Paris Descartes University

Study and research work.

Non-negative matrix factorization and image classification

Students:
Ilyes ZEMALACHE
Hamoud GUICHENITI

Supervisors:
Mohamed NADIF
Michael FEBRISSY

16th May 2019

Acknowledgments:
I would like to thank Mr Mohamed NADIF for having trusted us to work on this subject, and also to have made available to us all the good working conditions. I also thank Mr. Michael FEBRISSY who followed us through our project, and gave us advices that was useful during all stages of the realization of this latter.

Contents

Abstract

This study and research work consists in proposing a classification system for images from a data set.
First of all, we will create this data set ourselves by downloading 2000 images of 5 categories of animals, while respecting a distribution recommended by our supervisor, for the realization of a certain diversity in our photos.
Second, we'll do a small visualization with principals components analysis , which will give us an idea of how our data is distributed.
Then we will apply k-means and nmf with different values of k or rank, this latter will be determined according to several techniques that will be detailed below.
The purpose of our project is finally to make a comparative study between the two algorithms by applying two functions NMI and ARI, which will lead us to a general conclusion.

1 Introduction

1.1 General Introduction

Data partitioning (or data clustering) is one of the methods of data analysis. It aims at dividing a set of data into different homogeneous "packets", Where the data of each subset share common characteristics, which most often correspond to proximity criteria (computer similarity) that is defined introducing measures and classes of distance between objects. To obtain a good partitioning, it is necessary to: minimize intra-class inertia to obtain clusters as homogeneous as possible; maximize inter-class inertia in order to obtain well-differentiated subsets.
Data partitioning is a method of unsupervised classification (different from supervised classification where learning data is already tagged), so sometimes referred to as such.
Applications: There are generally three types:

- The segmentation of a database; it can be used to discretize a database.

  Segmentation can also be used to condense or compress the data of a spatial database (that means, to reduce the size of the data packets to be processed, in the given data set); for example, in an aerial or satellite image a SIG can treat forests, fields, meadows, roads, wetlands, etc. differently. here considered as homogeneous subspaces. A finer treatment can then be applied to subsets of these classes (eg deciduous, coniferous, artificial, natural forest, etc.). OLAP is a method that facilitates the indexing of such databases;

- The classification (in subgroups, subpopulations within the database), for example of a customer database, for the management of the customer relationship;

- The extraction of knowledge, which is generally done without a priori objective (serendipity factor, useful for the generation of hypothesis or predictive modeling), to bring out subsets and sub-concepts that may be impossible to distinguish naturally.

1.1.1 Machine Learning

Machine learning or statistical learning is a field of study of artificial intelligence that is based on statistical approaches to give computers the ability to "learn" from data, that is to say to improve their performance to solve tasks without being explicitly programmed for each. More broadly, this concerns the design, analysis, development and implementation of such methods. Machine learning usually has two phases. The first is to estimate a model from data, called observations, that are available and in finite numbers, during the design phase of the system. Model estimation involves solving a practical task, such as translating a speech, estimating a probability density, recognizing the presence of a cat in a photograph, or participating in the driving of an autonomous vehicle. This so-called "learning" or "training" phase is generally performed prior to the practical use of the model. The second phase corresponds to the setting in production: the model being determined, new data can then be submitted in order to obtain the result corresponding to the desired task. In practice, some systems can continue their learning once in production, provided they have a way to get a return on the quality of the results produced. According to the information available during the learning phase, learning is qualified in different ways. If the data is tagged (that is, the response to the task is known for that data), it is a supervised learning. We are talking about classification or classification if the labels are discrete, or regression if they are continuous. If the model is learned incrementally based on a reward received by the program for each of the actions taken, it is called reinforcement learning. In the most general, unlabeled case, we try to determine the underlying structure of the data (which may be a probability density) and this is unsupervised learning. Machine learning can be applied to different types of data, such as graphs, trees, curves, or simply feature vectors, which can be continuous or discrete.[6]

1.1.2 Classification of images

As an indication, almost 90% of the information received by the man is visual. The production of quality images, as well as their automatic (and if possible automatic) digital processing is therefore of considerable importance. Most scientific devices provide images (microscopes, telescopes, radiographs, magneto-nuclear resonance, ...) and many areas of applications use the image as a source of information and or visualization.
The image is actually a collection of information that, at first, was presented on a photographic medium that allowed the delayed processing of the fleeting phenomenon, a

fine analysis of recorded phenomena and of course archiving and illustration . Image processing is born from the idea and the need to replace the human observer by the machine. The image or signals from sensors were then digitized for processing by the computer. In a second step, the image has been coded and stored on different media. A very common task is the classification of images. In order to be able to use the images for mapping or for further analysis, it is often important to translate the frequency information contained in the images into thematic information on land cover or vegetation cover. You usually have two choices: supervised and unsupervised classification.[7]

1.1.3 Supervised classification

Supervised learning is a machine learning task consisting of learning a prediction function from annotated examples, as opposed to unsupervised learning. Regression problems are distinguished from classification problems1. Thus, it is considered that the problems of predicting a quantitative variable are regression problems whereas the problems of predicting a qualitative variable are classification problems. The annotated examples constitute a learning base, and the learned prediction function can also be called ”hypothesis” or ”model”. It is assumed that this learning base is representative of a larger sample population and the purpose of supervised learning methods is to generalize, that is, to learn a function that makes correct predictions about data not present in the training set.[8]

1.1.4 Unsupervised classification

This is what interest us throughout our project, unsupervised learning is mainly used in clustering,a method for grouping together a set of heterogeneous elements in the form of subgroups homogeneous or linked by common characteristics. . The machine then makes itself the reconciliations according to these characteristics which it is able to locate without requiring external intervention. From this ability to perform clustering also arises the possibility of developing a recommendation system (the system can for example recommend a book or a movie to a user based on the tastes of users sharing common characteristics) as well as the possibility of developing an anomaly detection system.[8]

1.2 Tools Used

We will use three main tools :

* Rstudio

* Jupyter NoteBook(Python)

* NMI/ARI

1.2.1 Rstudio

RStudio is a free, cross-platform development environment for R, a programming language used for data processing and statistical analysis. It is available under the free license AGPLv3, or under a commercial license, subject to an annual subscription. RStudio is available in two versions: RStudio Desktop, for a local execution of the software like any other application, and RStudio Server which, launched on a Linux server, allows access to RStudio by a web browser. RStudio Desktop distributions are available for Microsoft Windows, OS X, and GNU / Linux. RStudio was written in C ++, and its graphical interface uses the Qt programming interface. Since version 1.0 released in November 2016, RStudio integrates the ability to write notebooks interactively combining R code, text formatted markdown and calls to Python or Bash code.[3]

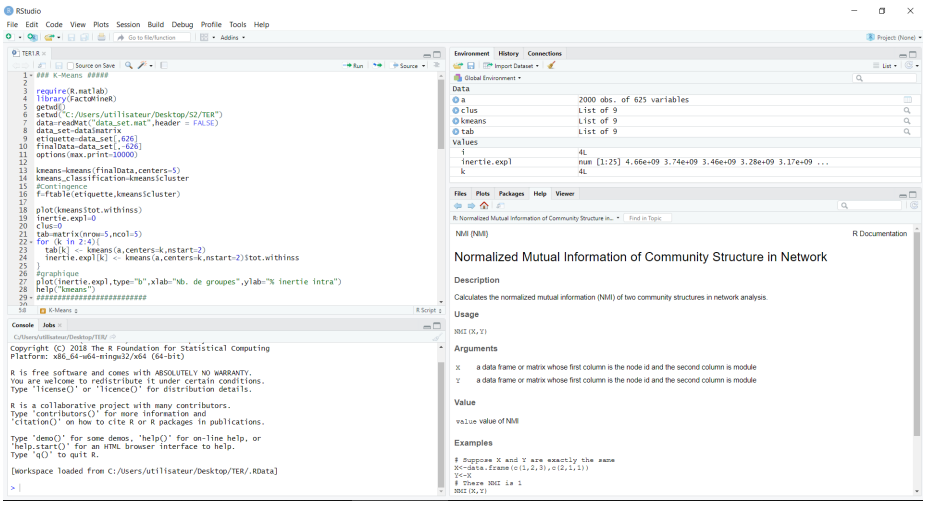

Figure 1: Rstudio

1.2.2 Jupyter-Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.[4]

1.2.3 NMI/ARI

Two functions to compute the Normalized mutual information and the adjusted rand index between two classifications.[1]

2 Data Construction

2.1 Resume

In this part, we built the data set from 0, this means that we used Flickr to download 2000 images, then we converted these images in gray level, and with Numpy we transformed each image in one vector, finally we applied the transformation of all the images into a matlab file of $2000 lines \times 625 columns$.

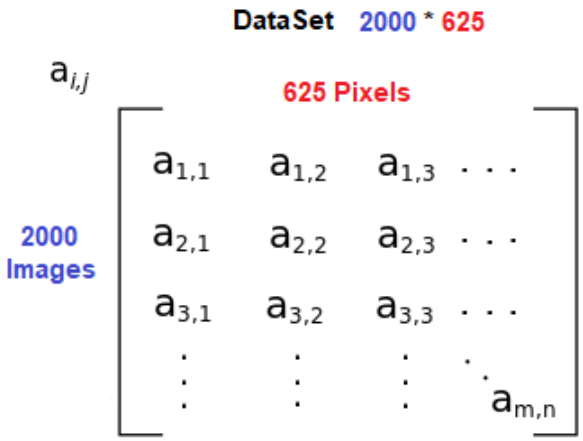

Figure 2: Illustration of the data

2.2 Flickr

Flickr, is a free photo and video sharing website, with some paid features. In addition to being a popular website for users to share their personal photos, it is also often used by professional photographers. In August 2011, the site passed the six billion mark of

photos hosted. In February 2017, the site hosts approximately thirteen billion photos for one hundred and twenty-two million members and two million groups

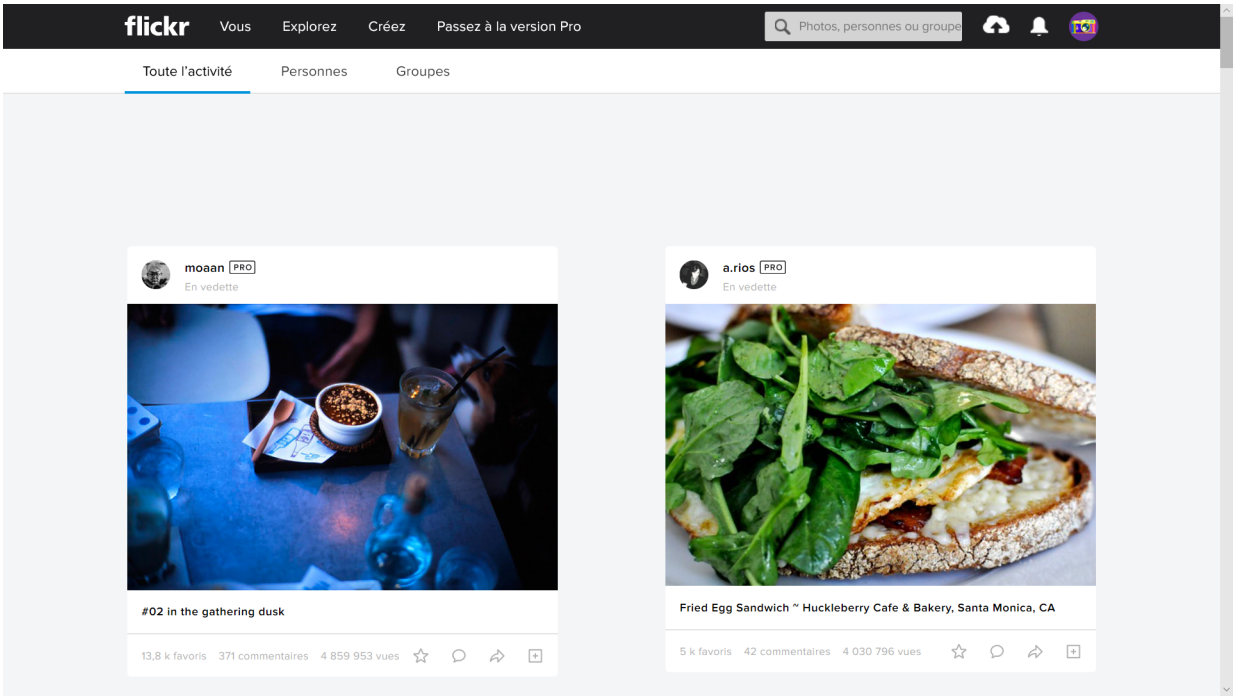

Figure 3: Flickr

2.3 API Flickr

To be able to use ”Flickr” in our piece of python code, we needed to have an account on the flickr website, and then we asked for an API, which allows us to download the images.

```
Entrée [1]: import flickrapi
            import urllib.request
            import os, sys
            from PIL import Image
            # Flickr api accessa key
            flickr=flickrapi.FlickrAPI('97c7f44a6015743ff2cbf9ea8fc6eb23', 'e7d6b6f428d87320', cache=True)
```

Figure 4: API

2.4 Preparation of the list of photos

In this part, we make a little code that will make us a total of 2000 photos of animals of 5 classes, using different keywords that will allow us to have a good diversity on the k classes.

As can be seen, the animals are:

- Cats, snakes, zebras, horses and Fishes.

Here is how the distribution is done:

- 600 Cats, including 300 photos containing several cats, and 300 others containing a cat.
- 300 Snakes
- 250 Zebras
- 600 Horses, 50% of them which we see only the head and 50% remaining represent a full horse.
- 250 Fishes, which 25% are shark, 25% pineapple fish, 25% fish and the rest is railed fish

```
nbr_cat=600
nbr_snake=300
nbr_zebra=250
nbr_horse=600
nbr_fish=250
data=[
            ['chat',int(0.5*nbr_cat)],
            ['multiple cat',int(0.5*nbr_cat)],
            ['zebra',int(nbr_zebra)],
            ['snake',int(nbr_snake)],
            ['horse',int(0.5*nbr_horse)],
            ['head horse',int(0.5*nbr_horse)],
            ['Cleidopus gloriamaris',int(0.25*nbr_fish)],
            ['shark',int(nbr_fish*0.25)],
            ['fish',int(nbr_fish*0.25)],
            ['stripped fish',int(nbr_fish*0.25)]
          ]
```

2.5 Uploading photos

Here we start downloading our 2000 photos, and store them in different folders depending on the keyword used.

```
for i in range(0,len(data)):
        photos = flickr.walk(text=data[i][0],
                     tag_mode='All',
                     extras='url_l',      # 75*75 s= small
                     per_page=data[i][1],           # number of pic per page
                            )
        urls = []
        for j, photo in enumerate(photos):
            url = photo.get('url_l')
            if(url!=None):
                urls.append(url)
            # get number of pic
            if j > data[i][1]:
                break

        # Download image from the url and save it to '00001.jpg'
        for k in range(0,len(urls)):
            dossier='/home/ilyessou/Bureau/TER/'+data[i][0]+'/'
            chemin=dossier+str(k)+'.jpg'
            if not os.path.exists(dossier):
                os.mkdir(dossier)
            urllib.request.urlretrieve(urls[k], chemin)

            #Resize the image and overwrite it
            image = Image.open(chemin)
            image = image.resize((200, 200), Image.ANTIALIAS)
            image.save(chemin)
```

Figure 5: Downloading

2.6 Conversion to gray level

To make the task easier , we made our study on gray-scale images, which limits the number of colors to only 256 colors,so the study will be done without colors(RGB) .

```
Entrée [65]:
folders=glob.glob('/home/ilyessou/Bureau/CopieDeDonnée/*') # enregistrer la liste de tous les dossiers d'annimaux
for i in range (0,len(folders)):           # parcourir tous dossier
    folder=glob.glob(folders[i])
    pictures=glob.glob(folder[0]+'/*')   # tous les photos du dossier i
    for j in range(0,len(pictures)):

        img=Image.open(pictures[j]).convert('L')        #conversion en niveau de gris
        dossier='/home/ilyessou/Bureau/ALL/'+folder[0].split("/")[5]+'/'    #split pour prendre le nom du fichier
        path=dossier+str(j)+'.jpg'
        if not os.path.exists(dossier):
                os.mkdir(dossier)
                img.save(path)
```

Figure 6: Conversion

2.7 Transformation into data set

Using the Numpy package,we transformed each image form a matrix (25*25) to one vector of 625 and we concatenate them vertically , So we will get 2000 lines and 625 columns, which will represent the pixels. Finally it's a data set in a matlab file of 2000×625.

```
Entrée [1]:
from PIL import Image
import csv
import os, sys
import numpy
import glob
import scipy.io as sio

#ImageFile.LOAD_TRUNCATED_IMAGES = True
```

```
Entrée [2]:
folders=glob.glob('/home/ilyessou/Bureau/GreyPics/*')
liste=[]
for i in range (0,len(folders)):
    folder=glob.glob(folders[i])
    pictures=glob.glob(folder[0]+'/*')   # tous les photos du dossier i
    for j in range(0,len(pictures)):
        img=Image.open(pictures[j])
        img=img.resize((25, 25), Image.ANTIALIAS)
        imgarr=numpy.array(img).reshape(625)
        #imgarr = numpy.insert(imgarr,i)
        imgarr = numpy.append(imgarr, i+1) # Pour rajouter une dernierre colonne pour supervisé
        liste.append(imgarr)
```

```
Entrée [3]:
matrix=numpy.array(liste)
```

```
Entrée [4]:
sio.savemat('data_set.mat',{'matrix':matrix});    #### .mat
```

```
Entrée [26]:
matrix.shape

Out[26]:
(2000, 626)
```

Figure 7: Data set

3 Les Algorithmes de partitionnement

3.1 Clusturing

Clustering is a statistical analysis method used to organize raw data into homogeneous silos. Within each cluster, the data is grouped according to a common characteristic. The scheduling tool is an algorithm that measures the proximity between each element based on defined criteria. To establish equilibrium, it minimizes the inertia within the classes and maximizes that between the subgroups in order to differentiate them well. The goal may be to prioritize or distribute the data. In French, the term ”grouping” or the expression ”partitioning of data” is commonly used. Clustering is mainly used to segment or classify a database (for example, sorting customer data such as age, occupation, place of residence, etc., to optimize customer relationship management) or to extract knowledge in order to try to find new information. subsets of data that are difficult to identify with the naked eye. In natural referencing, we use clustering to structure the key words of a site and create the basis of its semantic fabric from the search intentions collected on the search engine results pages. Spatial imagery compresses its data by organizing as clusters the different elements present on each image, such as forests, cities or agricultural areas for example. This makes it possible to reduce the size of the data packets that are, if not too heavy. To be applied, clustering relies on more or less complex algorithms, such as the k-means or k-medoids algorithms, or the algorithms of maximization of the expectation.[10]

3.2 K-Means

This is one of the most common clustering algorithms. It allows to analyze a dataset characterized by a set of descriptors, in order to group the ”similar” data in groups (or clusters). The similarity between two data can be inferred from the ”distance” separating their descriptors; So two similar data are two data whose descriptors are very close. This definition makes it possible to formulate the data partitioning problem as the search for K ”prototype data”, around which the other data can be grouped. These prototype data are called centroids; in practice, the algorithm associates each element with its closest centroid, in order to create clusters. On the other hand, the averages of the descriptors of the cluster data, define the position of their centroid in the descriptor space: this is at the origin of the name of this algorithm. After initializing its centroids by taking random data from the dataset, K-means alternates these two steps several times to optimize the centroids and their groups:

- Group each object around the nearest centroid.
- Replace each centroid according to the average of the descriptors in its group.

After a few iterations, the algorithm finds a stable division of the dataset: we say that the algorithm has converged. Like any algorithm, K-means has advantages and disadvantages: it is simple, fast and easy to understand; however, it does not allow finding groups with complex shapes.[9]

3.3 NMF

Non-negative matrix factorization (NMF or NNMF), also non-negative matrix approximation is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. Also, in applications such as processing of audio spectrograms or muscular activity, non-negativity is inherent to the data being considered. Since the problem is not exactly solvable in general, it is commonly approximated numerically.
NMF finds applications in such fields as astronomy,computer vision, document clustering, chemometrics, audio signal processing, recommender systems, and bioinformatics.[5]

$$X \approx WH$$

3.3.1 NMF Algorithms

NMF algorithms generally solve problem iteratively by building a sequence of matrices (W,H) that reduces at each step the value of the objective function F Beside some variations in the specification of F they also differ in the optimization techniques that are used to compute the updates for(W,H).

$$\min_{W,H \geq 0} \underbrace{[D(X, WH) + R(W, H)]}_{=F(W,H)}$$

- D is a loss function that measures the quality of the approximation.
- R is an optional regularization function, defined to enforce the wanted properties on matrices W and H .

4 Résultats PCA

4.1 PCA

Principal component analysis (PCA),(ACP In french) is a method of the family of data analysis and more generally of multivariate statistics, which consists of transforming linked variables (called ”correlated” into statistics) into new variables that are not correlated. others. These new variables are called ”principal components”, or main axes. It allows the practitioner to reduce the number of variables and make the information less redundant.

It is an approach that is both geometric2 (the variables being represented in a new space, according to maximum inertia directions) and statistical (research on independent axes explaining at best the variability - the variance - of data). When one wants to compress a set of N random variables, the n first axes of the principal component analysis are a better choice, from the point of view of the inertia or the variance.[2]

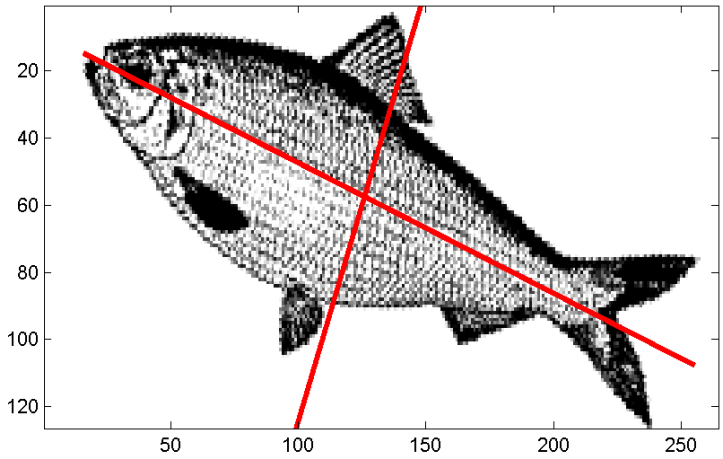

Figure 8: Illustration of the PCA

4.2 Interpretation

We applied PCA to our dataset, to get an idea of the distribution of our points, here is what we get as a result, see the figure below:

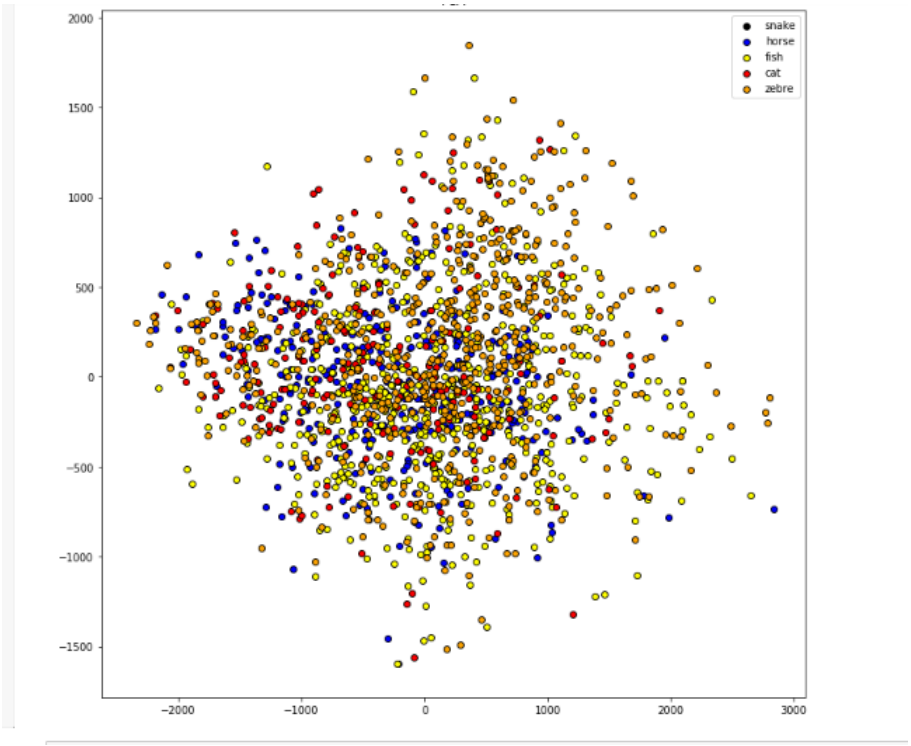

Figure 9: PCA

We notice an overlap in our classes, which really explains the results we obtained on the rest of our small research, such as the difficulty of determining the number of k and the rank for NMF.

5 Results Kmeans

5.1 Determining the number of cluster k

One of the complications, is to define the best number of cluster, for this several criteria have been proposed by the researchers, we will use two: the method of the Elbow and the coefficient of Silhouette.

- The Elbow Method: The idea behind this method is to run K-Means on all the data for a cluster number of k-values (for example k = 1 to 9) and calculate for each row k the total sum of intra inertia within the cluster. So one can choose a cluster number k so that the addition of another cluster does not give a better modeling of the data. This method is also called the ”elbow” method because if we display the information on a graph, we can choose the rank k at a level where the curve forms an ”elbow”.

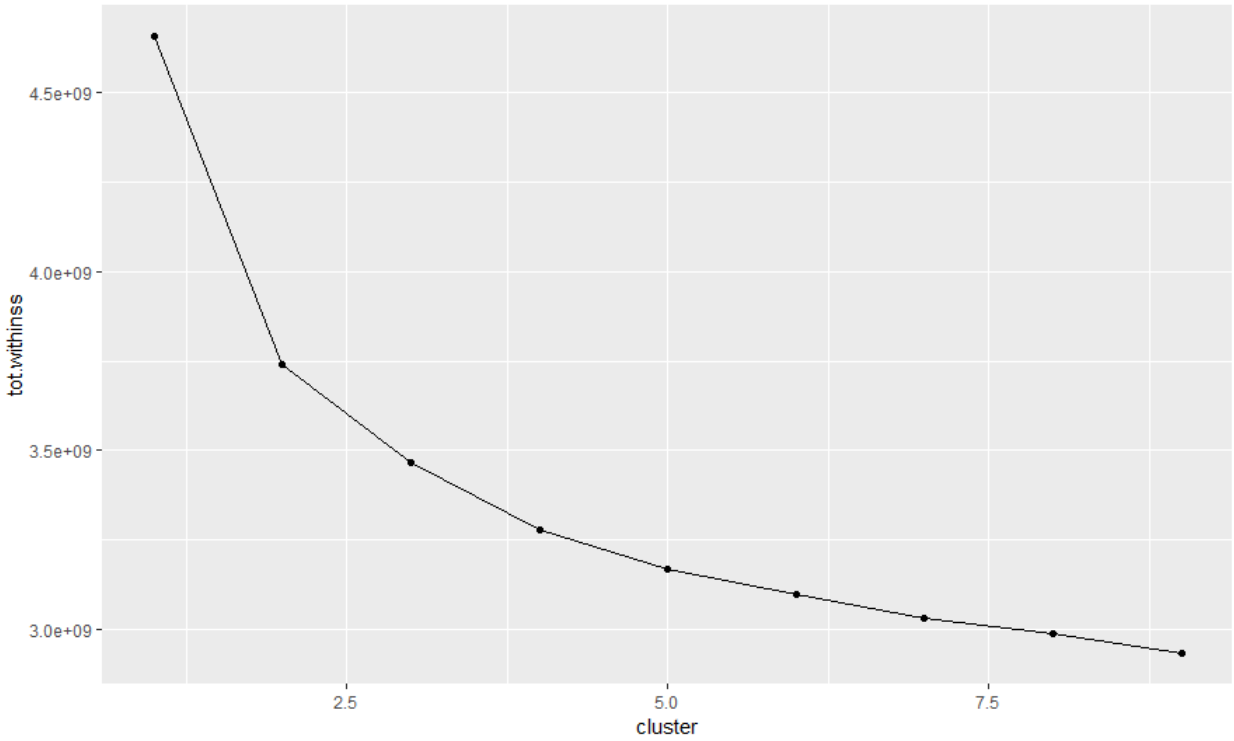

Figure 10: Kmeans avec critère du coude

We note according to the criterion of the elbow, that the ideal number of k is equal to 2, and even we can tolerate k = 3,4.
The silhouette criterion is applied, see Figure 8.

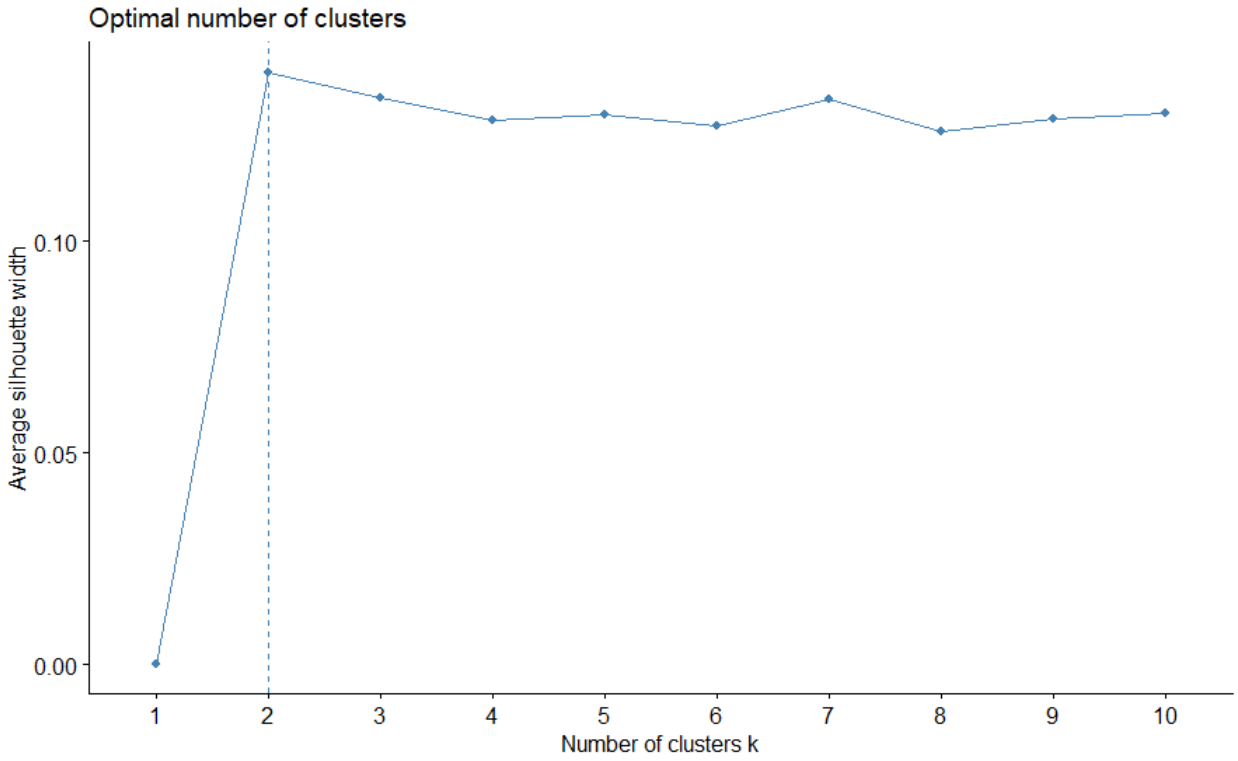

Figure 11: Kmeans with silhouette criteria

On the other hand we note, that the silhouette proposes us the rows 2, 5, 7, However these results are to be qualified because the highest value obtained for the silhouette is of 0.16.

5.2 Application of NMI/ARI

By calculating the NMI and the ARI here is what we obtained as a result: see figure 9

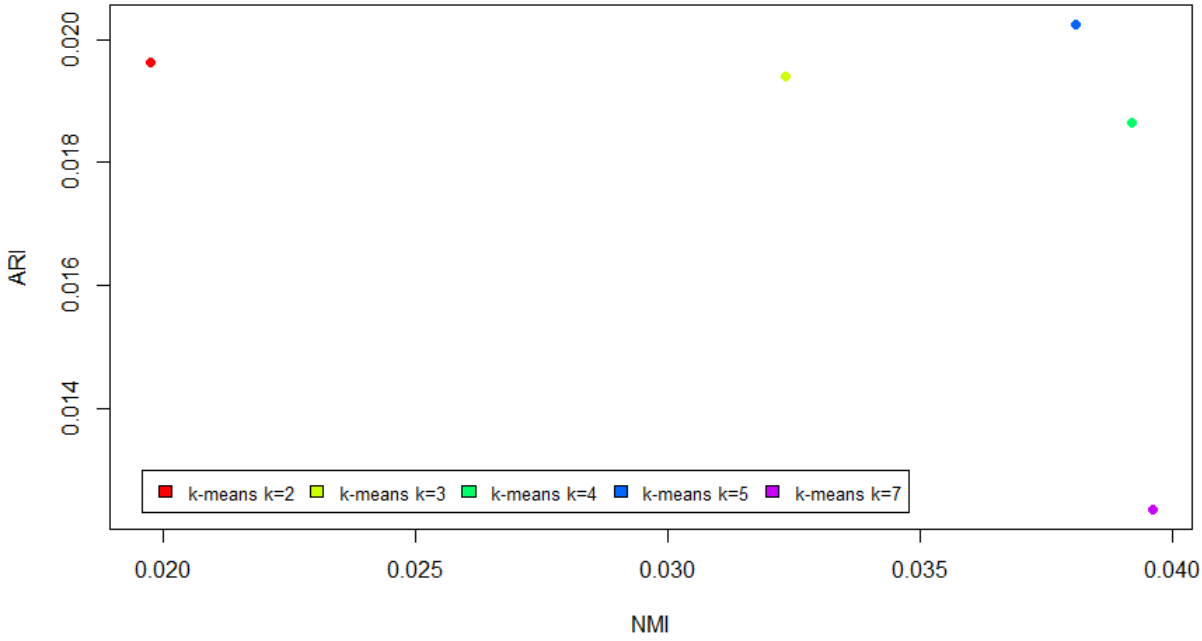

Figure 12: NMI/ARI in kmeans

The results in Figure 9 show us that the methods and ranks are more or less equal. The results are around 0.02 and 0.04 for the NMI and between 0.01 and 0.022 for the ARI. These results are weak and go in the direction of the results obtained with the silhouette.

6 Résultats NMF

6.1 Determination du rang k de la NMF

The factorization factor is a very important parameter of the NMF. As for the K-means cluster number. It can be difficult to determine the right number of clusters. In order to determine the number of clusters, one must calculate the qualities and choose the best value of these. there is several approaches have been proposed, in our case we will use 3 approaches:

- The first approach is to choose the first value of r where the cophenetic coefficient starts to decrease.

- The second approach is to choose the first value where the RSS curve presents an inflection point.
- The third approach is to choose k when the silhouette coefficientis the highest.

The first step was to determine the best rank for the NMF. For this, the NMF was calculated on a range of 2 - 10 intervals.

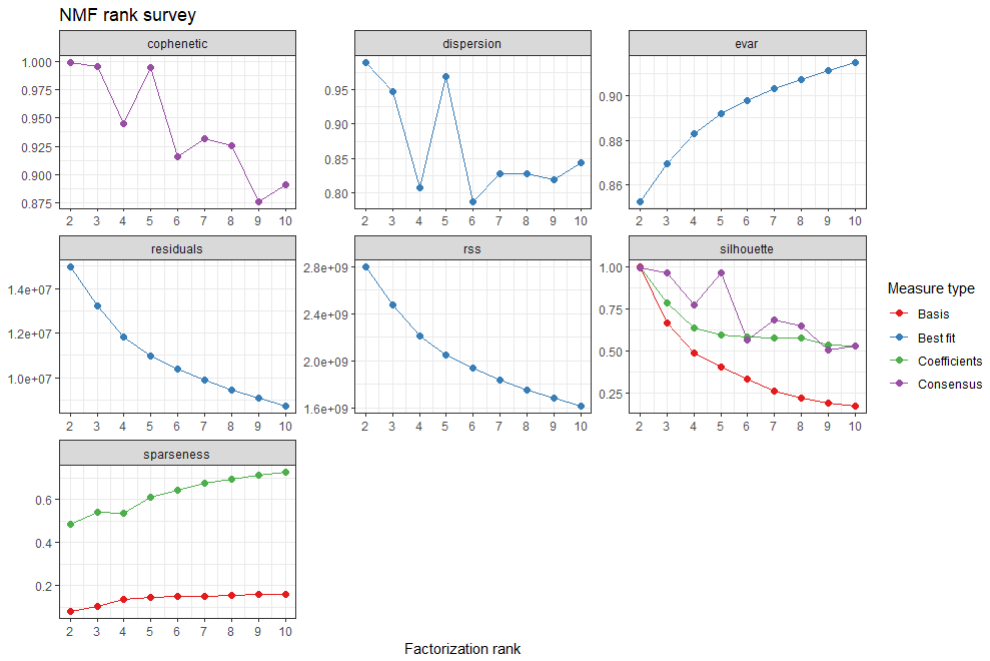

Figure 13: Graphics for NMF

we can see that the coefficient of cophenetic decreases in the ranks 3 and 5, the RSS makes an elbow in the ranks 3, 4, 5 rank and. In addition the silhouette has a maximum value in the ranks 3 and 5 so as the 3 and the 5 are present in the three approaches we decided to keep the 5 as an adequate rank .

6.2 Transformation from nmf result

After the application of nmf we tried to retrosforme the best 10 images by multiplying the two matrix (W and H) and we got this result :

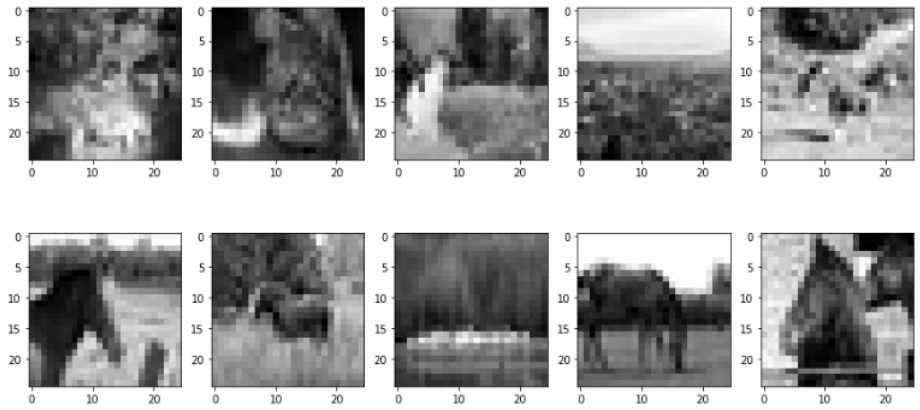

Figure 14: Images before the application of nmf

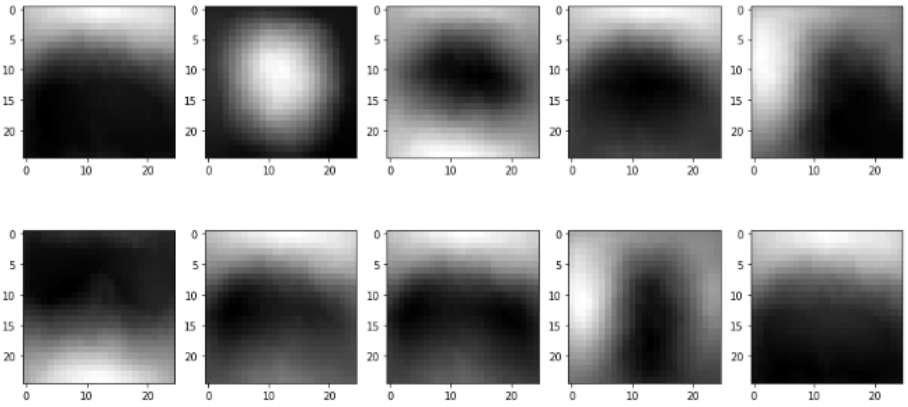

Figure 15: Images after the application of nmf

6.3 Application of NMI/ARI

After determining the best rank for the NMF and cluster number optimal for K-Means, we wanted to compare the partitions obtained between NMF and clustering. For this we calculated the ARI and the NMI :

6.4 Comparison of results between K Means and NMF

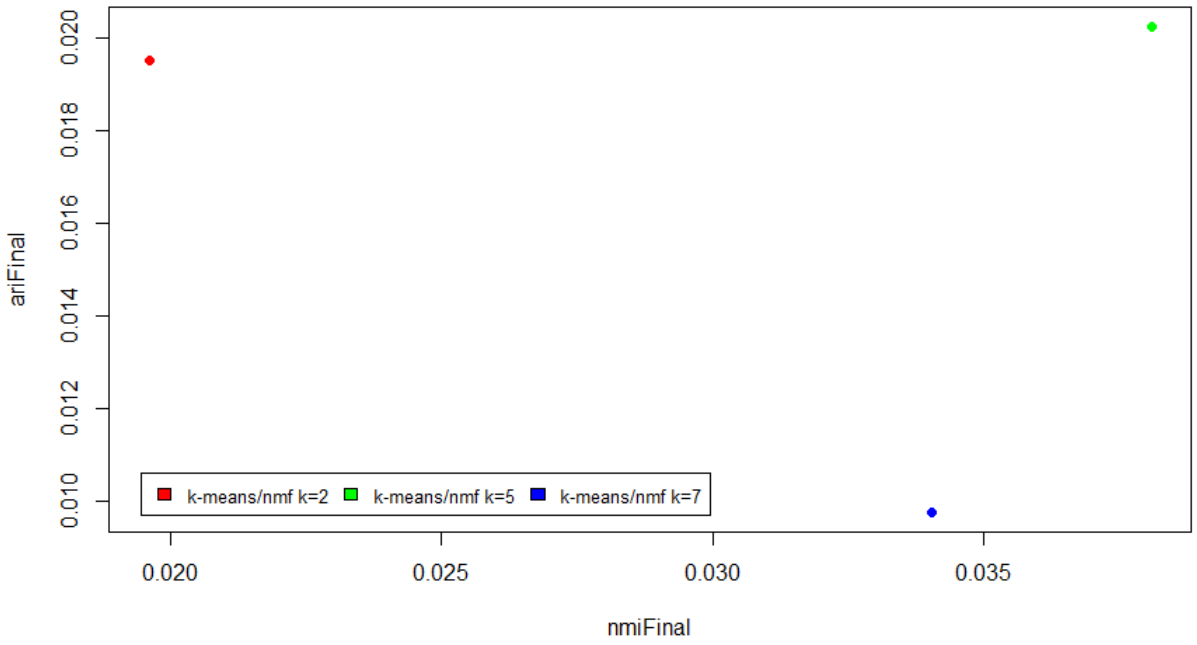

Figure 16: NMI/ARI kmeans/nmf

6.5 Interpretation

We can see that the ARI / NMI results are very weak, which means that the partitions obtained with the NMF and K-Means are very different and that there is no similarity between these partitions.
These results, however, are quite consistent with the results obtained throughout this report. Indeed the silhouette score of the different K-Means initialization was very low too, which may mean that there is no good separation between classes in the data. It is therefore very likely that several executions of K-Means give us very different scores.

7 Conclusion

To conclude, it should be noted that clustering is a very complicated field because we are travelling in the dark. First of all, we gave the different results and their explanations but it taught us several useful things as data science students. The results were poor even with the determination of better k for k-means and better rank (NMF). The problem comes from the data set itself, which is not adapted and corresponds to the weak points of the NMF. The coefficients are positive numbers certainly, but for each vector in the database, the amount of information is generally a small part that we use to reconstruct our points. Lines with too much variety in itself do not allow the NMF to find a cluster. In particular, it is necessary to realize that while NMF is widely used in science, its rigorous foundation has only been discovered for less than 30 years. Finally, we know very well that the NMF method works very well on images processing, however if we have obtained a bad result does not mean that this method is bad for such studies, but rather we must look for why this result is poor, In our case the construction of our data set plays a very important role, the distribution of groups, the grey level etc.... the application of any algorithm need a consistent data set to obtain a certain positive result. We affirm that k-means is not adapted to this kind of problem, we can also affirm that the NMF as a technique is not adapted to this problem when we have a data set that is not too clear.

References

[1] https://cran.r-project.org/web/packages/aricode/aricode.pdf

[2] P.C. Besse, PCA stability and choice of dimensionality, Statistics & Probability Letters13 (1992), 405–410

[3] https://fr.wikipedia.org/wiki/RStudio

[4] https://jupyter.org/

[5] Renaud Gaujoux, “An introduction to NMF package”, February 18, 2018.

[6] Christian Gagn ,”Learning and recognition”, septembre ,7,2018

[7] Asma Ouji ,”Segmentation and classification in images of scanned documents” , novembre ,8,2012

[8] ”Ciro Donalek”,” Supervised and Unsupervised Learning ” , April 2011

[9] ”Rico Rakotomalala”,”K-means Clustering”,Univeristé Lumiére Lyon 2

[10] ”P.J. FLYNN” ,” Data Clustering ”,The Ohio State University