Homework 1 – Machine Learning (2018/19)

Angelo Laudani s253177

PCA

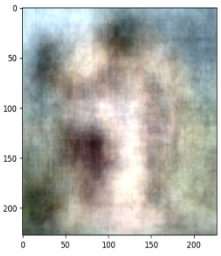
Introduction

The main objective of a PCA analysis is to identify patterns in data by detecting correlation between variables. This correlation may suggest that reducing the data dimensionality is useful. We can summarize PCA as the attempt of finding the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace. In this way we will still retain most of the information.

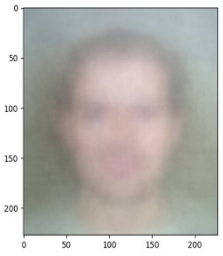
1 – PCA on Image

On a Dataset which counts 1087 images, distributed in the four classes of "dog", "guitar", "house" and "person" we can apply PCA in the attempt of reducing the dimensionality of our data. Each image has 154587 features, using PCA we can extract a different number of Principal Components and project a same image in the subspace created by those components.

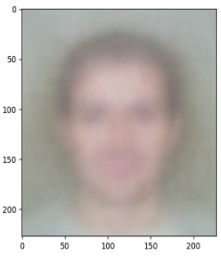
Original Picture



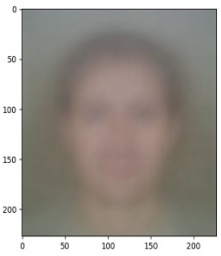
First 60 PC



First 6 PC



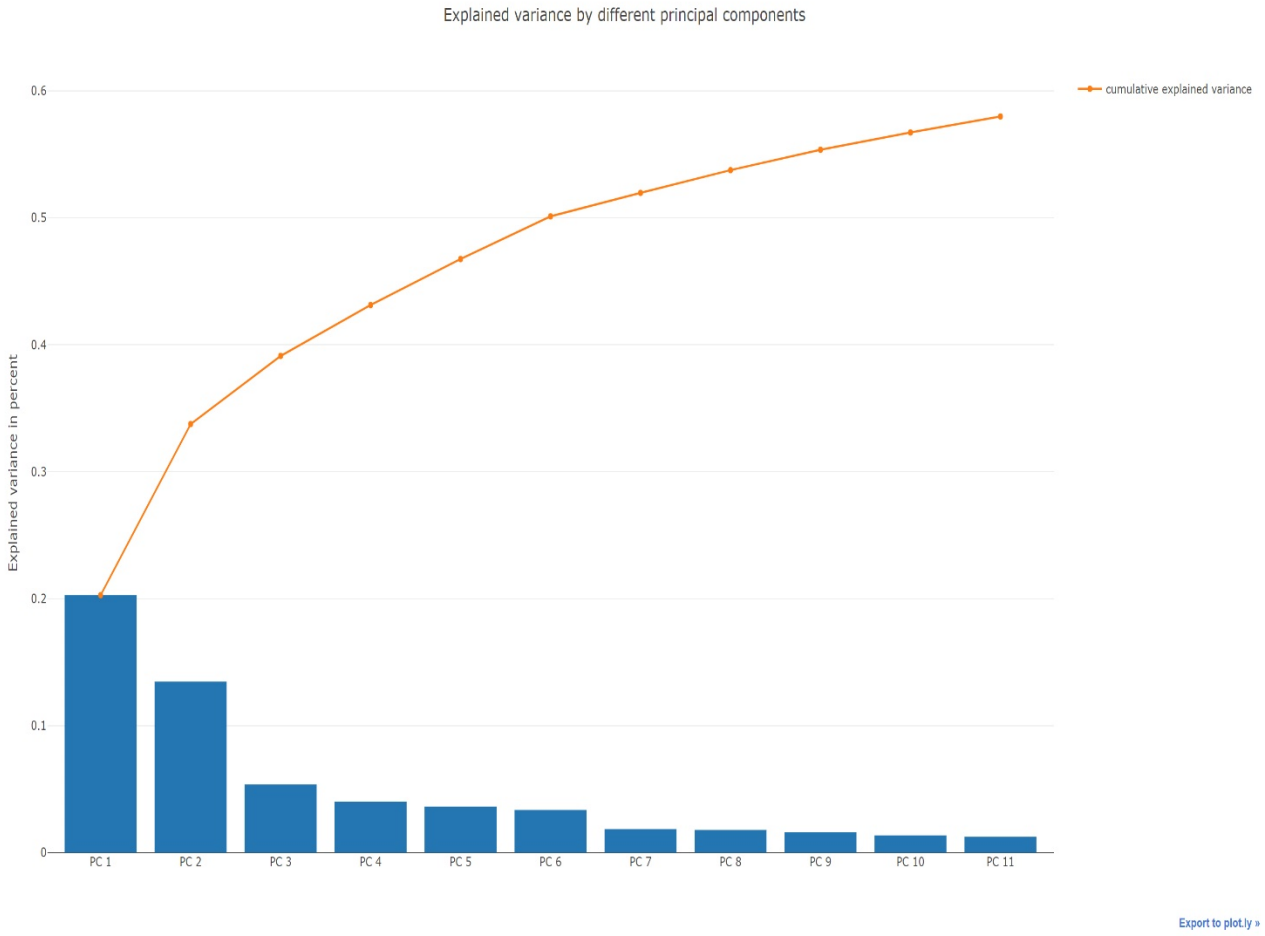
First 2 PC



Last 6 PC

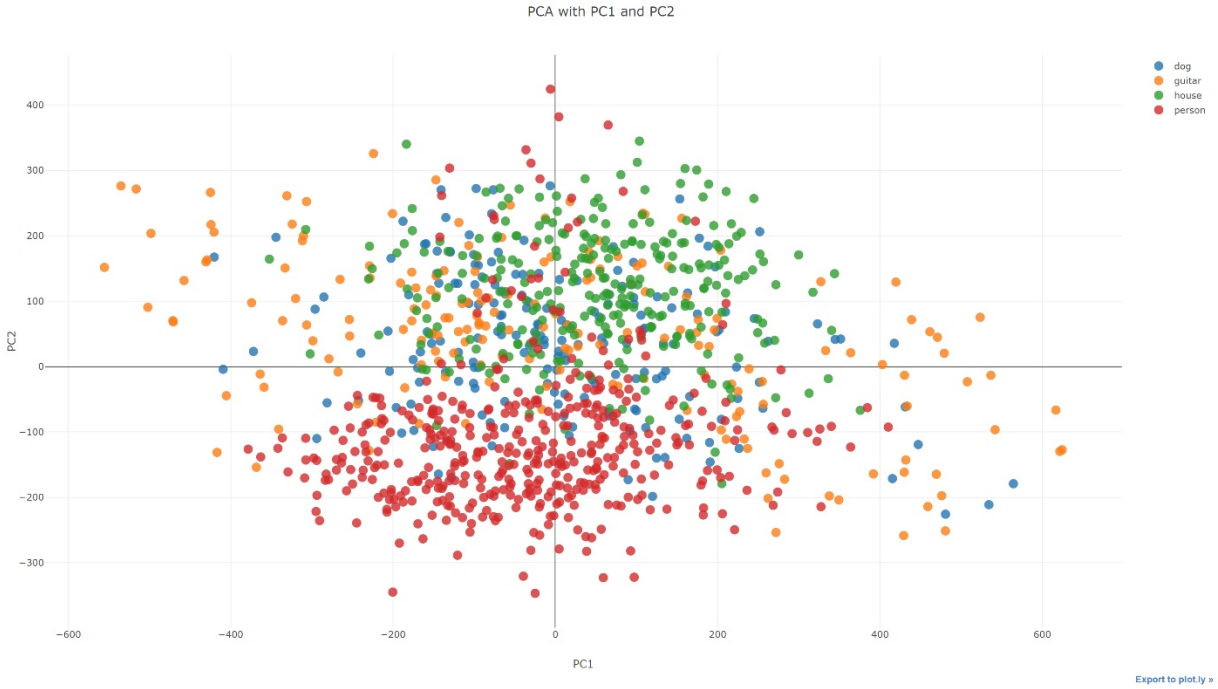
It is noticeable how the original image is more and more blurred out, till loosing completely its original information, “meaning”, when projected only using the first 6, 2 or last 6 components.

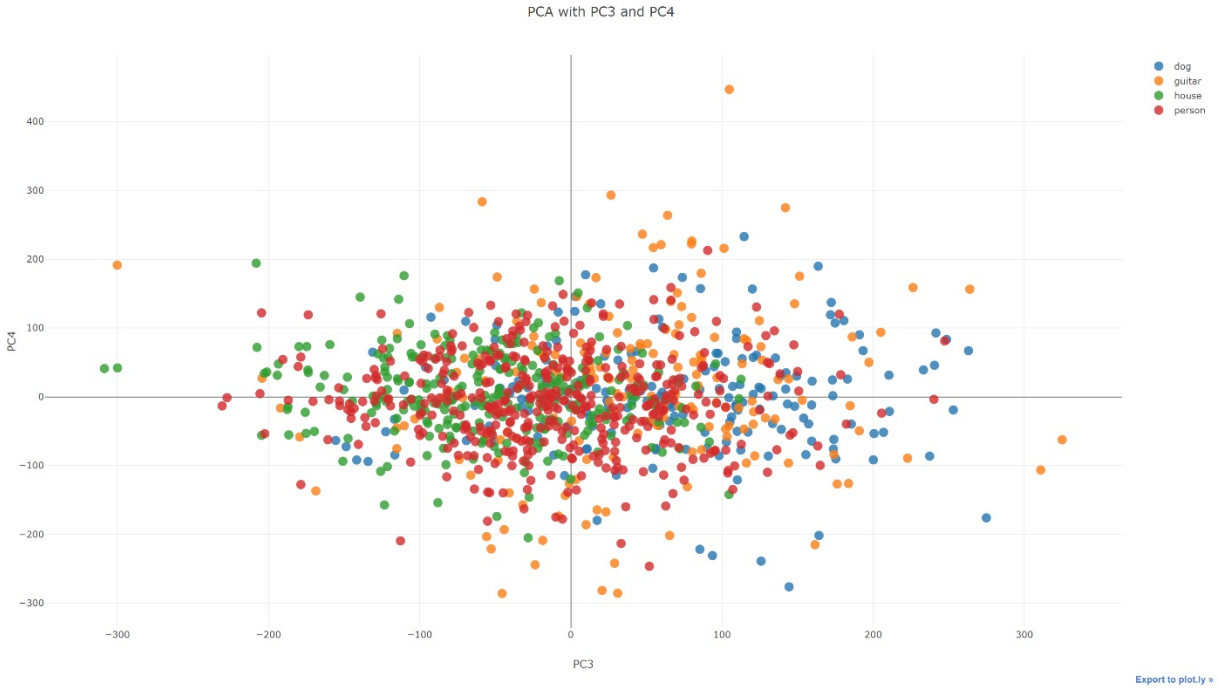
This can be easily explained by visualizing how much variance each principal component retains. Just in the two PC the 33.75% of the variance, with the next components adding almost no significant value to the total. Only with the first 60PC we retain almost 77% of the variance, which makes the image still somehow recognizable. The last 6PC explain only the 0.007% of the variance.

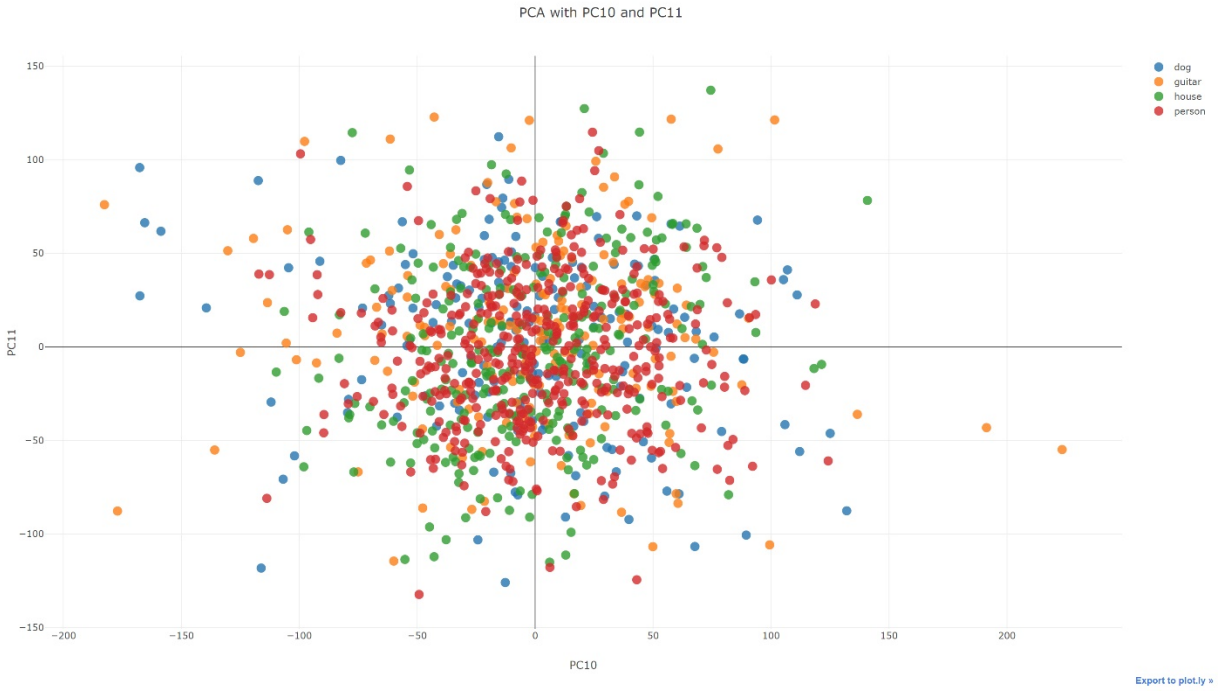


2 – PCA Projection

By projecting the standardized data on a scatterplot along their principal components we can visually explain why the dog image, when projected with less and less components, resembles the face of a person. In the following plots, the “person” images are more concentrated towards the first principal components, retaining the most significant part of information of the dataset in terms of variance.


PCA with PC1 and PC2
dog
guitar
house
person
400
300
200
100
0
−100
−200
−300
PC2
−600
−400
−200
0
200
400
600
PC1
Export to plot.ly »


PCA with PC3 and PC4
dog
guitar
house
person
400
300
200
100
0
−100
−200
−300
PC4
−300
−200
−100
0
100
200
300
PC3
Export to plot.ly »



3 – Naive Bayes Classification

The naïve Bayes Classifier is a simple probabilistic classifier, with a strong independence assumption between the features. It is a supervised learning model, Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. By dividing the dataset in train and test set, we can perform a classification on the images projected with different Principal Components:

- **Unmodified Dataset:** 94 mislabelled points out of 435, accuracy of 78%
- **First 2 Principal Components:** 161 mislabelled points out of 435, accuracy of 62%
- **3rd and 4th Principal Components:** 237 mislabelled points out of 435, accuracy of 45%