Master in Artificial Intelligence

INTRODUCTION TO MACHINE LEARNING

Work 2

Authors:
Ruizhe Yu Xia
Jan Rodríguez Miret
Jose Juan Galán Montero

Course 2020-2021

Contents

1 Introduction **2**

2 PCA **2**

2.1 Vowel Dataset 2

2.2 Waveform Dataset 2

2.3 Adult Dataset 3

2.4 Sklearn Comparison 3

3 Effects of Applying PCA before K-means **4**

3.1 Vowel Dataset 4

3.2 Waveform Dataset 5

3.3 Adult Dataset 6

4 Visualizing the Data: Comparison **8**

4.1 Vowel Dataset 8

4.2 Waveform Dataset 9

4.3 Adult Dataset 11

5 Conclusions **12**

6 Execution Guide **13**

References **13**

1 Introduction

In this second practical work of the subject, our goal is to see the effect of dimensionality reduction and to compare different visualizations methods. More specifically, we are going to implement our own PCA to get the principal components of the datasets, execute and compare the results with K-Means algorithm and then plot them using t-SNE and the same PCA technique, in a lower-dimensional space.

In this case, we have chosen to represent the data in 2D plots, rather than 3D, because they are much faster to compute.

For this second work, we have changed the Connect-4 dataset for the Vowel dataset, as we did not get interesting results from the first one and the algorithms take too much time to compute. Vowel dataset contains 11 classes and contains mixed data (both numerical and categorical features), so we had to apply one-hot encoding.

2 PCA

We implemented the covariance method of computing the PCA. In this section we will discuss the first few principal components of each dataset. Additionally, we will compare our results to the sklearn implementation of PCA [1] and Incremental PCA [2].

2.1 Vowel Dataset

In the first few principal components (PC) there are clearly dominant features. For the first we find that the most dominant feature is the ninth feature, contributing -0.784 to the PC, which is a normalized vector. For the second PC we find that the 3rd and 11th features are dominant with -0.530 and 0.580 contribution. The 9th and the 3rd features are relaated to the sound, while the 11th is related to the gender of the speaker.

2.2 Waveform Dataset

For the first PC we find that the most dominant feature is the ninth feature, contributing 0.846 to the PC. For the second PC we find that the 21st feature is dominant with 0.803 contribution. For the third PC the 27th feature is dominant with -0.821 contribution. These positions indicate relevant places in the wave that account for the most variance.

2.3 Adult Dataset

For the first PC we find that there is a group of features that contribute the most, from 16th to the 21st features. They each have an absolute value of 0.1 to 0.6. These features correspond to the one hot encoded feature of education. This indicates that education is a relevant feature.

For the second and third PC's we find that there is no clear dominant feature. Both of them have many features that have the same degree of magnitude.

2.4 Sklearn Comparison

We compared our results with the results obtained with the PCA implementation of sklearn. We see that they are essentially the same except for a global sign on some components. This is to be expected, because the normalization condition only restricts the magnitude of the eigenvectors, but one can choose any of the two directions.

On the other hand we compared our results with the result of the Incremental PCA implementation of sklearn, we obtain bigger differences between our PCA and sklearn's Incremental PCA that between our PCA and sklearn's PCA (see table 1). However, these differences are also extremely small.

	Vowel	Waveform	Adult
PCA	1.13e-13	1.18e-11	5.68e-11
Incremental PCA	1.23e-12	2.06e-11	9.23e-11

Table 1: Difference norm of absolute values of the whole transformed dataset between our PCA and sklearn's PCA and Incremental PCA.

Upon inspection, sklearn's PCA uses Singular Value Decomposition to compute the PC's, while we use the covariance method. This difference in methods accounts for the small differences between the two. The difference is so small that they can be considered equivalent.

On the other hand, the Incremental PCA is used to compute the PCA when the data set does not fit in memory. While the difference between our PCA and this PCA is bigger than to the regular PCA, the difference is also really small so they can also be considered equivalent.

3 Effects of Applying PCA before K-means

To study the effect of dimensionality reduction in the k-means algorithm, we tried different configurations for each dataset, combining different values of k with different numbers of components to mantain after applying PCA. Some conclusions can be extracted from the results.

3.1 Vowel Dataset

The metrics obtained for different values of k and p (number of components conserved after PCA) are shown in the tables below:

Table 2: Adjusted rand index

k/p	No PCA applied	6 components	4 components	2 components
k=7	0.104	0.132	0.158	0.015
k=11	0.163	0.178	0.141	0.082
k=15	0.161	0.165	0.165	0.085

Table 3: Foulkes-Mallows scores

k/p	No PCA applied	6 components	4 components	2 components
k=7	0.225	0.246	0.264	0.173
k=11	0.243	0.257	0.238	0.173
k=15	0.233	0.233	0.237	0.165

As it can be observed, both external metrics seem to improve for our real k (11) if we keep our 6 components with higher variance. However, it decreases when we go over this number. If we reduce the components to 4 or 2, the metrics are reduced and the real k is surpassed by one that we know it's not correct. However, the value of the metrics for the other k's also diminishes when surpassing p=6.

The internal metrics seem to improve their values when any reduction of the dimensionality is applied, but we can't extract any clear conclusion from here as they differ in the way that they increase for each k and metric.

Table 4: Silhouette scores

k/p	No PCA applied	6 components	4 components	2 components
k=7	0.212	0.244	0.279	0.576
k=11	0.175	0.255	0.286	0.439
k=15	0.195	0.264	0.308	0.428

Table 5: Davies-Bouldin scores

k/p	No PCA applied	6 components	4 components	2 components
k=7	1.605	1.404	1.207	0.598
k=11	1.645	1.260	1.117	0.653
k=15	1.521	1.193	1.065	0.760

Table 6: Adjusted rand index

k/p	No PCA applied	20 components	10 components	5 components
k=2	0.343	0.342	0.341	0.342
k=3	0.251	0.251	0.251	0.251
k=4	0.279	0.279	0.279	0.279

Table 7: Foulkes-Mallows scores

k/p	No PCA applied	20 components	10 components	5 components
k=2	0.618	0.618	0.617	0.618
k=3	0.501	0.502	0.502	0.502
k=4	0.497	0.497	0.496	0.496

3.2 Waveform Dataset

The external metrics don't seem to change after applying PCA.

Table 8: Silhouette scores

k/p	No PCA applied	20 components	10 components	5 components
k=2	0.155	0.206	0.282	0.362
k=3	0.129	0.180	0.266	0.368
k=4	0.096	0.138	0.216	0.321

Table 9: Davies-Bouldin scores

k/p	No PCA applied	20 components	10 components	5 components
k=2	2.145	1.758	1.370	1.096
k=3	2.249	1.797	1.331	0.983
k=4	2.671	2.120	1.530	1.093

However, the internal metrics show a bigger improvement for k=3 than for the other values (incorrect ones). It can be observed that without PCA, this value has not the best Silhouette score neither the Davies-Bouldin one. But if PCA with 5 components is applied, then it becomes the better value among the three. Moreover, this improvement is notable, making the results of these metrics much better than when using the raw data.

Taking into account the graphics shown in the next sections, the inmutability of the external metrics tells us that the clustering is working in the same way and finding the same clusters in every configuration, with slightly better results when applying PCA, specially in the internal metrics. This means that the clustering is indeed working properly and dividing the data into coherent groups, but these groups do not coincide with the real labels.

3.3 Adult Dataset

Table 10: Adjusted rand index

k/p	No PCA applied	20 components	10 components	5 components
k=2	0.175	0.032	0.175	0.175
k=3	0.041	0.186	0.052	0.130
k=4	0.003	0.097	0.012	0.007

Table 11: Foulkes-Mallows scores

k/p	No PCA applied	20 components	10 components	5 components
k=2	0.644	0.644	0.644	0.644
k=3	0.489	0.608	0.496	0.555
k=4	0.428	0.483	0.445	0.436

The internal metrics improve their values independently from the value of k. External metrics show no significant change when decreasing the number of components. The reason

Table 12: Silhouette scores

k/p	No PCA applied	20 components	10 components	5 components
k=2	0.167	0.192	0.267	0.372
k=3	0.167	0.134	0.269	0.289
k=4	0.158	0.115	0.261	0.367

Table 13: Davies-Bouldin scores

k/p	No PCA applied	20 components	10 components	5 components
k=2	2.061	2.043	1.495	1.137
k=3	2.396	2.093	1.936	1.356
k=4	2.665	2.340	1.909	1.430

is similar to that in the waveform dataset, the clusters do not vary much when applying PCA, but in contrast, the clusters here do not seem to have any sense, as it can be observed in the later figures. These results are probably due to the high number of nominal attributes of this dataset, and the difficulty to determine the distance between points. PCA results are not expected to be very good under this conditions.

4 Visualizing the Data: Comparison

One of the main reasons we may want to apply dimensionality reduction, besides from a faster computation and maybe a better performance for the distance-based clustering algorithms, is to be able to visualize the data into a lower-dimensional space (2D or 3D).

In this section we compare the different ways we can visualize the data and the effects of the dimensionality reduction. For each dataset, we show both PCA and t-SNE [3] projections (noted between parenthesis) for the original labels, (a) and (b); the labels predicted with k-Means without any dimensionality reduction, (c) and (d); and the labels predicted with k-Means after applying PCA with 5 components, (e) and (f). The reason why we have chosen 5 is because variance ratio explained by the components decreases rapidly, and also this is the overall best p parameter found in last section.

We have to point out that we are plotting and comparing with the real labels (images (a) and (b)) but in an unsupervised learning environment we usually don't have this information. Due to that, we are extracting conclusions that we couldn't otherwise.

Each color represents a cluster, a group of similar samples. We must recall that colours themselves don't indicate a specific class or meaning, but rather a group (e.g. not 'dog' and 'cat', but 'group 1' and 'group 2'). In other words, we can have different permutations of colours and it will still be the same clusterization.

4.1 Vowel Dataset

If we take a look at Figure 1, we can clearly see that samples are distributed into two dense main clusters, for both PCA and t-SNE visualizations. There is some pattern that polarizes a lot the samples into two subgroups.

However, if we take a closer look, we don't see a clear separation of each real class, as we have samples of the same real label in both main structures (see (a) and (b)).

When we apply PCA dimensionality reduction, we get slightly different results. For example, comparing the right cluster-structure of images (c) and (e), we can see that in the first one there are only 5 clusters (notice some greener points), and in the latter there are clearly 6 clusters, resulting in a different grouping. The same happens with t-SNE visualizations (d) and (f).

Even after PCA, K-Means is not able to classify the samples properly in this dataset, at least comparing to the real ones (which we usually don't have access in an unsupervised learning environment).

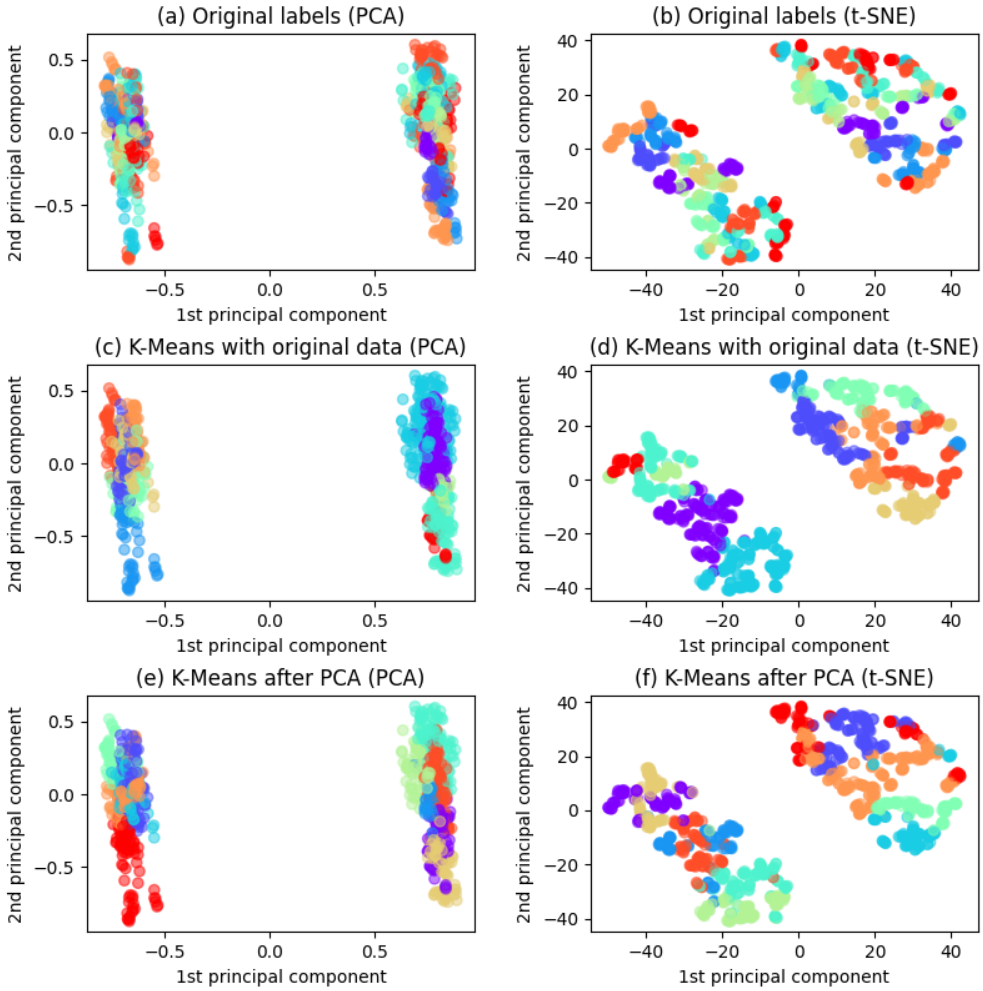

Figure 1: Comparison for vowels dataset applying dimensionality reduction

4.2 Waveform Dataset

As we can see in Figure 2, the original samples are quite well distinguished after projecting them in the two main components, especially for PCA (a), where each class correspond roughly to a an edge of a triangle and the intersections between classes, where clusters mix, at the vertexes. The same pattern is observed when projecting with t-SNE (b), though not so clearly.

When we apply the K-Means (see Figure 2 (c) and (d)), we can see that the centroids

are placed in the vertexes of the aforementioned triangular-structure, resulting in a bad performance. This is due to the fact that the vertexes are the 3 farthest points of the space, and concentrate most of the samples nearby. Therefore, we can conclude that K-Means is not a suitable algorithm to get the pattern of real labels that exists in the data.

If we execute the k-Means algorithm with the PCA dimensionality reduction as shown in (e) and (f), it appears that we get the same results (with color permutations) than (c) and (d). That makes sense, as we get the same centroids in the vertexes and thus the same problem explained before.

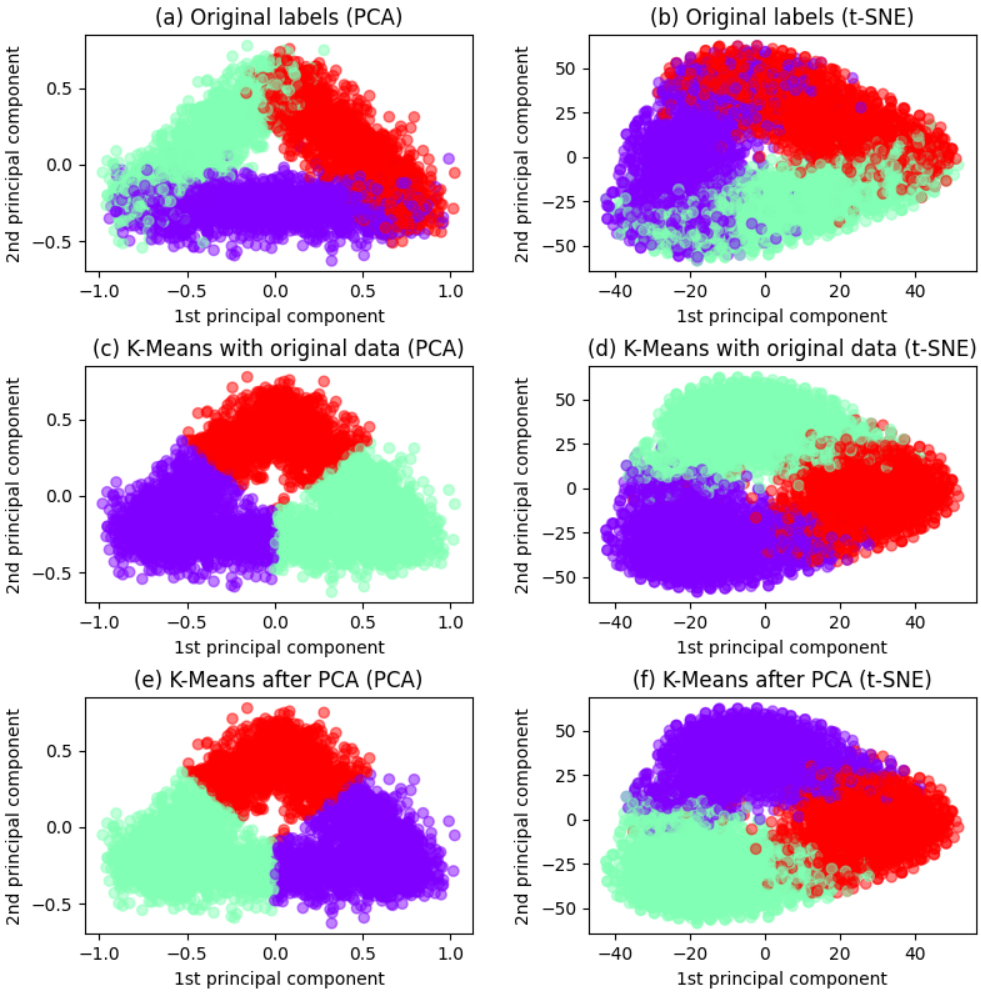

Figure 2: Comparison for waveform dataset applying dimensionality reduction

4.3 Adult Dataset

Looking at the Figure 3, we can see that visualizing the data with PCA and t-SNE results in two very different plots. For the former, we can see 5 well-defined high-density clusters (see (a), (c) and (e)), while the latter shows a rather homogeneous distributed data, like in (b).

We may conclude with quite certainty, if we look only at the PCA visualizations, that k=5 is the most suitable, but we will be somehow mistaken, as we can see that this is not true when we look at the t-SNE, where distances between samples are more preserved than PCA.

What we can see is that, neither PCA nor t-SNE visualizations in 2D makes us think that there are indeed 2 clusters (the real '$\leq$50k' and '>50k'). It is true, though, that when we plot the original labels with t-SNE (b), there is some predominance of the red cluster on the top-left and right corner and the purple cluster on the bottom, but it is not significant.

The results appear to be the same when applying PCA before K-Means, as images (c) and (e), and (d) and (f) look identical to the naked eye.

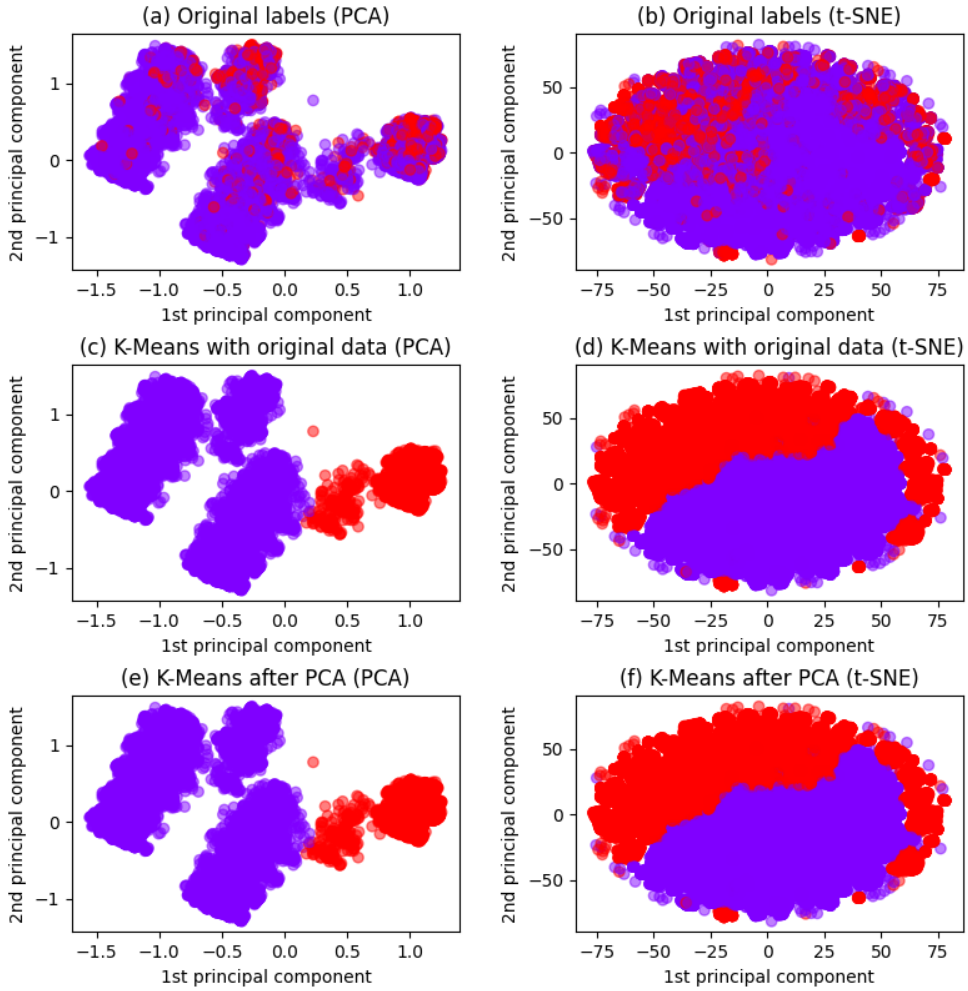

Figure 3: Comparison for adult dataset applying dimensionality reduction

5 Conclusions

After applying PCA, we were able to know more underlying information of the datasets. We have found out that waveform is well explained with the two first principal components of PCA but different clusters lie on the edges of a triangular-structure, rather than the vertexes. K-Means (at least without more tuning) is not good a good algorithm to obtain the same clusters as the real ones, but without knowing the real labels, the result makes a lot of sense. Our conclusion is that now we know a good representation for this dataset and therefore we

know what are we looking for. One approach we could try is to set the center of each edge as the centroid of a cluster and then associate each sample to the closest centroid (to its corresponding edge).

The adult dataset on the other hand, shows barely any improvement, probably due to the high number of nominal attributes. Finally the vowel dataset responds positively after applying PCA, as the external metrics get better values for the correct value of k after applying it and giving us real information about the distribution of the data.

The number of components were chosen trying to keep an overall high variance, so we don't loose much information. As said before, the effect of dimensionality reduction is different among the three datasets. Having a positive impact in the vowel one (but not being able to reduce more than 6 components) and having not so much impact in the other two, although for different reasons.

To conclude, we have learnt to visualize data that we weren't able before. This can be a key step in order to understand better the data we are working on and therefore take future decisions based on that. Furthermore, we have seen an alternative for when we have data with a huge number of dimensions. Now, we can apply this approach before executing an algorithm and see if the results improve.

6 Execution Guide

In order to test the different visualizations you can use the Python main.py file by entering the following on the command line:

```
python main.py
```

Once the main is executed, you will be asked to specify the dataset and the visualization you want to test. The visualization selected will be executed with the best parameters values we found during our research.

The plots will be appearing as they are computed. It might take a while, especially for adult dataset.

References

[1] *sklearn.decomposition.PCA*, accessed November 16, 2020. `https://scikit-learn.org/`

stable/modules/generated/sklearn.decomposition.PCA.html.

[2] *sklearn.decomposition.IncrementalPCA*, accessed November 16, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.IncrementalPCA.html.

[3] *sklearn.manifold.TSNE*, accessed November 16, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.