HW3 Report : Unsupervised Learning

Subject :
CS 7641 Spring 2017 - OMS

Author :
Melisande Zonta Roudes
GT account name : mzr3

The goal of this project is to see the effect of the clustering algorithms and of dimensionality reduction methods. Expectation Maximization (EM) and KMeans were used in clustering and Principal Component Analysis (PCA), Independent Component Analysis(ICA), Randomized Projection(RP) and finally as a personal choice, Linear Discriminant Analysis were tested. The whole was applied to train a neural network.

1 Dataset Choices

The datasets that I chose in the first assignment were not adapted to this one, indeed the "Letter Recognition" dataset with its 26 classes and its 20000 samples was to difficult to run with the clustering algorithms and the "Pima Indians Diabetes Data Set" gave too much bad results. Hence I took two new datasets "Breast-Cancer-Wisconsin" and "Digits".

1.1 First Dataset

The "Breast-Cancer-wisconsin" comes from the results of the University of Wisconsin Hospitals. As we can't deduce from the diagnosis, this is a binary dataset with 10 attributes (medical parameters) and 683 instances. The output is either 2 for benign and 4 for malignant. One feature was removed because it corresponded to a sample code number which is an id number.

1.2 Second Dataset

The second dataset called "Digits" is a dataset included in the sklearn library. It is interesting in the way it is a really different dataset from the previous one, since the attributes are image's pixels of 8×8 images of a digit (between 1 and 9). it is composed of 1797 instances and thus 64 features.

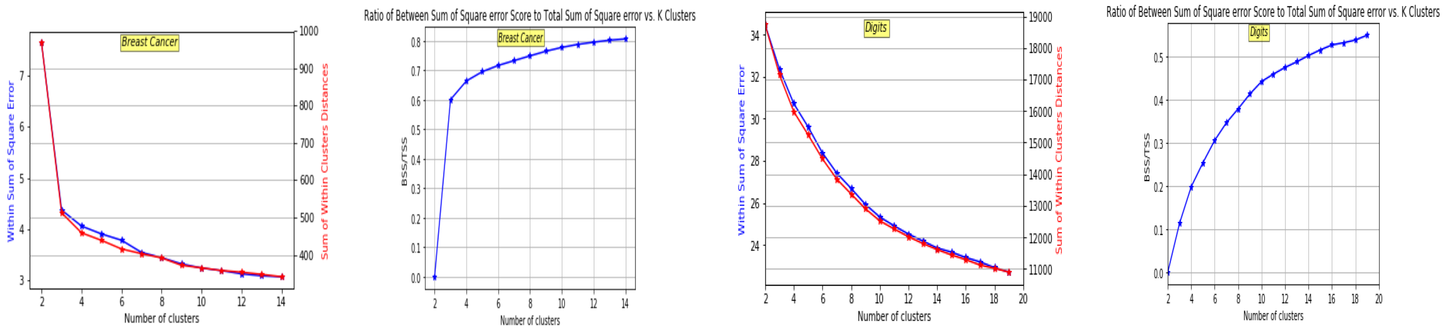
2 Clustering

Clustering is an unsupervised method of classification which has for task to find the dataset's hidden structure from unlabeled data. That's why the aim of the clustering method is to group instances that are similar to each other. In order to implement this part, functions from Scikit Learn were used.

2.1 K-Means

Theory K Means is one of the simplest unsupervised learning algorithm that solve the clustering problems. The procedure aims to classify a dataset through a certain number of clusters. The main idea is to define k centroids (one per cluster). These points should be placed cautiously because this algorithm is really sensitive to the initialization and the clustering can be different according to this placement. The goal is to place them as far away from each other as possible. And thus maximizing the distance inter clusters and minimizing the intra one. It assigns each point to one of the k clusters based on a specific distance metric such as manhattan distance (norm 1) or euclidean norm (norm 2). With a loop, the k centroid change their location step by step until no more changes are done. The clusters centers keeps being reassigned.

Results Two specific metrics of K Means were used to demonstrate the choice of the appropriate number of clusters. As we spoke of the importance of the within clusters distances, we plot the within cluster sum of square error according to the euclidean distance as well as the sum of the within cluster distances with the Manhattan distance. Furthermore, we represent the ratio of Between Sum of Squares and Total Sum of Squares.



a b c d

Figure 1: *a.* WSS and SWCD for Breast Cancer Dataset . *b.* BSS over TSS for Breast Cancer Dataset. *c.*WSS and SWCD for Digits Dataset. *d.* BSS over TSS for Digits Dataset.

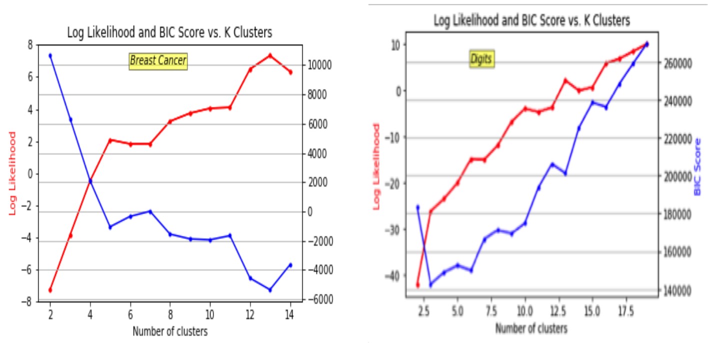
Breast Cancer In order to choose the appropriate K we compare the two metrics mentioned above on an increasing number of clusters from 2 to 14. It will allow us to check if the first intuition of 2 clusters is right. Using the "Elbow Method", the curves 3 a. gives K = 3 or 4 as well as b. provides 3 clusters.

Digits The same measures were done on a varying number of clusters from 2 to 20. The "Elbow Method" is more difficult to apply on the 3 c. but the 3 d. provides a slight change at 10 clusters.

2.2 Expectation Maximization

Theory The Expectation Maximization is a probabilistic distribution based algorithm. As it can K Means assign in a hard way a data point to one particular cluster on convergence and as it makes use of the euclidean or manhattan norm when optimizing the centroid coordinates, in contrast EM soft assigns a point to clusters by giving a probability to any centroid and it is based on the Expectation ie the probability of the point belonging to a particular cluster. This makes K means biased towards spherical clusters. The model implemented with Scikit Learn is Gaussian Mixture Modeling (GMM) which is the most popular variant of EM.

Results Again two metrics were used to analyze EM. As it is the main characteristic of this algorithm, the Log Likelihood is plotted as well as the Bayesian Information Criterion (BIC). Indeed the BIC score is independent of the prior and can measure the efficiency of the parametrized model in terms of predicting the data and above all can be used to choose the number of clusters according to the intrinsic complexity present in a particular dataset. The current behaviour of the log likelihood is the increasement of this one while k is increasing. The curve flattens for high values of k.



a b

Figure 2: *a.* Log Likelihood and BIC score for Breast Cancer Dataset . *b.* Log Likelihood and BIC score for Digits.

Breast Cancer While the Log Likelihood increases with K, whereas BIC score decreases. The curve 2a. is hard to exploit, an inclination could lead us to conclude to K = 5/6. But the elbow method currently used for these indicators are hardly appliable.

Digits The curve 2b. is again rather impossible to interprete.

Thus the metrics chosen for EM don't provide much interesting conclusions.

2.3 Conclusion

In order to compare our two clustering methods, 2 metrics in common were chosen. The first one is silhouette width which is the average measure of the degree of confidence in a particular clustering task. Values are between -1 and 1 with 1 being the indicator of a well-clustered dataset. The second one is a mean between two metrics (homogeneity and completeness) indeed we represent the V-measure. The completeness measure reaches its maximum if all the data points that are members of a given class are elements of the same cluster and homogeneity measures if all of its clusters contain only data points which are members of a single class. Hence we are searching for the maximum of those two both metrics.

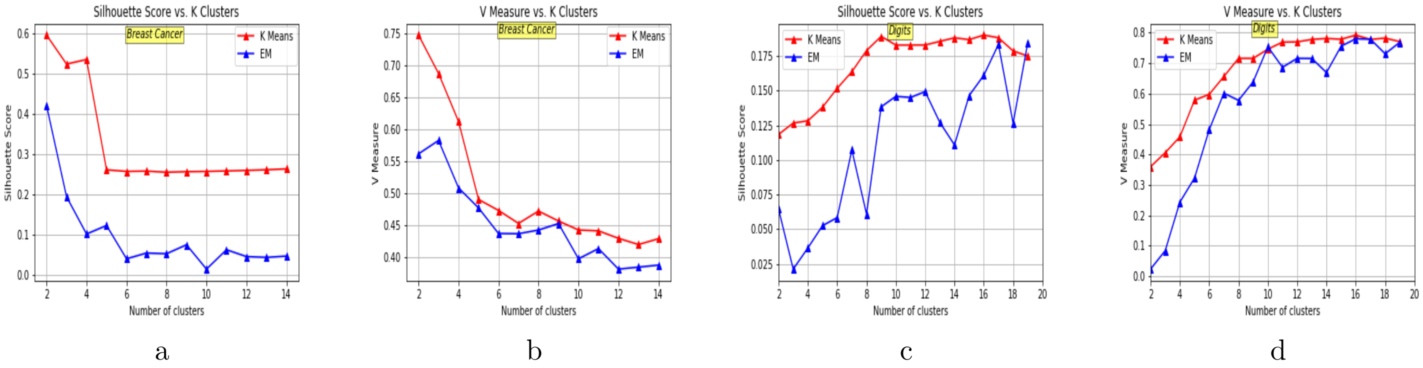


Figure 3: *a.* Silhouette curves EM/K Means for Breast Cancer Dataset . *b.* V Measures curves EM/K Means for Breast Cancer Dataset *c.*Silhouette curves EM/K Means for Digits Dataset . *d.* V Measures curves EM/K Means for Digits Dataset

Breast Cancer According to what we said before, we can deduce from the curve 3a. that the maximum is reached of the two indicators for K = 2. for both EM and K-Means. The V Measure provides two contradictory results indeed EM provides the optimal number of clusters K = 3 whereas K Means gives K = 2.

Digits The results for this dataset are much less satisfying since the Silhouette Score gives K = 9 for K-Means and K = 17 or 19 for EM where as V Measure agreed on K = 19 for both methods.

3 Dimensionality Reduction

The aim of the dimensionality reduction methods is to restructure the input before trying to classify it with learning algorithms. These methods are motivated by the "curse of dimensionality", indeed when the number of features is too important, it becomes really hard to compute any classification.

3.1 Principal Component Analysis

Theory The PCA is a statistical procedure that uses a transformation to convert a set of observations that are possibly correlated into linearly uncorrelated variables which are called principal components. It results into a smaller number of components than the original number of features. Moreover the first principal component has the largest possible variance and it goes on in a decreasing way for the others components. The resulting vectors are an uncorrelated orthogonal basis set.

Step 1 : Dimensionality reduction As the variance has an important role in the PCA method, the eigen values were represented as well as the cumulative distribution. In order to see the efficiency of this algorithm, a fast classification algorithm as Support Vector Machine is tested. The accuracy reported after reduction of the dataset is compared to the one obtained directly with the original dataset. This measure is made with an increasing number of clusters from 2 to the max number of features. Finally, it would be deceiving not to visualize our clusters onto our new principal components.

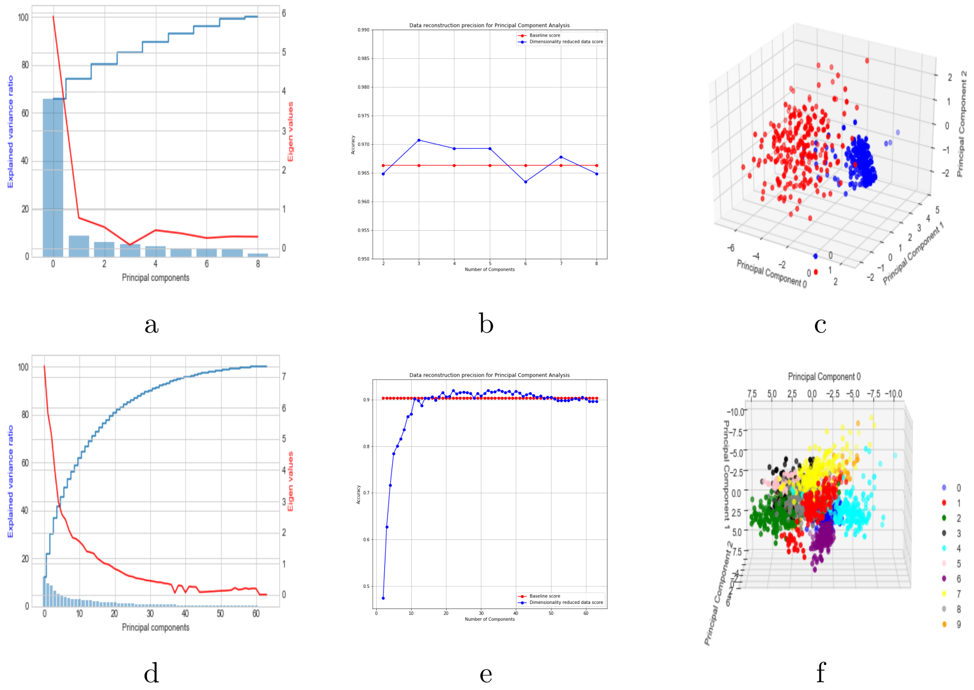


Figure 4: *a.* Eigen Values and Variance for Breast Cancer Dataset. *b.* Learning curve for Breast Cancer Dataset *c.*Representation 3D of reduced Breast Cancer Dataset *d.* Eigen Values and Variance for Digits Dataset. *e.* Learning curve for Digits Dataset *f.*Representation 3D of reduced Digits Dataset

Breast Cancer As we can see on the bar plot 4a. the first component reaches more 60% hence to obtain 90% 5 components are enough. We can also observe the behaviour of this algorithm through the link between the eigen values and the cumulative variability.Indeed, a singular value decomposition is performed and the eigen values are sorted in order to select then the eigen vector which correspond. The accuracies provide others hints on the efficiency of our reduced data set indeed e can see that with the three components, we can obtain an accuracy as big as the original dataset and its 8 features. The representation in 3D shows us the separation between the labels 0/1 which demonstrates that PCA is really efficient for this dataset

Digits The bar plot 4a shows we could reduce the dataset to 30 components (90%) and the accuracy would prove that only 10 components would be enough to reach the accuracy of the entire number of features. The 3D representation shows the distinct 10 clusters of our dataset which is a success for our projection method.

Step 2 : Clustering Now we have proceeded in the dimensionality reduction, we can apply a clustering algorithm. It will allow us to visualize if the optimal number of clusters has changed.

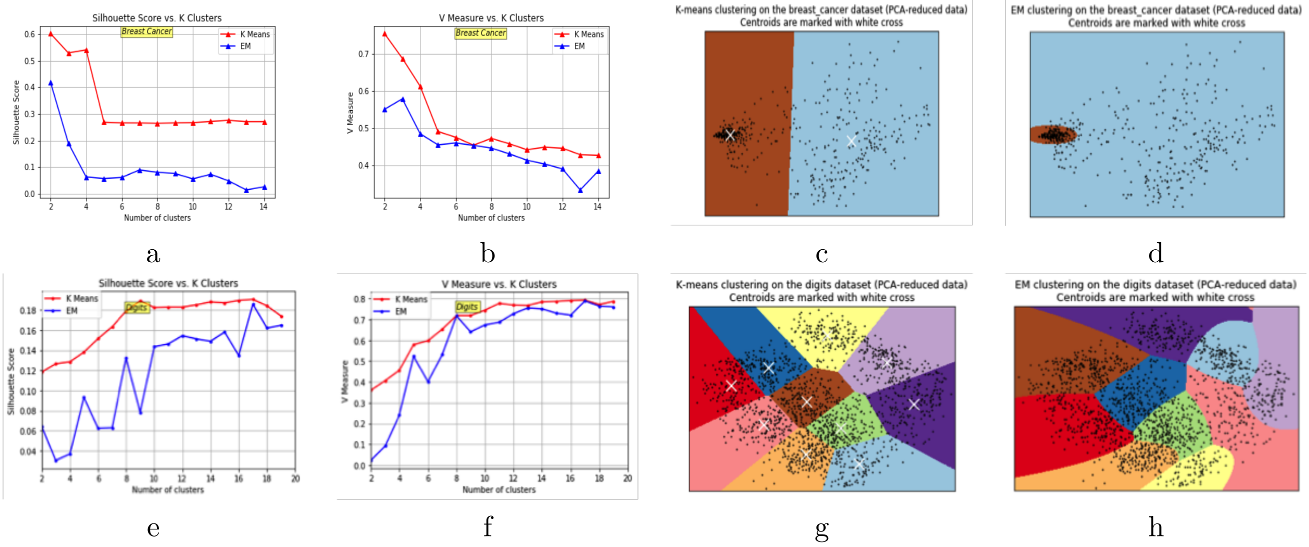


Figure 5: *a.* Silhouette curves EM/K Means for Breast Cancer Dataset . *b.* V Measures curves EM/K Means for Breast Cancer Dataset *c.* K Means *c.* EM *e.*Silhouette curves EM/K Means for Digits Dataset . *f.* V Measures curves EM/K Means for Digits Dataset *g.* K Means *h.* EM

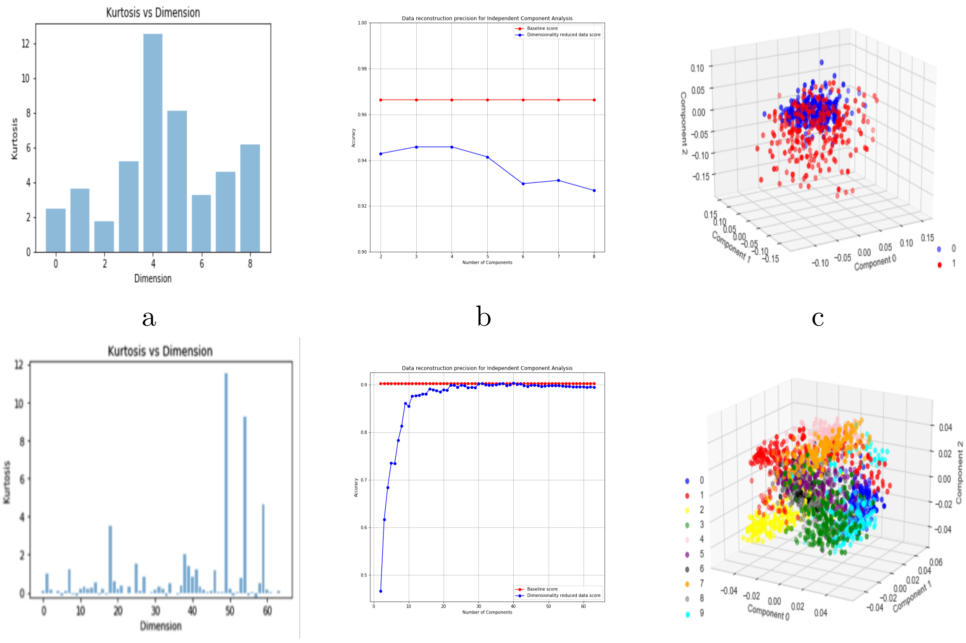
Breast Cancer The curve 5a. shows that the number of clusters is K = 2 ,the V Measure metric gives K = 3 for EM and 2 for K Means. We can visualize in 2D the clustering, indeed the K Means provide the zones centered each on a centroid. The points are quite spread especially the label 0. The EM modelisation is itself quite different since the cluster brown (label 1) is really condensated and small.

Digits The Silhouette score seems to provide K = 10 for K Means but K = 17 for EM, the V Measure score don't provide good results again since the maximum is detected for 17 in both cases. We can remark from the 2D representations, that K means provide hard boudaries between the 10 classes whereas EM is softer due to its probabilistic computation since weights are allocated to each class.

3.2 Independent Component Analysis

Theory Independent Component Analysis is a computational method for separating components. This is done by assuming that the subcomponents are non -Gaussian signals and that they are statistically independent from each other. The two broadest definitions of independence for ICA are the minimization of mutual information and the maximization of non gaussianity.

Step 1 : Dimensionality reduction As we defined above, the non gaussianity measure is important and is represented by the kurtosis. Indeed, kurtosis is a descriptor of the shape of a probability distribution. This a measure of the "tailedness" of the probability distribution. Hence as kurtosis measure the gaussianity of a distribution, if we want to maximize the non gaussianity, we will minimize the kurtosis. This is what we will do by selecting the first components that have the minimum value of kurtosis. As before, we will test the efficiency of this reduction on SVM.



a b c

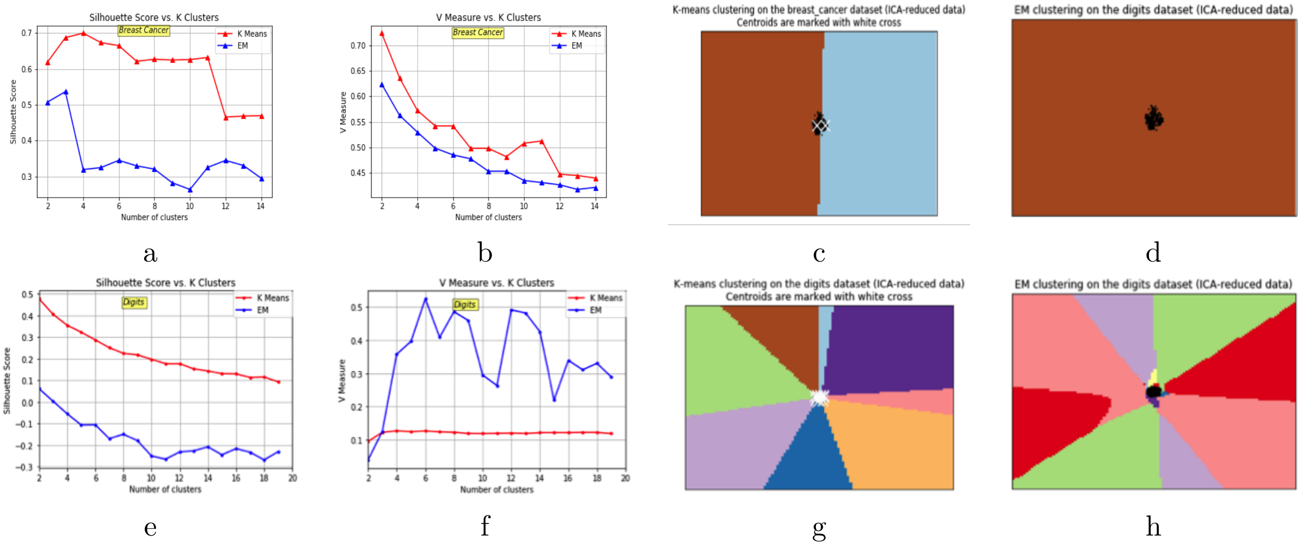
d e f

Figure 6: *a.* Kurtosis for Breast Cancer Dataset. *b.* Learning curve for Breast Cancer Dataset *c.*Representation 3D of reduced Breast Cancer Dataset *d.* Kurtosis for Digits Dataset. *e.* Learning curve for Digits Dataset *f.*Representation 3D of reduced Digits Dataset

Breast Cancer By looking at the bar plot 6, we can visualize some minimum values which will be reported on the representation. The accuracy of our reduced dataset is below the baseline one even if we can remark that the maximum is reached for only 3 components. The 3D representation does not provide distinct clusters so the result is not as conclusive as PCA.

Digits The same reasoning is made on the kurtosis of digits. The accuracy provide an interesting result since we see that only 20 components are sufficient. The 3D plot for digits show a mix of many classes hence the result is not satisfactory.

Step 2 : Clustering As before, we apply our clustering algorithms.



a b c d e f g h

Figure 7: *a.* Silhouette curves EM/K Means for Breast Cancer Dataset . *b.* V Measures curves EM/K Means for Breast Cancer Dataset *c.* K Means *c.* EM *e.*Silhouette curves EM/K Means for Digits Dataset . *f.* V Measures curves EM/K Means for Digits Dataset *g.* K Means *h.* EM

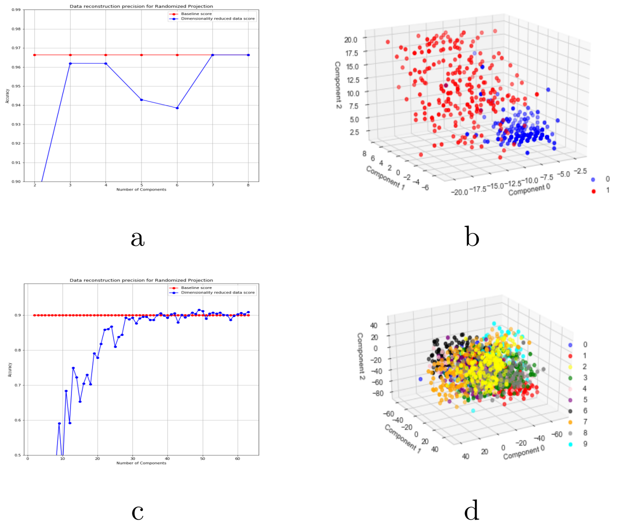
Breast Cancer The curve 7a. shows that the number of clusters is K = 4 for KMeans and K = 3 for EM , the V Measure metric gives K = 2 for EM and K Means. We can visualize in 2D the clustering, the points are condensed on the line separating the two classes and the centroids are quite overlapping. The EM result is more astonishing since no label 1 can be seen.

Digits The Silhouette score seems to provide K = 2 for K Means and K Means which is really anormal, the V Measure score is quite oscillatory but a maximum can be detected on K = 6. The 2D representations show again centered points but the classes are distinct.

3.3 Randomized Projection

Theory Random Projection is a technique used to reduce the dimensionality of a set of point which lie in Euclidean space. Random projection are powerful methods known for their simplicity and less erroneous output compared with other methods. Random projection preserve distances well. Random projection is a simple and computationally efficient way to reduce the dimensionality of data by trading a controlled amount of error for faster processing times and smaller model sizes. The dimensions and distribution of random projection matrices are controlled so as to approximately preserve the pairwise distances between any two samples of the dataset.

Step 1 : Dimensionality reduction As it is usually the case, it is advised to compute several iterations of the algorithm to obtain results that are less "random". Again, it is interesting to see if those kind of algorithm provide good results on classification task.



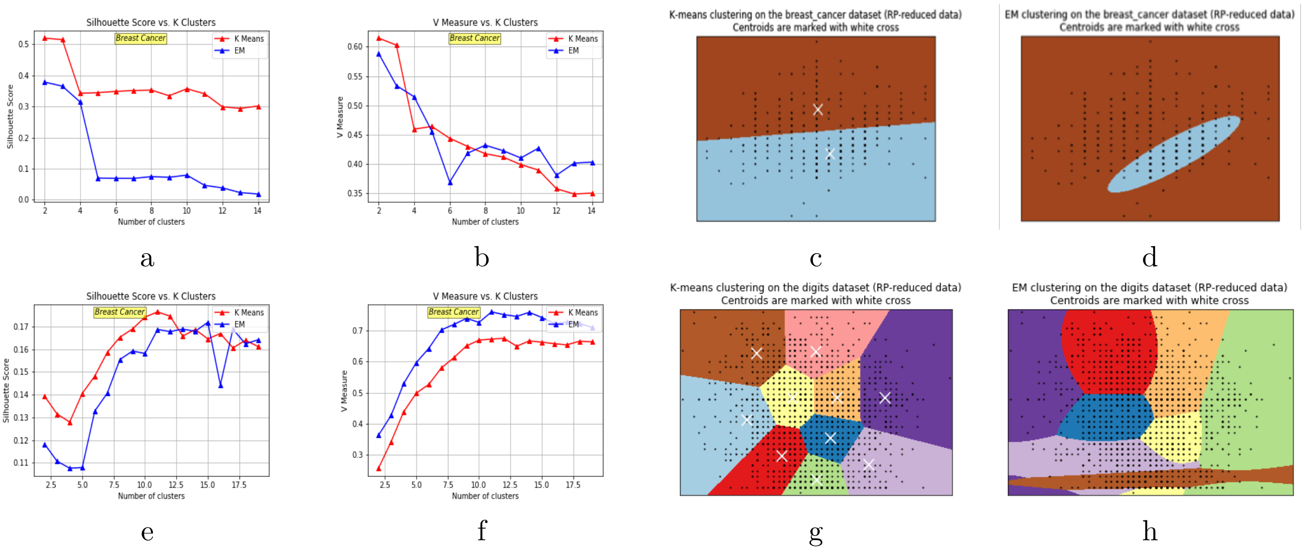
a b c d

Figure 8: *a.* Learning curve for Breast Cancer Dataset *b.*Representation 3D of reduced Breast Cancer Dataset *c.* Learning curve for Digits Dataset *d.*Representation 3D of reduced Digits Dataset

Breast Cancer We can visualize that with 7 components out of 8, the max accuracy is reached and that 3/4 are not so bad. The visualization provided shows distinct clusters, the red one is well spread and the blue one more condensed.

Digits The accuracy provide an interesting result since we see that only 25 components are sufficient. The 3D plot for digits show a mix of many classes hence the result is not satisfactory which is normal since it is a random algorithm.

Step 2 : Clustering The clustering methods are then applied.



a b c d

e f g h

Figure 9: *a.* Silhouette curves EM/K Means for Breast Cancer Dataset . *b.* V Measures curves EM/K Means for Breast Cancer Dataset *c.* K Means *c.* EM *e.*Silhouette curves EM/K Means for Digits Dataset . *f.* V Measures curves EM/K Means for Digits Dataset *g.* K Means *h.* EM

Breast Cancer The curve 7a. shows that the number of clusters is K = 2 , the V Measure metric gives K = 2 for EM and K Means. We can visualize in 2D the clustering, the points are spread on the 2 classes and classes are half each. The EM result provides a cluster of 1's surrounded by 0's. **Digits** The Silhouette score seems to provide K = 10 for K Means and EMwhich is really great, the V Measure score furnishes a plateau around K = 10. The 2D representations show again distinct classes for K Means as well as for EM with more uncertain boundaries.

3.4 Linear Discriminant Analysis

Theory Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant. LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data.PCA on the other hand does not take into account any difference in class. The Fisher Criterion wants to minimize the difference of the mean between two classes over the sum of the standard deviations. Hence a computation of the d-dimensional mean vectors for the different classes from the dataset is done.Then the scatter matrices are calculated (in-between-class and within-class scatter matrix) then the eigenvectors and eigenvalues which will be sorted by decreasing eigenvalued. The eigenvector matrix is finally used to transform the samples onto the new subspace.

Step 1 : Dimensionality reduction The same test on SVM were computed for this algorithm as well as the representation of the clusters.

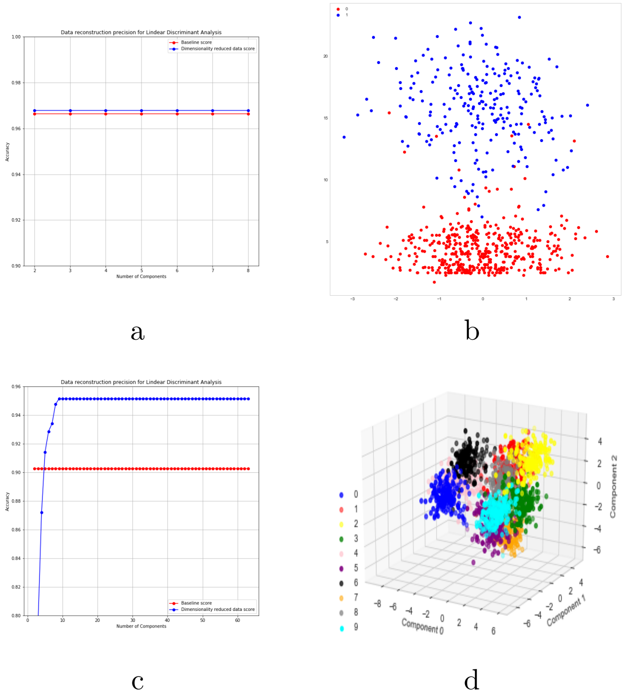


Figure 10: *a.* Learning curve for Breast Cancer Dataset *b.*Representation 3D of reduced Breast Cancer Dataset *c.* Learning curve for Digits Dataset *d.*Representation 3D of reduced Digits Dataset

Breast Cancer By looking at the eigen values, I remarked that only one component has been kept in the transformation that's why we can see that the accuracy is constant since it is the one for one component. The visualization provided shows distinct clusters, the red one is well spread and the blue one more condensed. It was represented in one dimension according to the results.

Digits The accuracy provide an interesting result since we see that only 8 components are sufficient to go higher than the baseline accuracy. The 3D plot show distinct classes which proves that this algorithm provides great results.

Step 2 : Clustering The clustering algorithms are applied to the reduced dataset.

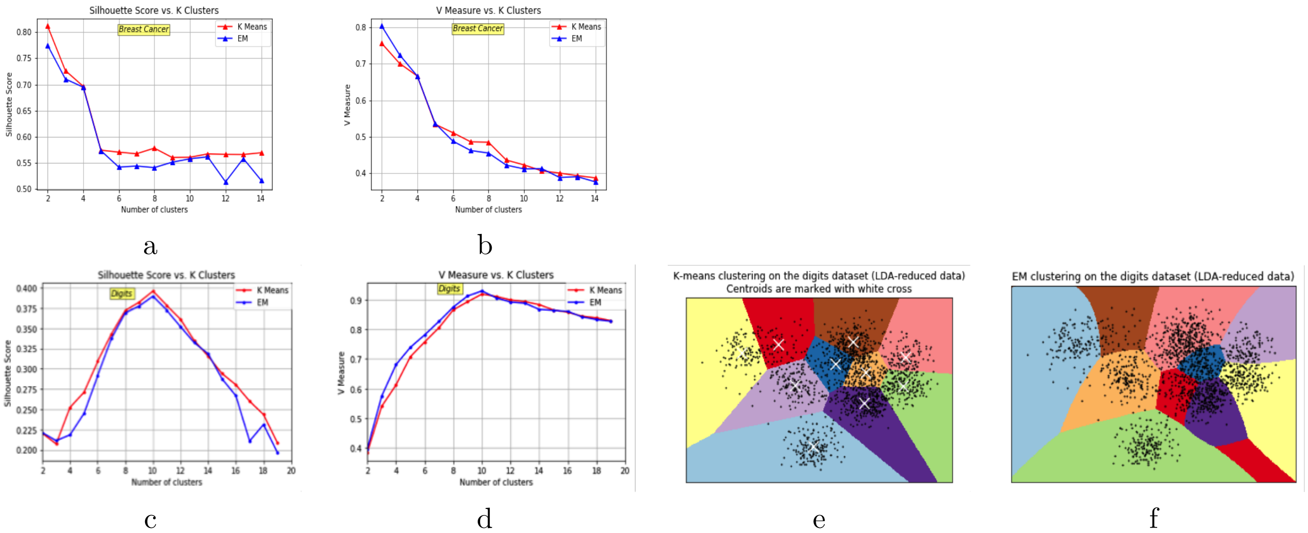


Figure 11: *a.* Silhouette curves EM/K Means for Breast Cancer Dataset . *b.* V Measures curves EM/K Means for Breast Cancer Dataset *c.*Silhouette curves EM/K Means for Digits Dataset . *d.* V Measures curves EM/K Means for Digits Dataset *e.* K Means *f.* EM

Breast Cancer The curve 7a. shows that the number of clusters is K = 2 , the V Measure metric gives K = 2 for EM and K Means. We can visualize in 2D the clustering, the points are spread on the 2 classes and classes are half each. The EM result provides a cluster of 1's surrounded by 0's. **Digits** The Silhouette score provides the best result we had for now K = 10 for K Means and K Means which is really great, the V Measure score provides also K = 10. The 2D representations shows again distinct classes for K Means as well as for EM with more uncertain boundaries. The results are really clean and satesfactory, we can see the amount of points in each cluster.

4 Application on a Neural Network

After studying clustering and dimensionality reduction to the 2 datasets that were not the ones I chose for the first assignment. I decided to study the performances of neural network by training it on the dataset Pima Indians Diabetes. The implementation was done on Scikit learn again by reusing the code of neural networks of the first assignment. The Multi Layer Perceptron function was used. As a reminder we will keep in mind the learning curve we obtained in the first assignment on this dataset.

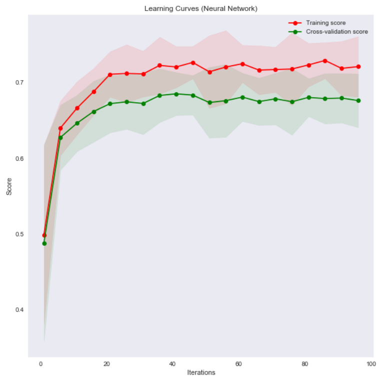
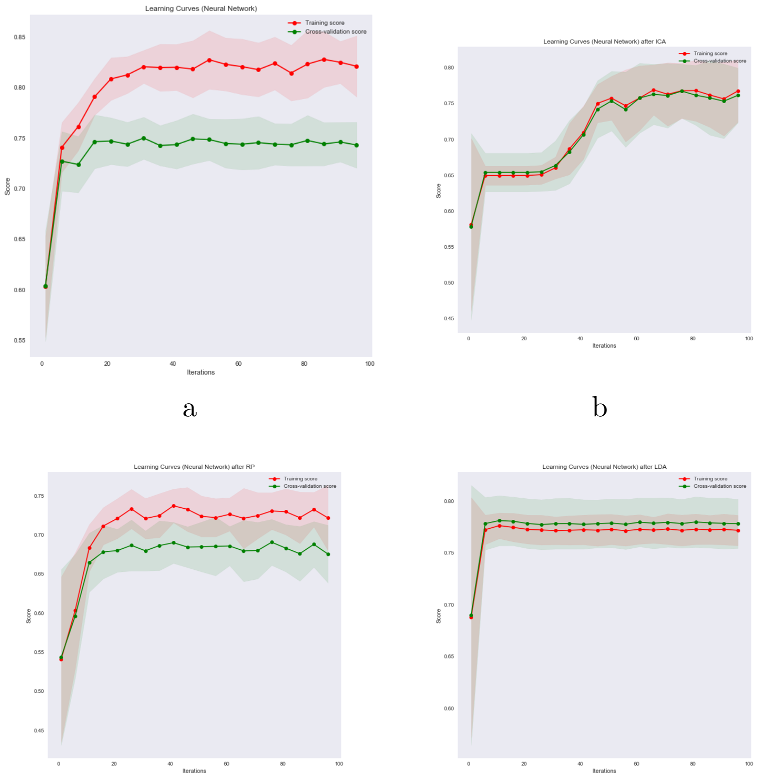


Figure 12: learning curve baseline

4.1 Dimensionality Reduction on a Neural Network

We will apply the 4 dimensionality reduction algorithms on our dataset in order to train it.



a

b

c

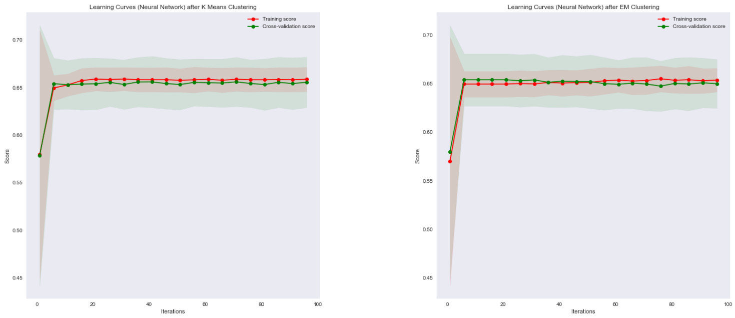
d

Figure 13: *a.* Learning curve after Principal Component Analysis . *b.* Learning curve after Independent Component Analysis . *c.* Learning curve after Randomized Projection . *d.* Learning curve after Linear Discriminant Analysis

The learning curves had been plotted over the iterations in order to notice an evolution of the accuracy over more iterations. We can see on the curve 14a. that the training score is around 80% which is more than the 70% of the baseline and the testing score is also better. The curve 14b. shows not good results, just training and testing accuracies which are really closed. Although the maximum at 100 iterations reached 75% We can see on the curve 14c. that the results seem like the ones for pca but with lower accuracies. Finally, the curve 14d. seems to be the best with training score and testing score high for this dataset and close from each other.

4.2 Clustering on a Neural Network

The two clustering algorithms are applied in order to train the Neural Network



a

b

Figure 14: *a.* Learning curve after K Means Clustering . *b.* Learning curve after Expectation Maximization

The two learning curves demonstrate that the clustering applied without a previous dimensionality reduction provide bad results, as we can see the accuracies close to 65%. But we can observe that the two algorithms provide quite the same results.

5 Conclusion

This assignment allows to handle the effect of both the dimensionality reduction algorithm and the clustering ones. It allows to decrease the processing time and avoid the "curse of dimensionality". I had been surprised of the ameliorations of the results of my accuracies concerning a dataset with a usual low accuracy.