Work 1. Clustering and factor analysis exercise

Marc Albert Garcia Gonzalo, Miquel Perelló Nieto

October 28, 2012

1 Introduction

In this excercice we have implemented a generic system for analyzing data sets. The system includes the preprocessing of the data, clustering using K-means, and dimensionality reduction using PCA.

2 Project environment

This project has been developed using Matlab 7.12 for Linux.

These are the files of the project:

- *run* : Shell script to run the code from command line
- *main.m* : Main program, which executes all the functions
- *kmeans.m* : Custom K-means algorithm implementation
- *pca.m* : Custom Principal Component Analysis algorithm
- *arffload.m* : Function to load data from ARFF to Matlab
- *distance.m* : Function to calculate distances

To execute the project on a UNIX system, next command needs to be called:

```
./run input_file number_of_clusters images_format
```

For example:

```
./run db/iris.arff 3 .png
```

3 Preprocessing

The preprocessing of the data includes next tasks:

- Loading of data from ARFF files.
- Imputation of missing values.
- Standardization of the data

To load the data from ARFF files (Weka format), we used the existing code from the "dataformat" project [1]. The original script needed some minor modifications to work, and also to process missing values.

Imputation of missing values has been implemented using native Matlab's *knnimpute* method.

All dimensions have been standardized to have mean zero and standard deviation one.

4 Clustering

Clustering of the data has been done using a custom version of the K-means algorithm.

The K-means algorithm has two different parts which are iterated until convergence:

- Assignation of individuals to clusters.
- Position cluster centroids to the center of the individuals.

To do that, first thing is to set the initial position of the clusters. While there are many different approaches to do so, we used the less costly, which is initializing the centroids randomly using the position of individuals.

K-means is guaranteed to converge in a local minimum, so is not deterministic for different initializations. As our initialization is random, K-means can converge in different positions for our algorithm.

5 PCA

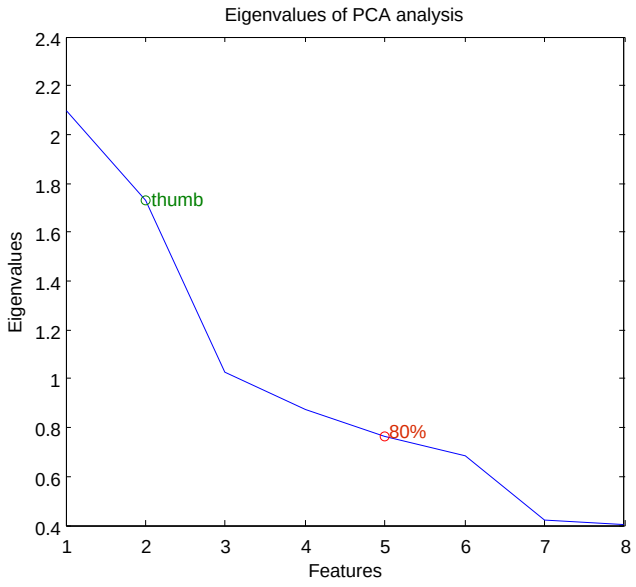
Principal Components Analysis (PCA) has been implemented with a custom version of the algorithm.

PCA creates new dimensions for representing the data, by rotating the original axis. This rotation is done in a way that the first component (dimension) is the one of all possible ones that has a higher variance. Next one is the one having more variance of all the orthogonal to the first one. Next is the one which maximizes the variance of all which are orthogonal to the two first. And so on.

The idea is have a new set of dimensions where the first ones have as much variance as possible, and the last ones as less as possible. This way, we can reduce dimensionality by using only the N first ones.

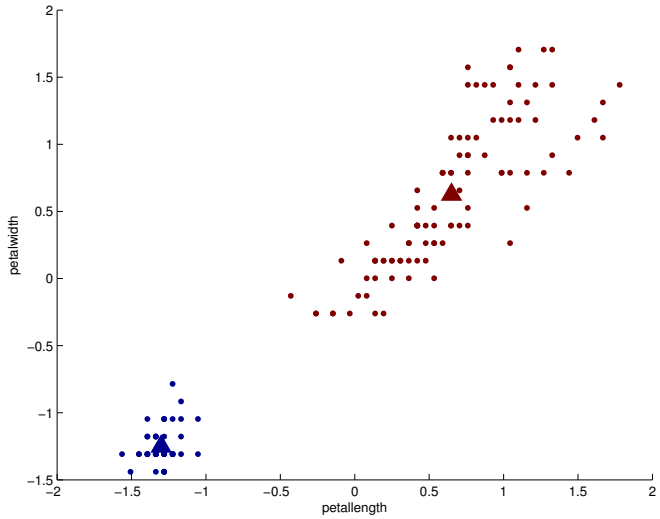
There are different criteria to select how many dimensions to use, and it is specific to the problem we are working on. Most common criteria are probably choosing all the dimensions with and eigenvalue higher than the mean of all eigenvalues. Also the last elbow criterion, which represents the eigenvalues graphically, and selects the ones until the last elbow, after which the lost of variance is minimum, as the variance on the dimension is also minimum.

Next, there is a representation of the eigenvalues of the data *Diabetes*:

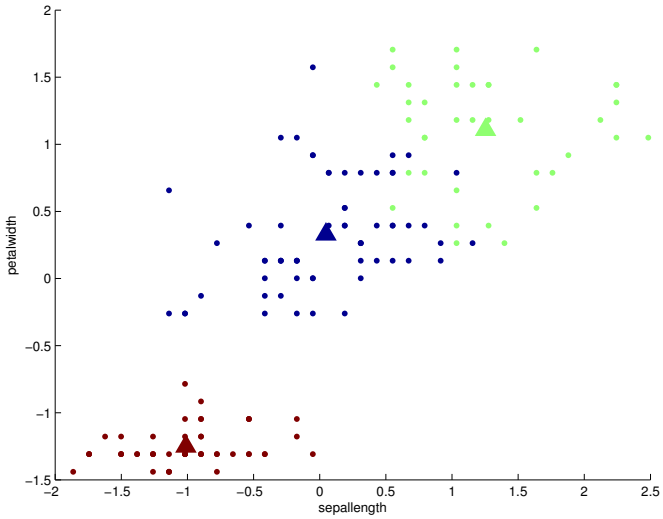


6 Iris

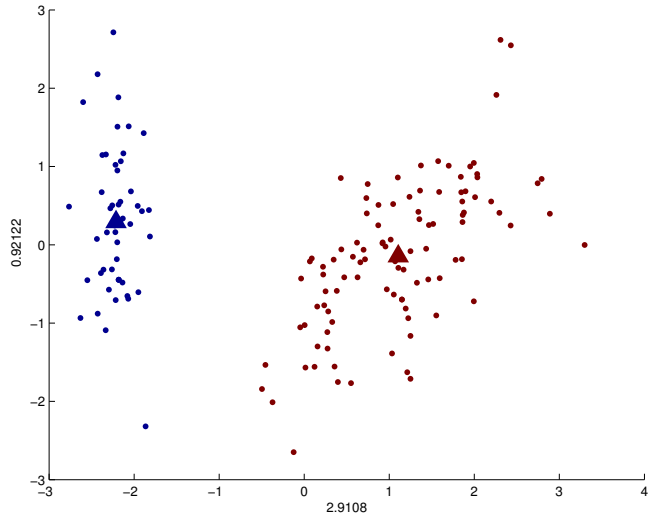
The Iris data set contains information about individuals of three different classes of the Iris specie. Attributes contain information about sepal and petal length and width. There are 150 instances.



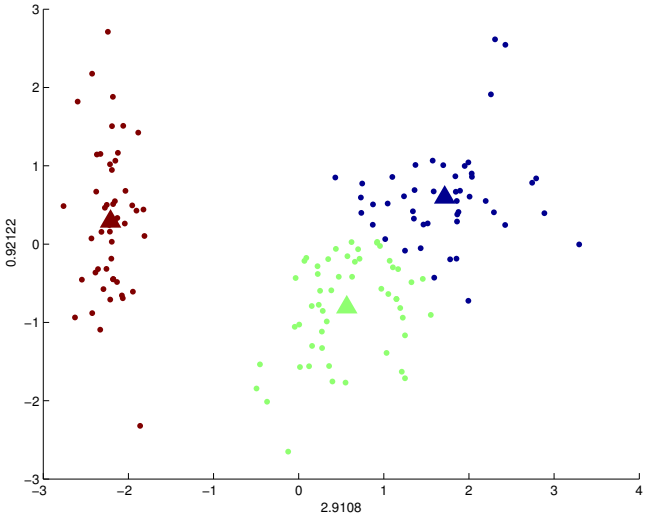
(a) K-means with 2 clusters



(b) K-means with 3 clusters



(a) Two clusters with the PCA best features.



(b) Three clusters with the PCA best features.

While it is known that the Iris data set is composed of three different classes of plants, when performing clustering we see that it probably makes more sense to divide data in two different classes.

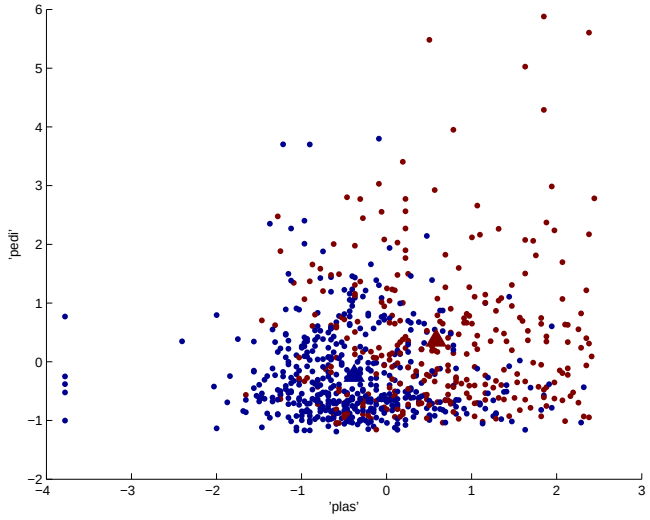
Probably, if we compare our clusters with the classes, we would observe that one of the clusters contain two different classes, the two more similar ones.

After performing PCA, we see how individuals are more spread out, and it still exists one cluster which is farer from the rest of the data.

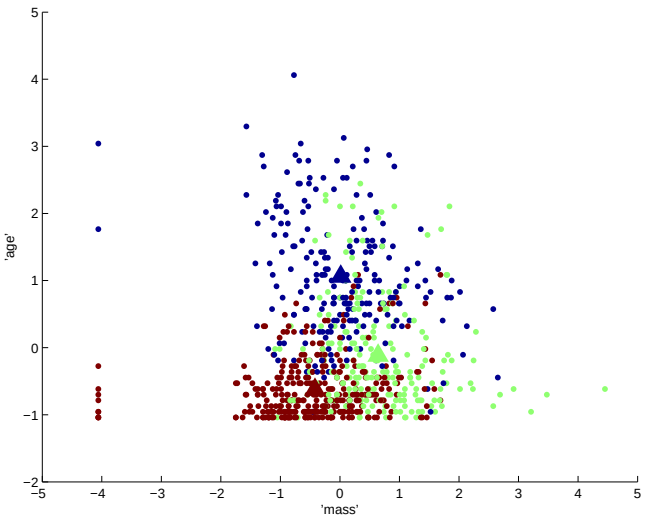
7 Pima Indians Diabetes

This data set contains information about female Pima Indian individuals. Pima Indians have a higher rate of diabetes than normal, and this data set is used to study the causes.

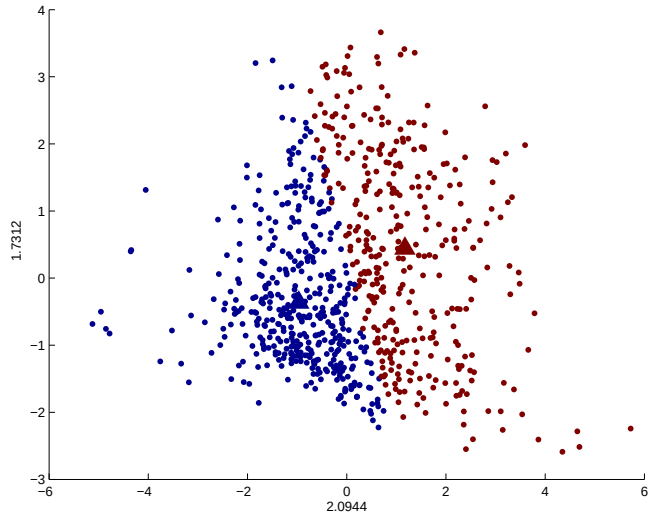
There are 768 individuals on the data set. Attributes contain medical information such as the age, the number of times the individual was pregnant, and the results of several medical tests.



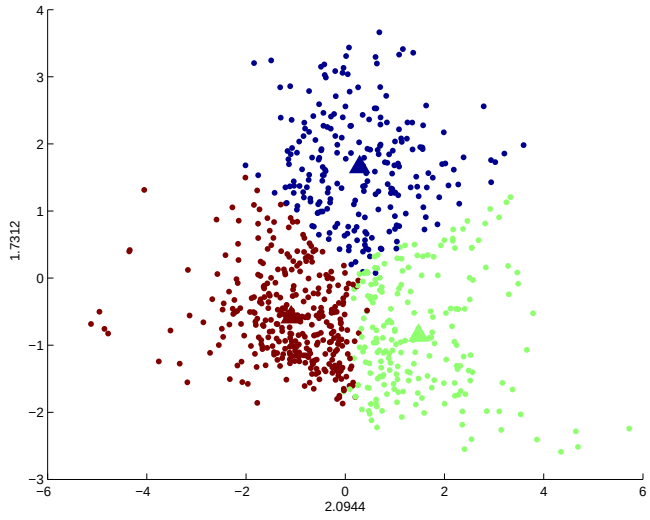
(a) K-means with 2 clusters



(b) K-means with 3 clusters



(a) Two clusters with the PCA best features.



(b) Three clusters with the PCA best features.

In the first approach executing K-means algorithm to all features is so difficult to see any difference between predicted clusters. In that case you can select two or three clusters indifferently, but the cloud of points remain nearby.

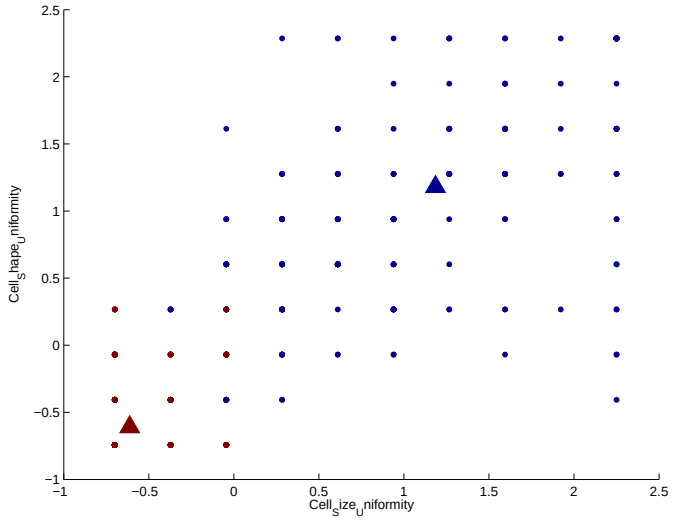
On the other hand, if we compute the PCA algorithm and select the features with highest eigenvalues then it is easy to see a division in two and in three clusters.

8 Breast Cancer

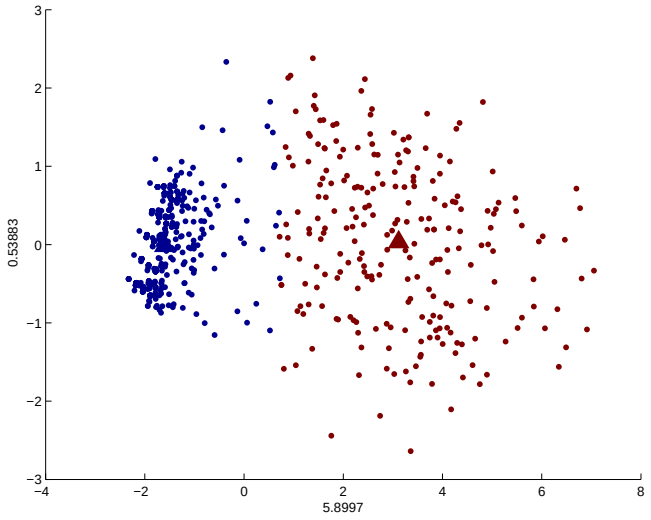
This data set contains information about the results of tests on breast cancer based on visual characteristics of the result of a Fine Needle Aspiration. Data is gathered by taking measures from a microscope scan.

There are 699 samples. Attributes are visual features of the cells, such as radius, perimeter, area, symmetry, etc.

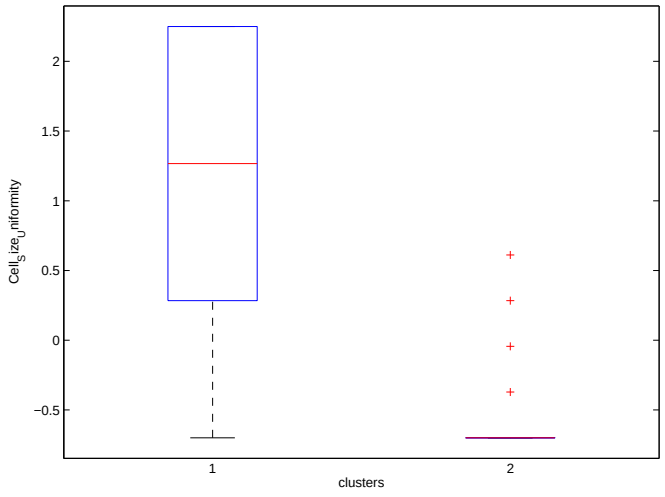
Data set contains missing values.



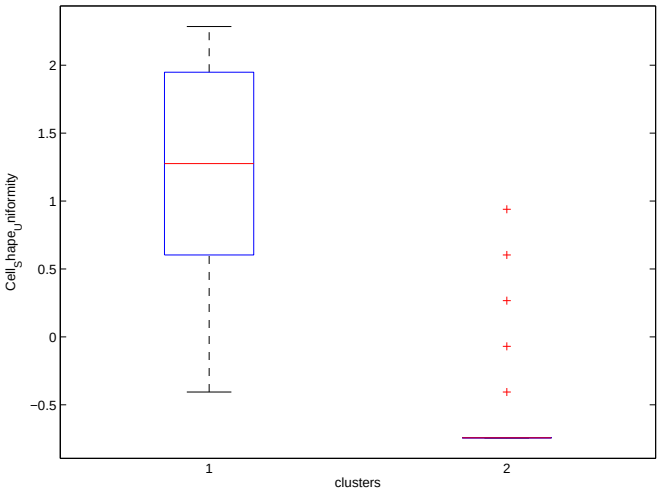
(a) Clustering with original dimensions



(b) Clustering with principal components



(a) Cell size uniformity by cluster



(b) Cell shape uniformity by cluster

Clustering has been implemented with 2 to 5 number of clusters. Best results have been observed for 2 clusters.

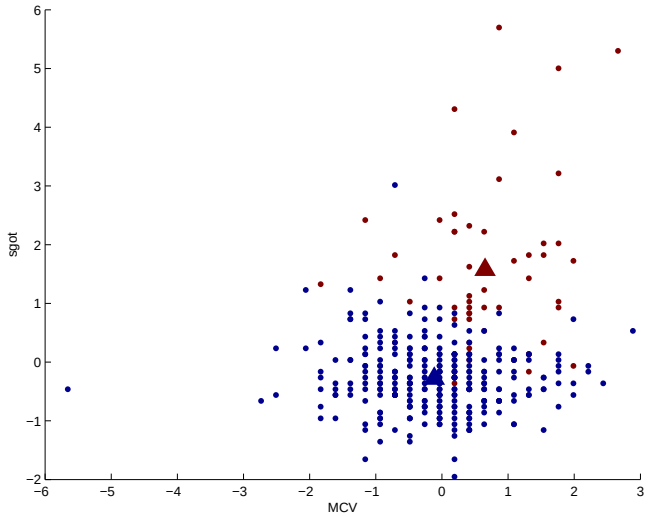
From the graphics we can see how instances are distributed mostly randomly on the space. After clustering we observe that one of the clusters contains the individual with lower values in most variables, while the other contains the individuals with higher values. This is specially true for the variables *cell size uniformity* and *cell shape uniformity*, as is shown in the scatter plot, and in the box plots of both variables for cluster.

After performing PCA, we observe some structure for the two principal components. While K-means is able to recognize the structure, and to associate one cluster to each group, it falis to associate some individuals, as the clusters need to have the same size on the space.

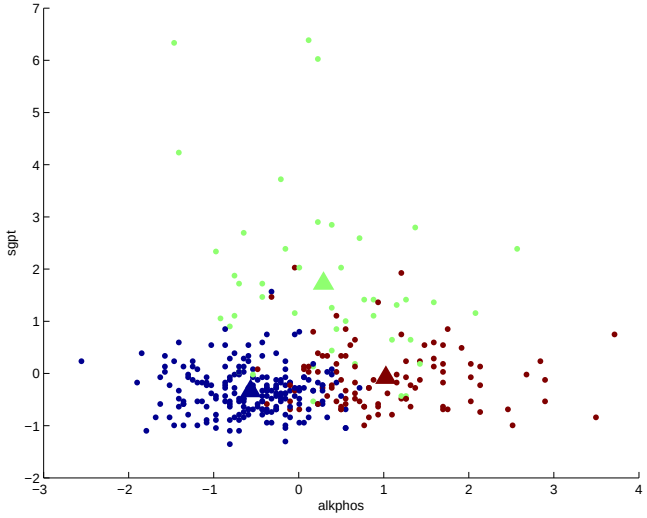
9 Liver disorders

The BUPA liver disorders data set contains information about blood tests on male individuals, which could be related to lever disorders.

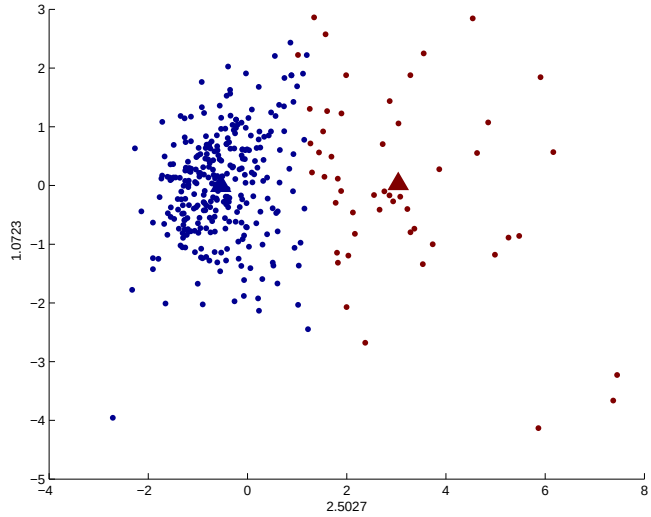
There are 345 instances, and the attributes are the results of the blood tests, as well as other medical information about alcohol consumption and corpuscular volume.



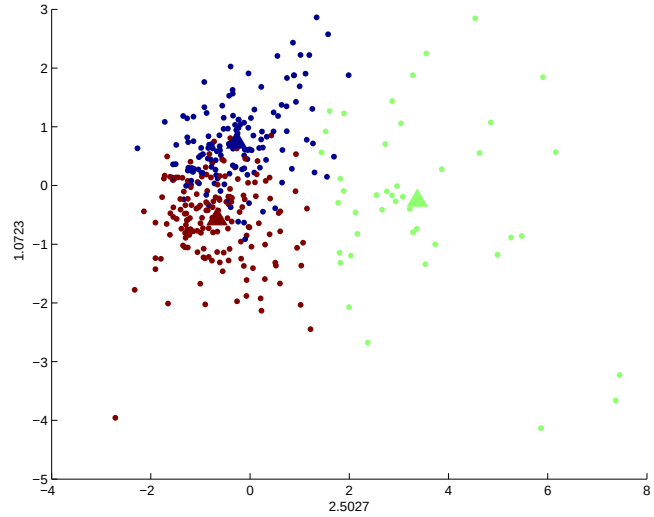
(a) K-means with 2 clusters



(b) K-means with 3 clusters



(a) Two clusters with the PCA best features.



(b) Three clusters with the PCA best features.

We can observe in the first image that this two features *mean corpuscular volume* and *aspartate aminotransferase* can make a separation between two clusters clearly. If we try to make three clusters it is most difficult to see an easy aproach, but in the features *alkaline phosphotase* and *aspartate aminotransferase* it is possible to divide in this three spaces.

If we compute the PCA the new features another good space, in that the division in two clusters seems easier than in three, like in the other case is difficult to say that it really have more than two clusters.

References

[1] The ”dataformat” project https://ml01.zrz.tu-berlin.de/trac/dataformat/browser/trunk/matlab/arffload.m.

[2] Wikipedia: Principal component analysis http://en.wikipedia.org/wiki/Principal_component_analysis

[3] Wikipedia: k-means clustering http://en.wikipedia.org/wiki/K-means_clustering