Homework
Principal Components Analysis

Francesco Della Santa [*]

Computational Linear Algebra for Large Scale Problems

Politecnico di Torino, A.Y. 2023/2024

Contents

[*] Dipartimento di Scienze Matematiche, Politecnico di Torino, Turin, Italy

1 How the Homework Must be Prepared and Uploaded

Here, you find all the necessary information for preparing the homework:

- The homework can be done alone or in teams of two people. Groups with three or more people are not allowed;
- All the files of the homework must be compressed into one file (e.g., *.tgz* or *.zip* file) named as *Studentsurname_HWpca*. For example: *Dellasanta_HWpca.zip*. In case of a group of two students, write both the surnames in alphabetical order: *StudentsurnameoneStudentsurnametwo_HWpca*. For example: *BerroneDellasanta_HWpca.zip*.
- The compressed file for this homework must contain *exactly* two files:
  1. A jupyter-notebook *Studentsurname.ipynb* (or *StudentsurnameoneStudentsurnametwo-.ipynb* for groups) that is your report for the homework (see Section 2 of this document for more details);

     Attention: the only python modules and packages allowed are the ones used in the laboratories; i.e., the ones specified in the file *modandpacks_cla4lsp.txt* (uploaded on the web page of the course). The teacher must be able to run your code using the environment created for the laboratories!
  2. A PDF version of the jupyter-notebook above, named *Studentsurname.pdf* (or *StudentsurnameoneStudentsurnametwo.ipynb* for groups).

     Very important: this file is the official and "printed" version of your report! Then, be sure that the plots are well represented and all the comments refer to visible codes/pictures/tables etc.
- The compressed file must be uploaded within the beginning of the exam session (ask to professor Berrone for more details);

1.1 How to Prepare the Report

The report of this homework is a document where the student (or the group of students) writes and comments the addressed problem, the performed procedures, and the results of the exercises adding the codes. For this reason, the report must be written using a jupyter-notebook, since it allows to "merge" the textual descriptions and argumentations with python codes.

General suggestions for writing the report:

- Split the report in Sections corresponding to the steps/exercises of Section 4;
- **Always** justify your choices! For example, explain why you decide to use some particular type of encoding for categorical data instead of another one.
- Unless specifically required by the exercise, tables and plots are optional. Nonetheless, they are welcome *if they help to read the report and understand your analyses.*

2 The Homework Files

Here, we introduce the dataset used for this homework and we list the exercises and the required analyses.

In particular, all the files required for this homework (included this document) are available in the folder *DellaSanta/HWpca* on the web page of the course.

2.1 Dataset Descritpion

For this homework, we consider a dataset containing customers' information with respect to their spending habits and their family status. For a detailed description of the dataset columns, see the Appendix A.

2.2 The Files

Inside the homework's folder, the students can find one file: *cla4lsp_customers.csv*. This is the file containing all the data necessary to the homework, consisting in a dataset of customers described by a set of features useful for profiling them.

Some characteristics to be known concerning the dataset are:

- The data file consists of 2 240 rows and 29 columns;
- The data file contains missing values;
- the data file contains both numeric and categorical data.

3 Simulation of a Real-World Problem

Given the dataset, the students have to use the Principal Component Analysis (PCA) to reduce the dimensionality of the problem and, then, have to identify meaningful clusters of customers (if existing) using the k-Means algorithm.

Remark 3.1 (Simulating the typical situation in a company)**.** With this homework we want to simulate (for example) the situation where a company wants to analyze the market and identify meaningful profiles of customers. Typically, all the information gathered from the history of sold/selling products must be summarized with few and easily-comprehensible concepts, in order to help the managers with future decisions. Then, in these situations, very often the number m of chosen Principal Components (PCs) is very low, giving more importance to the dimensionality reduction then the preservation of the information.

Given the remark above (Remark 3.1), let us assume that the dataset is the result of a data-collecting procedure performed by the company you are working for. Then, your manager asks you to:

1. Summarize the available data in order to make them easier to be interpreted. In particular, summarize the information in the with at most $m = 5$ features but (if possible) with at least 33% of preserved information.
2. Identify "customer profiles" according to the summarized features. Specifically, identify a minimum of 3 up to 10 profiles.

Remark 3.2 (Learn to look for tools in the documentations)**.** Some exercises may require the use of functions and/or tools that we have not seen during the laboratories; e.g., how to generate a sequence of n random integers between n_1 and n_2, $n_1 < n_2$, apply a function to an entire column of a DataFrame, work with date-time data, etc.. In these cases, part of the exercise consists in looking for a solution among the built-in tools and/or among the tools of the installed third-party packages.

4 The Homework Exercises, in Practice

Translating the instructions of the manager in Section 3 into an academic exam, the specific exercises of the homework are the following:

0. **Preparation (Setting the Random State):** before starting with the exercises, initialize a random state variable rs equal to the minimum of the ID student numbers of the group members. For example, if the group consist of two students with ID numbers 12345 and 67890, the value is $rs = \min\{12345, 67890\} = 12345$. If the group consists of only one student, use the ID student number as random state.

   The random state rs **must** be used to set the *numpy random seed* at the beginning of the code and in every python functions you call during the exercises (if a random procedure is used).

   Specifically, before all the other operations in your code, write the command

   ```
   numpy.random.seed(rs)  .
   ```

   Concerning the functions characterized by a random state, you must specify it. For example:

   ```
   km = KMeans(random_state=rs)  .
   ```

1. **Exercise 1 (Loading and Preparing the Data):** load the file *cla4lsp_customers.csv* as a pandas DataFrame (DF). Then:

   1.1. store in the variable *df_tot* the df obtained from the csv file.

   1.2. create a sub-DFs *workdf*, extracted from *df_tot*, such that it contains 2/3 of the original dataframe's rows (randomly sampled);

   1.3. let us denote (see Appendix A) with:

   - *labels*: the columns *NumDealsPurchases*, *AcceptedCmp1*, . . ., *AcceptedCmp5*, *Response*, *Complain*, *Recency*;
   - *features*: all the other ones, except for *ID*, *Z_CostComtact*, and *Z_revenue* (i.e., discard these columns).

   1.4. Remove *randomly* from *workdf* one feature column among the spending habits or the purchasing habits ; i.e., remove one column among this list: *MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumWebPurchases, NumCatalogPurchases, NumStorePurchases.*

   1.5. Clean the dataset *workdf* from missing values in the *feature columns* (if needed).

2. **Exercise 2 (Encoding of Categorical Data):** Analyze and prepare *workdf* for the PCA. In particular, apply a proper encoding of the categorical features. Once applied the encoding, store into a variable *Xworkdf* the sub-DF obtained from *workdf* selecting the feature columns (updated to the new encoding).

3. **Exercise 3 (Preprocessing and full-PCA):** Preprocess the data, before applying the PCA:

   - create two DFs *Xworkdf_std* and *Xworksf_mm*, created using a StandardScaler and a MinMaxScaler (min = 0, max = 1), respectively, applied to *Xworkdf.*
   - analyze and comment a comparison of the variances of *Xworkdf* with the variances of *Xworkdf_std* and *Xworkdf_mm.* What do you observe from this analysis?
   - Apply the "full" PCA[1] to the DFs *Xworkdf*, *Xworkdf_std*, and *Xworkdf_mm* and plot the curve of the cumulative explained variance. Looking at the results, improve the analysis and comments made at the previous step.

[1] i.e., no dimensionality reduction.

4. **Exercise 4 (Dimensionality Reduction and Interpretation of the PCs):**

   Apply the PCA to both[2] *Xworkdf_std* and *Xworkdf_mm*, selecting m PCs such that

   $$m = \min\{m', 5\}\,, \tag{1}$$

   where m' is the minimum number of PCs that explains 33% of the total variance. Plot the barplots of percentage of explained variance, with respect to the PCs.

   Then:

   - Given the PCs of *Xworkdf_std* and *Xworkdf_mm*, give them an interpretation and, therefore, a *name*. Tables and/or plots are welcome;
   - After the interpretation, for both the DFs represent a score graph with respect to the first ℓ PCs, where $\ell = 2$ if $m = 2$ and $\ell = 3$ if $m \geqslant 3$. In particular, write the names of the PCs (chosen by you) on the axes of the plots;
   - **Optional:** make more than one score graph, coloring the dots with respect to any *label* you consider meaningful.
   - analyze and comment the results.

5. **Exercise 5 (k-Means):** Run the k-Means algorithm on the two DFs, with respect to the "PC-space". Select the best value of $k \in \{3, \ldots, 10\} \subset \mathbb{N}$ using the *silhouette coefficient*.

   Optional: "play" with the other parameters of the *KMeans* class of scikit-learn.

6. **Exercise 6 (Clusters and Centroid Interpretation and Visualization):** Comment the centroids of the best clustering for both the DFs. In particular, give to each centroid a *name* or a *meaningful brief description* that characterizes the average customer in the cluster represented by the centroid.

   Moreover, plot the score graph of exercise 4 together with the centroids. In particular, show the different clusters using different colors and/or markers for the dots.

7. **Exercise 7 - Clusters and Centroids Evaluation:** For both the DFs, perform an internal and an external evaluation of the clusterings obtained. In particular:

   - Measure the silhouette scores of the clusters (*internal evaluation*);
   - perform an *external evaluation* of the clusters analyzing and plotting the distribution of the *labels* (that you retain more interesting) inside each cluster.

     Attention: before the external evaluation, check how the labels selected by you are distributed into the dataset.
   - Comment the results. Compare the results obtained from *Xworkdf_std* and *Xworkdf_mm* and comment them.

[2] two different PCA objects for the two DFs, obviously; analogously, two different choices of m.

A Dataset's Details

A.1 Labels

1. NumDealsPurchases: Number of purchases made with a discount
2. AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
3. AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
4. AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
5. AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
6. AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
7. Response: 1 if customer accepted the offer in the last campaign, 0 otherwise
8. Complain: 1 if the customer complained in the last 2 years, 0 otherwise
9. Recency: Number of days since customer's last purchase

A.2 Features

1. Year_Birth: Customer's birth year
2. Education: Customer's education level
3. Marital_Status: Customer's marital status
4. Income: Customer's yearly household income
5. Kidhome: Number of children in customer's household
6. Teenhome: Number of teenagers in customer's household
7. Dt_Customer: Date of customer's enrollment with the company
8. MntWines: Amount spent on wine in last 2 years
9. MntFruits: Amount spent on fruits in last 2 years
10. MntMeatProducts: Amount spent on meat in last 2 years
11. MntFishProducts: Amount spent on fish in last 2 years
12. MntSweetProducts: Amount spent on sweets in last 2 years
13. MntGoldProds: Amount spent on gold in last 2 years
14. NumWebPurchases: Number of purchases made through the company's website
15. NumCatalogPurchases: Number of purchases made using a catalogue
16. NumStorePurchases: Number of purchases made directly in stores
17. NumWebVisitsMonth: Number of visits to company's website in the last month

A.3 Other Columns

1. ID: Customer's unique identifier
2. Z_CostContact: *not specified*
3. Z_Revenue: *not specified*