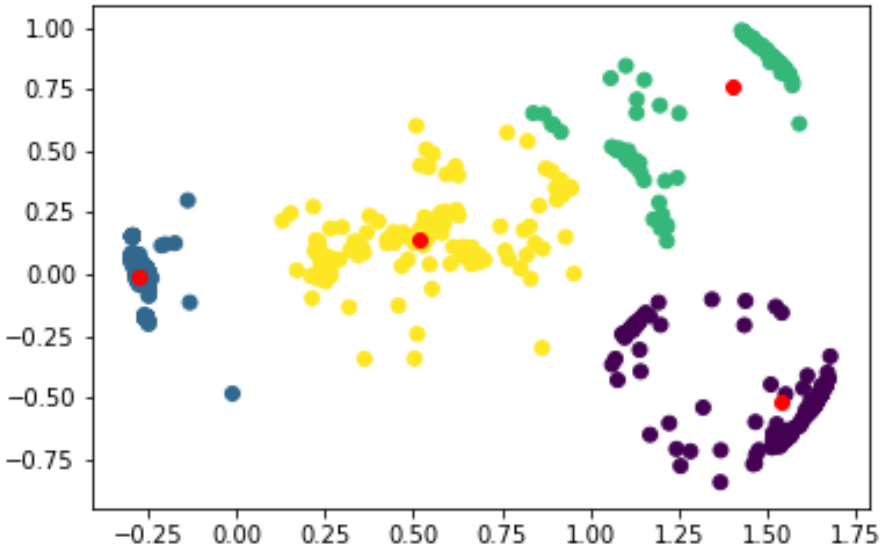

Figure 4: Plot of the data after K-means (centroids in red)

Friday, 26 October 2018

1 Feature Selection done well

Done by Juan José Corroto.
We have repeated the same process as before: pick the day 1 data from the whole database, but this time, we are going to perform some analysis on the features in order to remove those features that gives us less value due to redundancy, instead of removing some columns randomly. First, we still remove all columns that have to do with the *Fast Fourier transform*, since we don't know how it works and we can not extract knowledge from them. The same with the *Middle Sample* columns. After that we remove *NaN* columns and rows as we were doing proviously.

Now we are going to see the correlation between the features:

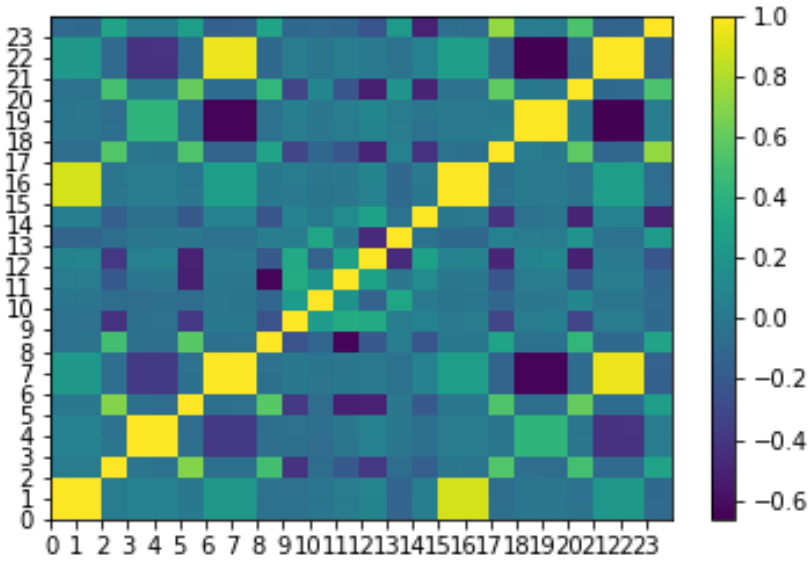

Figure 1: Correlation of the 24 features we now have

We can see some variables with a very high correlation in figure 1. There are some couple of features with a correlation near to 1 (The maximum value), like features 0 and 1, or 21 and 22. This features are the mean and median values of every axis, so we remove

all median values from our features.

Having a very similar median and mean values gives us some information: The mean of a value is an stimator strongly affected by outliers, while the median is not. If the median and the mean have similar values, this means little ammount of outliers in our data (It does not mean they are non existent).

From the correlation matrix we can also see some features with a very big correlation: the mean values of each axis between both sensors. This means the value of x for the Accelerometer and the LinearAcceleration are highly correlated, so we remove one of those (In this case, the value of the Accelerometer is dropped). The result can be seen in Figure 2.

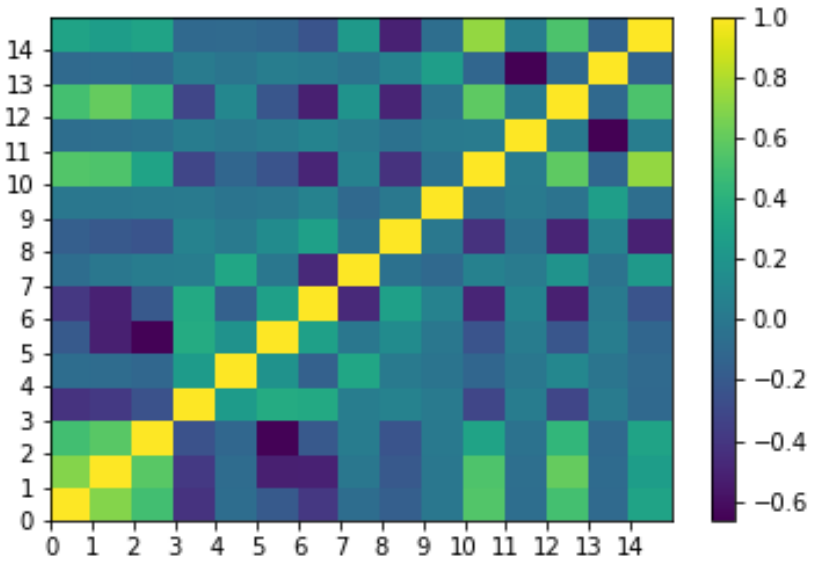

Figure 2: Correlation of the 15 features we now have after removing some

After removing these features, we are going to use hierarchical clustering to see if we have more redundancies in the values instead of the correlation. As seen in the Figure 3, there are little clusters of features, but the most similar ones are features 0, 1, 2 and 12. Those are the variances of the accelerometer. Since those features seem to have very similar values, we can use only one of them.

2 Repeating the process

Done by Juan José Corroto.

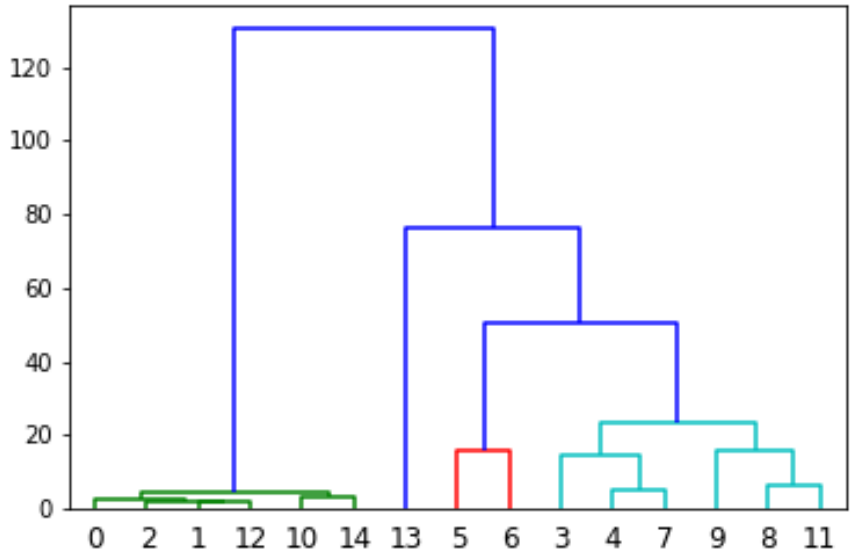

Figure 3: Features dendogram

With our new dataset finished from the previous experiment, we are going to repeat the process of our last day. First we apply PCA with an *explained_variance_ratio* of **0.76** and **0.14**. This ratio is very similar to the one we had in the previous iteration, and pretty good considering we have removed features that were highly correlated.

The data is plotted in Figure 4, and we can see some differences with respect to the plot we had in the previous day.
We run K-means several times again. We get very good values of Distortion and Silhouette again, thought this time they are even better for 4 clusters, so we select again this number of clusters and run K-means, with the plotted results shown in Figure 7.

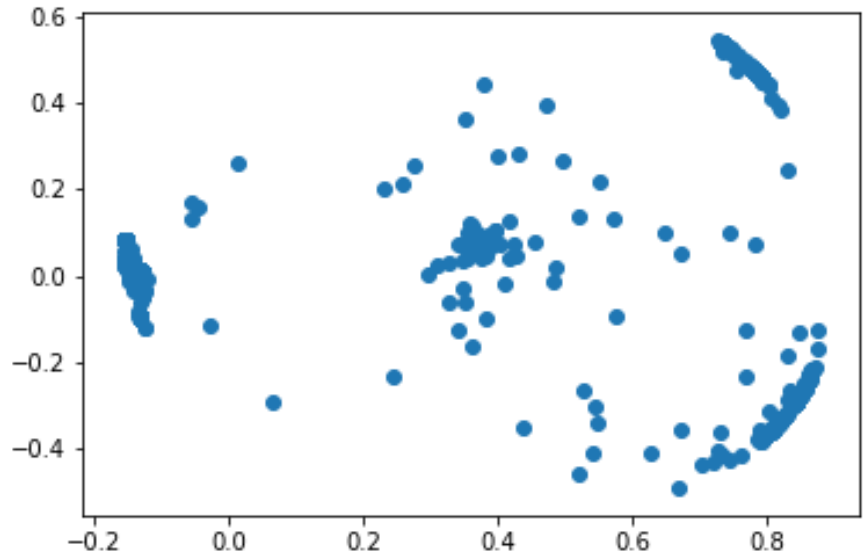

Figure 4: Plot of the data after PCA

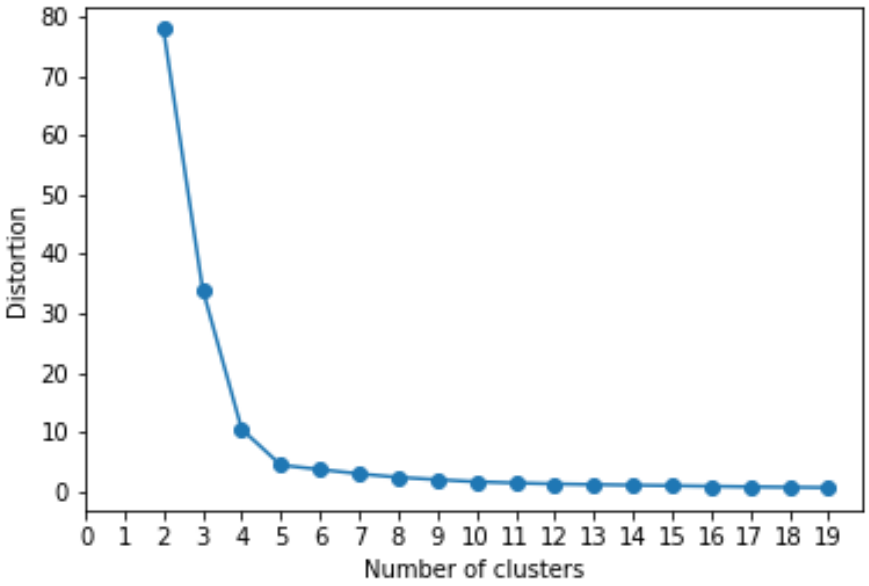

Figure 5: Distortion of K-means with different number of centroids

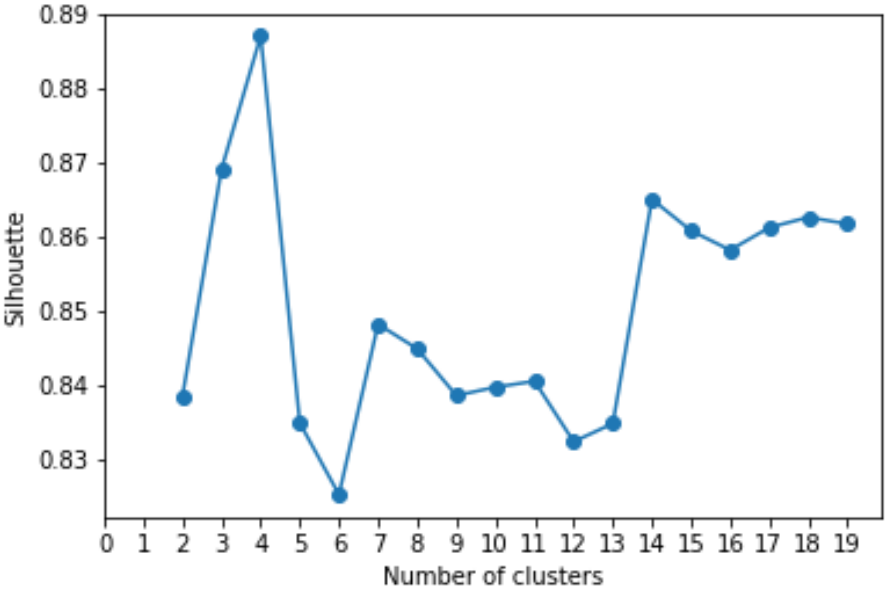

Figure 6: Silhouette of K-means with different number of centroids

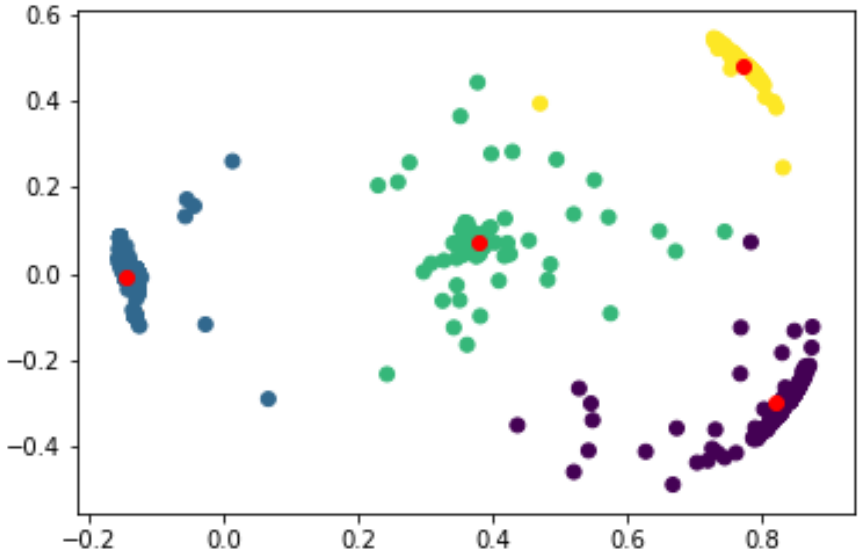

Figure 7: Plot of the data after K-means (centroids in red)

Saturday, 27 October 2018

1 Using other sensors

Done by Álvaro Guerrero del Pozo.
In this experiment, we have used what we have learnt until now, this time applied to other sensors. Again, we load the dataset, but just keep the data of the first day. Then, we drop any column that contains no information (i.e is null) and rows with null values. Then, we just keep columns related to *GyroscopeStat* and *RotationVector*
As before, we now plot correlations between features:

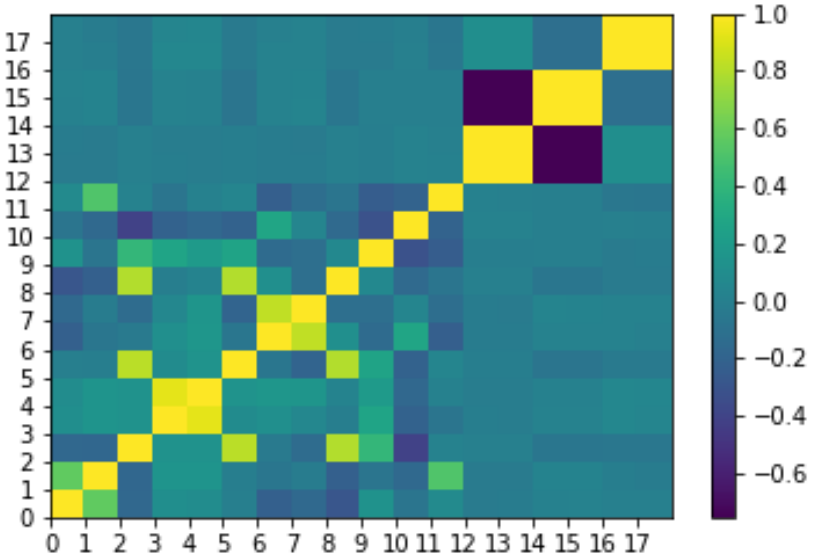

Figure 1: Correlation of the 18 features

As expected, *Mean* and *Median* values of each axis, of both sensors are highly correlated, so we can safely drop one of them, as the only add redundant informacion. We choose to remove *Median* values. But, there is an exception: X axis of the *Gyroscope*. The correlation between *Mean* and *Median* isn't high enough, so we don't remove any of them. Also, we can see that there is a surprisingly high (inverse) correlation between *Y Mean* and *X Mean* of the *Rotation Vector*.

It's not as high as other values (most of the previous ones had a correlation of near to 1), but still, with a correlation of around **-0.75**, we have decided to remove *Y Mean*. As a result, we are left with 12 features, whose correlations can be seen in the next figure:

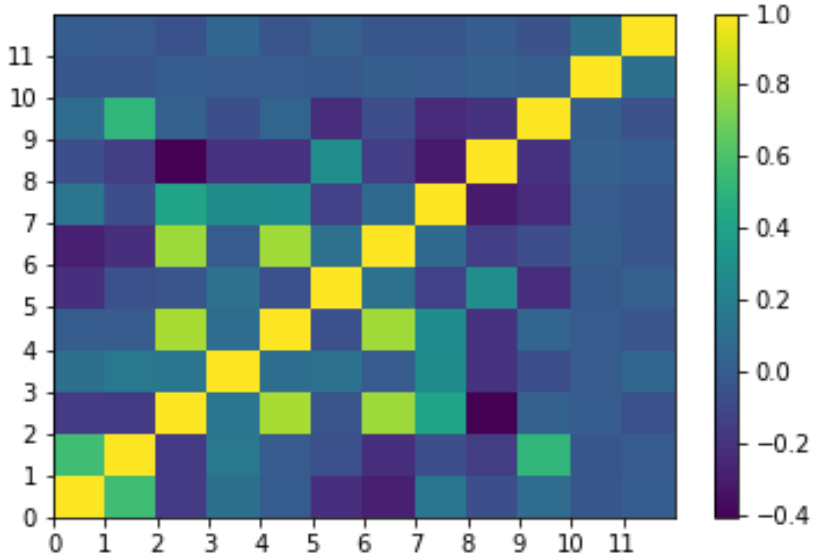

Figure 2: Correlation of the 12 features

After that, we plot the dendogram of the features (Figure **??**).

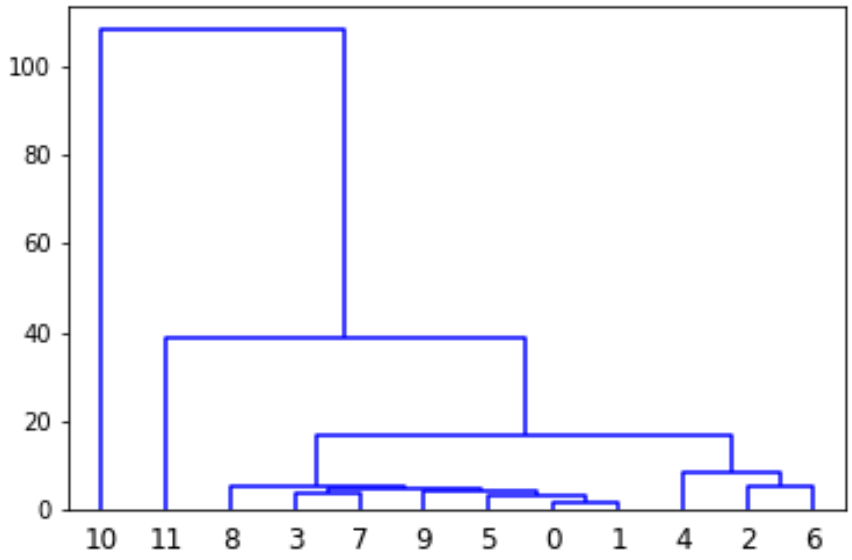

Figure 3: Dendogram

As expected, the most similar features are *Mean* and *Median* of the x axis of the gyroscope. We won't remove any of them, as we consider of interest the fact that only in this case, *Mean* and *Median* differ this much. We want this fact to influence the next steps or experiments.

Now that we have our data only with the features that are considered relevant, we proceed to visualize the data, by using *k-means*.

First, we apply PCA, and obtain an *explained_variance_ratio* of **0.809** and **0.128**. It's a high result, so we expect to obtain a good visualization. The plot of PCA is:

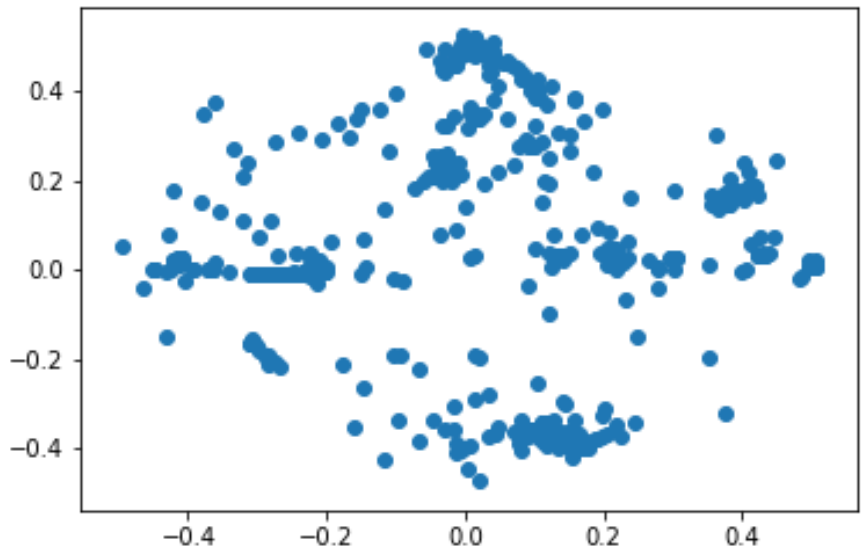

Figure 4: Plot of the data after PCA

After that, we run *k-means* several times, from 2 to 20 *centroids* and plot the Distortion **??** and Silhouette **??**.

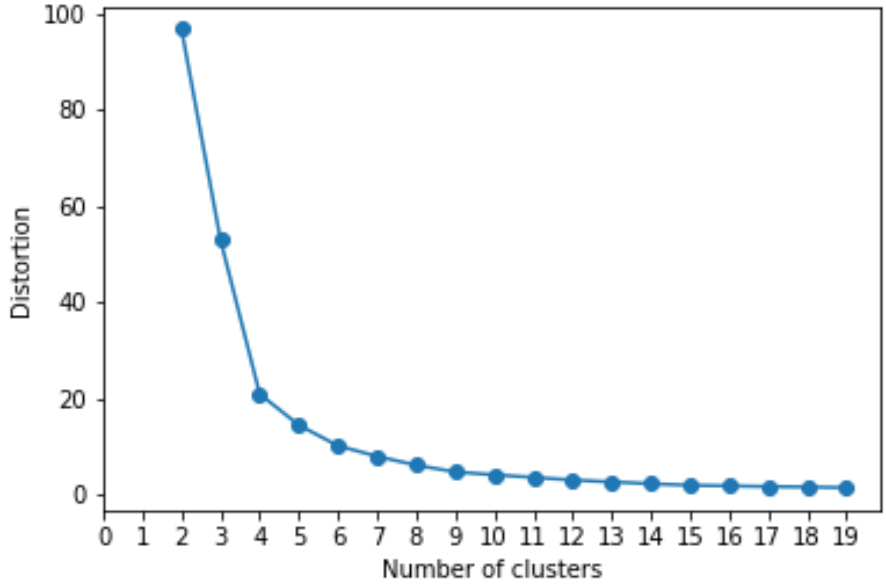

Figure 5: Distortion

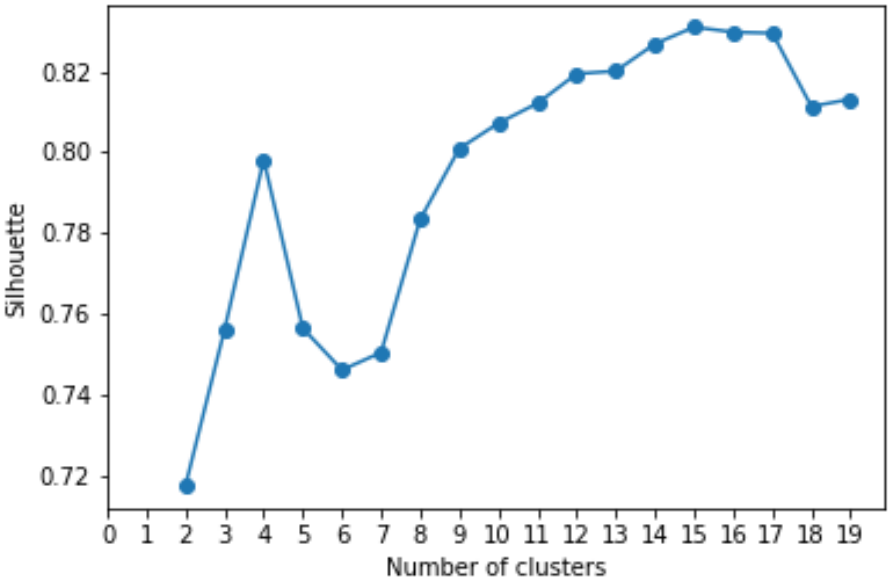

Figure 6: Silhouettes

We choose 4 as the best number of *centroids*, as there is a maximun in the *silhouette*, and the *distortion* is low. The resulting plot is shown below.

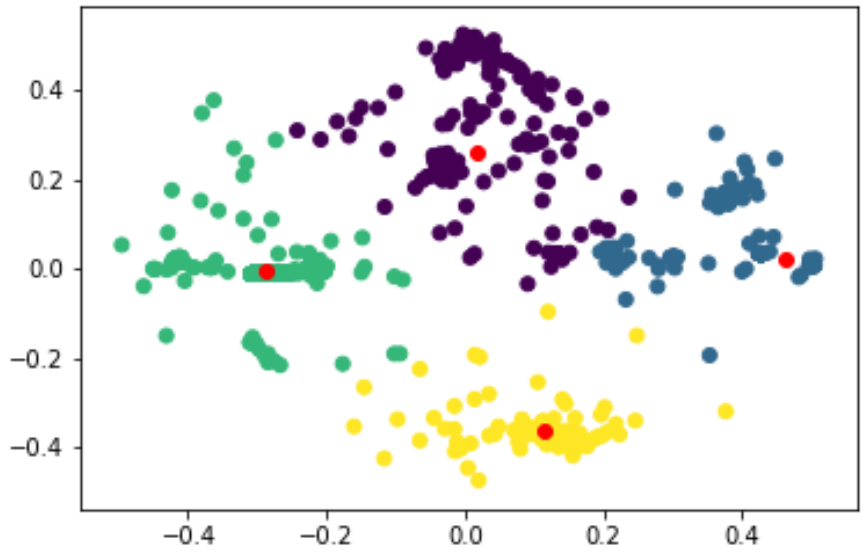

Figure 7: Plot of the data after K-means (centroids in red)

As we can see, the centroids have been placed where there are higher concentrain of points, but there is a decent amount of points that would be assigned to the neightbour cluster, should the centroids change slightly. Is a matter for another experiment to interpret this results, and decide wether they are good enough or not.

Friday, 26 March 2010

1 This shows a sample table

Groups	Treatment X	Treatment Y
1	0.2	0.8
2	0.17	0.7
3	0.24	0.75
4	0.68	0.3

Table 1: The effects of treatments X and Y on the four groups studied.

Table 1 shows that groups 1-3 reacted similarly to the two treatments but group 4 showed a reversed reaction.

Saturday, 27 March 2010

1 Bulleted list example

This is a bulleted list:

- Item 1
- Item 2
- . . . and so on

2 example

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

3 example2

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Formulae and Media Recipes

Media

Media 1

Compound	1L	0.5L
Compound 1	10g	5g
Compound 2	20g	10g

Table 1: Ingredients in Media 1.

Formulae

Formula 1 - Pythagorean theorem

$a^2 + b^2 = c^2$