Aristidis Chronarakis 389

13/06/18

T6- Machine Learning

Project #2

Exercise 1A

The purpose of this exercise is to classify face images. The whole implementation was made in Python using the Machine Learning library (scikit-learn). First of all, the whole image dataset with the faces is divided on two equally datasets, one for the training and the other for the testing. The training dataset is used as input on a PCA model with which we reduce the dimension of the input using different values for the level of reduction. After, we make also the needed reduction for the dimensions of the test dataset, based on the previous reduction on the training dataset. The dimension-reduced training dataset is used for the "fit" on a k-NN model, and then the dimenison-reduced test dataset is used on this model, in order to find the k nearest neighbors for each element of this set. On the k-NN model are used both the euclidean and the cosine as distance metrics and different values for the k.

Results Overview

The results from the k-NN classifier show us that we can get also good results using the dimension-reduced dataset from the PCA. The level of how good are the results is affected by the dimension-reduction variable (M). Using for the dimension-reduction variable, a value which lead us to have a relative "slight" dimension reduction help us to "drop" the noise from our images and thus get better results on the k-NN. On the other hand using an M which lead to a small dimension for our images lead us to "drop" from our images valuable information and thus lead us to get worse results on the k-NN classifier. Final, it seems that we get better results using the cosine related distance metric.

Exercise 1B

The purpose of this exercise is the clustering of face images. The whole implementation was made in Python with using the Machine Learning library (scikit-learn). First of all, the whole image dataset is used as input on a PCA model with which we reduce the dimension of the input using different values for the level of reduction. Then we construct two model for the clustering, one based on k-Means and the other based on Gaussian mixture models. Both of the models takes as input the

number of needed clusters (10 on our case) and the dimenison-reduced dataset and give us for each element of the set a label which indicates the cluster in which belongs. The evaluation of the clustering on both cases uses the purity as metric.

Results Overview

As on the previous exercise the results are also affected from the the PCA variable (M) for the same reason. About the purity of the 10 clusters that our purity measurement implementation give us said that k-means gives slight better results than using the option of Gaussian Mixture Models.

The full code for the both implementations can be found on Github

https://github.com/chron56/ML-microprojects