University of Macedonia

Department of Applied Informatics

7Thesemester | Academic Year: 2022 – 2023

Machine Learning

3theTask (Unsupervised Learning – Clustering)

In this work you will develop combined projection models in a new space (PCA) / clustering techniques, which you will use on the dataset data **fashion-mnist**. They are provided at the following link:https://github.com/zalandoresearch/fashion-mnist .

For this exercise you will deliver a python file in which you will present the differences in clustering results when using raw data and when using data, when projecting into a new space, through principal component analysis.

The code you develop must do the following:

1. It will load fashion-mnist data.
2. It will separate the data into three sets: train, validation & test data.
3. It will perform principal component analysis (PCA) on the train set data. Caution: during PCA choose any proportion of variance you wish to preserve, as long as it is in the interval [85%, 100%].
4. It will use the existing transformation (derived from the train data) and apply it to the validation data. Now use the inverse transform to bring them back to the original space.
5. Print random images from the validation set (at least one from each class) as well as the reconstructed ones as they came from the inverse transformation.
6. It will use the existing transformation (derived from the train data) and apply it to the test data.
7. It will use three different clustering techniques (whichever you want) to create subgroups, on:
    a. No test data processing
    b. PCA transformed test data
8. It will calculate the performance indicators presented in the workshop and one more of your choice.
9. It will show indicative clustering results for random images.

Using the results of the algorithm, and graphs that you will make in excel, you will write a report in which you will present your conclusions, you will make comparative evaluations.

The report should answer the following questions:

When do you get better clustering performance scores? when you have raw data or when you have pca projected data?

Based on the clustering performance scores, what is the ideal number of clusters for your dataset? Do you get the same result when you have raw and pca projected data?

Which clustering technique gives you the best results? In the end, which is the best combination of PCA or raw data / clustering approach?

Present the clustering results of the best possible combination, based on the clustering metrics. Practically speaking, are the results acceptable? If not, what could you improve/do differently?

Instructions:

A. Assignments are in groups of up to four (4) people. Each person can submit work to only one group at a time.

B. Assignments should be uploaded to eClass in a zip (not rar) file by the deadline. No extension will be granted.**Caution**: Each group assignment will be submitted by only one group member (you choose who/who).

C. Each assignment must be accompanied by:

- One and only one file.py will contain the answers to the queries
- One**report**inpdf with the following information: o
  - Cover: 1 page, includes the details of the students in the group, course name, date, department and other relevant details.
  - o Summary table of contents, images, and other graphics that cite in the report.
  - o Introduction section: 1 page, describe the problem (*without* copying exactly the speech of the exercise)
  - o Methods applied: from 2 to 10 pages, describe the methods you used and list the relevant results. Make sure it's clear which question you're referring to.
  - o Conclusions: 1 page, based on the results what do you recommend, which model performs better, what could be done to further improve the performance.
  - o The report will contain graphical representations of all kinds and tables of evaluation of the results that must be accompanied (each) by at least one paragraph with commentary.

Make sure that:

- The code must necessarily be accompanied by appropriate comments.
- An editorial and spelling check has been done on the report you will submit.
- The sentences should be understandable and short in length.
- Images to***do not***have arisen fromprint screen. If the program creates an image, save it normally (jpg or png) before using it.
- The graphs should include names on the axes and a legend. Purpose is to understand what it shows, at a glance.
- If something is not specified, you have the right to make any implementation that suits you. Make sure you can explain exactly what you did in the code.
- The libraries you will use *must* be able to be installed via it pip.

- The code *must* run onGoogle Colab.

<u>Delivery Due Date:</u> **January 23, 2023**. No extension will be granted.