Lab Book

Sherlock and Moriarty

Alfonso Bárragan Carmona
Javier Monescillo Buitrón
Roberto Plaza Romero

alfonso.barragan@alu.uclm.es
javier.monescillo@alu.uclm.es
roberto.plaza@alu.uclm.es

Start October 6, 2018

Contents

Monday, 15 - 10 - 2018

1 Repository of the project

To start working on the project of the subject "Machine Learning Techniques", we have created a repository in GitHub for collaborative work, you can see the repository here:

```
https://github.com/RoberPlaza/MachineLearningLAB
```

2 Approach to the BIG dataset

Today, we met in the library to begin our research over the sherlock and moriarty dataset, in the early, we make a lighty reading of the description and the article of sherlock and moriarty, we a less we try to open with excel (we know that we can't open the file by that way, but we wanna try).
We are thinking other kind of ways to explore the dataset, the most posible options are to take a proportion of the elements from the dataset (1 element of every 10, in a regular method or with a random method, for example) or making a bash script because our teacher says that bash can do wonders when dealing with files and the management associated with them.
After the first hours of contact, we have opted to cut the database via python programming. For reasons of efficiency and speed we will do it staying with the tenth line of every ten, and once we have a csv of a more manageable proportions, we will proceed to a more internal analysis to see how to manage the large database really. In order to make it as fast as possible, we are even trying to make programming in threads, for the future could be very profitable.
We did it and with a beer we're done for the day.

Thursday, 23 - 10 - 2018

1 Begining the milestone one

We begin with the treatment of the T2 dataset of sherlock and moriarty, our beginning is partly directed by the indications given by francisco, we are beginning with doing a sampling of a single day instead of putting ourselves first with the complete dataset.
Once this has been done, we have decided to take rows of three in three (one minute), and make the average of the values poured by the sensors. We do this because a lot of the sensors are dependent on the user's situation and we assume that it won't vary substantially at twenty-second intervals.
In order to undertake a first approach in an efficient way, we are going to take the sampling of a specific day. We will deal with the choice of that day, randomly or by some specific criterion. But that is a problem for the machine learning laboratory group of the future.

Saturday, 27 - 10 - 2018

1 Some problems with missing values

Today we have gathered as a team to try to treat the lost values in T2 csv, and the result has been a failure, we have had problems with the fields that had "NULL" values. This is because when we try to do the filtering we replace these fields with basically "", which is nothing. So we have done several tests with pd.dropna(), and we have not been able to remove them.

2 Success in normalizing csv

In the end we were able to eliminate the lost values, thanks to the following code:

```
def normalize_filtered_data(path):
    file = pd.read_csv(path, low_memory=False)

    exclude = ['UserID', 'UUID', 'Version', 'TimeStemp', "
       RotationVector_cosThetaOver2_MEAN", "
       RotationVector_cosThetaOver2_MEDIAN", "
       RotationVector_cosThetaOver2_MIDDLE_SAMPLE"]
    df_ex = file.loc[:, file.columns.difference(exclude)]
    df_ex = df_ex.replace(" NULL", np.NaN)

    df_ex = df_ex.dropna()

    min_max_scaler = preprocessing.MinMaxScaler()
    df_norm = min_max_scaler.fit_transform(df_ex)

    return df_norm
```

We couldn't remove the fields to "NULL" because they had a space in front of them, so using dropna(), the np.NaN values are replaced and deleted.

As you can see we remove the columns that are not necessary for the PCA, and then we use the normalized min max scaler. The next job we'll have to do is to think what kind of clustering algorithm we'll use. We will also request a meeting with the teacher to see if all the work is correct.

And now it's time to drop ourselfs to sleep, and continuing tomorrow.

Thursday, 1 - 11 - 2018

1 Deciding which features we should delete

After a long period of work, thinking about what we had to do and meeting with the subject teacher in his office, we met to do the following activities:

- Update the organization of the project
- Treat features in preprocessing
- Decide which features to remove, according to the description of the csv file

At this moment, we are deliberating which are the most appropriate characteristics depending on the correlation matrix, correlation is something we have also done in this meeting.

We have updated the **.py** files and this is the following organization of the project:

- MachineLearningLAB
  - docs
  - milestone1
    - plots
    - data
    - 1_preprocessing.py
    - 2_normalize.py
    - 3_clustering.py
  - milestone2

In preprocessing.py we have the preprocess of both rows and columns, in normalize.py we have normalized the csv file, deleting categorical variables and applying minMaxScaler and finally in clustering.py we will try to apply several clustering algorithms.
After several hours of deliberation, we have decided to discard the following columns:

- Fast Fourier Transform.
- All columns of the axes of the y.
- All summary columns for medians, variance, middle sample.
- Columns that have to do with the rotation vector and the orientation probe.

(Reasons why here)

After these modifications, we have made a dimensional reduction by principal component analysis, obtaining two principal components of [0.52] and [0.27] respectively.

Once reached this step, we reached the point of choosing which could be the most suitable clustering for our data, at the moment in view of the PCA, we have tested first with the k-means and more or less does it correctly, but we are open to a dbscan because we have a large density of data and may group better than the k-means.

But we will arrange that tomorrow, for today we have already had enough Principal Components that represent 100 percent of our hard work.

Friday , 2 - 11 - 2018

1 At the end of milestone 1

It should be noted that in recent days of meetings, specifically the previous and today, have been online.
At the beginning of the afternoon we have created the powerpoint related to the video and we have decided which parts are going to occupy each one and what is going to explain.
In terms of code, we moved the replacement of null values to preprocessing.py, as we were doing it in normalize.py.

Then we have saved the images of the similarity matrix and the correlation tests using plt.savefig().
And now comes the most important thing: We have added the DBSCAN clustering
After many deliberations we have decided to discard the accelerometer readings because they were redundant with the linear acceleration readings.
We are also interpreting the groups that pour the different clustering algorithms, at the moment we have KMeans, KMeans++ and DBSCAN.

For the next meeting that will probably be tomorrow we will finish interpertar the results, choose which clustering we are going to stay.

We can portray all our effort and hard work, with the next data scientist spell: Skidaddle skadoodle your data is now knowledge

Saturday , 3 - 11 - 2018

1 Fixing some problems and starting the video