Homework
Principal Components Analysis

Computational Linear Algebra for Large Scale Problems

Politecnico di Torino

A.Y. 2022/2023

Contents

1 How the Homework Must be Prepared and Uploaded

Here, all the necessary information to prepare the homework for the exam:

- The homework can be done alone or in group. Groups with three or more people are not allowed;
- The compressed file must be uploaded within 5 days from the *oral exam* date;
- 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 3 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.

2 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 3.3.2;
- **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.*

3 The Homework

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.

3.1 Dataset Descritpion

The dataset used in this homework is a preprocessed and cleansed[1] version of a dataset extracted from bikez.com on April 30^{th} 2022, using a custom scraper in order to enrich an existing used motorcycle dataset for a hackathon competition.

For a detailed description of the dataset columns, see the Appendix A.

3.2 The Files

Inside the homework's folder, the students can find one file: *cla4lsp_bikez_curated.csv*. This is the file containing all the data necessary to the homework, consisting in a dataset of motorcycles described by many technical features.

Some characteristics to be known concerning the dataset are:

- The data file consists of 38 472 rows and 27 columns;
- The data file contains missing values;
- the data file contains both numeric and categorical data.

3.3 Exercises

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 motorcycles (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 products. 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.*

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, $n_1 < n_2$, splitting a string with respect to a substring, or apply a function to an entire column of a DataFrame. In this cases, part of the exercise consists in looking for a solution among the functions/tools of the installed third-party packages.*

3.3.1 Simulation of a Real-World Problem

Considered 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 35% of preserved information.
2. Identify "motorcycle profiles" according to the summarized features. Specifically, identify a minimum of 3 up to 10 profiles.

[1] indeed, the original "raw" dataset is great for practicing data cleaning skills.

3.3.2 The Homework Exercises, in Practice

Translating the instructions of the manager in Section 3.3.1 into an academic exam, the specific exercises of the homework are the following:

1. **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)
   ```

   .

2. **Exercise 1 (Loading and Preparing the Data):** load the file *cla4lsp_bikez_curated.csv* as a pandas DataFrame (DF). Then:

   (a) store in the variable *df_tot* the df obtained from the csv file.

   (b) select a random integer r among $0, 1, 2$, and create a sub-DFs *workdf*, extracted from *df_tot*, such that it contains only data corresponding to years with reminder r, if divided by three;

   (c) let us denote (see Appendix A) with:

   - *labels*: the columns *Brand*, *Model*, *Year*, *Category*, *Rating*;
   - *features*: all the other ones.

   Then, remove *randomly* from *workdf* two columns selected among the features: *Front/Rear breaks*, *Front/Rear tire*, *Front/Rear suspension*.

   (d) Clean the dataset *workdf* from missing values in the *feature columns* (if needed).
   Hint: for categorical data, sometimes the missing values can be interpreted as another category. This is useful in particular when, e.g., the rows with missing values are too much to be removed.

3. **Exercise 2 (Encoding of Categorical Data):** Analyze and prepare *workdf* for the PCA. In particular:

   - apply a proper encoding of the categorical data.
     Hint: Some categorical features in the dataset are particular. These features consist in lists of characteristics represented by a string and separated by the characters *'. '*. Then, for a proper encoding, it is suggested to do as in the example of Table 1.
   - once applied the encoding, store into a variable *Xworkdf* the sub-DF obtained from *workdf* selecting the feature columns (updated to the new encoding).

4. **Exercise 3 (Preprocessing and 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*, with a special focus on the non-categorical features. What do you observe and what can we infer from this analysis?

FeatureA
charα. charβ
charα. charβ. charγ
charγ
$\vdots$

$\rightsquigarrow$

FeatureA-α	FeatureA-β	FeatureA-γ
1	1	0
1	1	1
0	0	1
$\vdots$	$\vdots$	$\vdots$

Table 1: Special encoding example

- Apply the "full" PCA[2] 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.

5. **Exercise 4 (Dimensionality Reduction and Interpretation of the PCs):**
   Apply the PCA to both[3] *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 35% 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 $m \geq 3$. 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.

6. **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.

7. **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* that describes the average motorcycle in the cluster represented by it.

   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.

8. **Exercise 7 - Optional (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 (e.g., check if the motorcycle categories are balanced in the dataset).

[2] i.e., no dimensionality reduction.
[3] two different PCA objects for the two DFs, obviously; analogously, two different choices of m.

- Comment the results. Compare the results obtained from *Xworkdf_std* and *Xworkdf_mm* and comment them.

A Dataset's Details

A.1 Labels

1. Brand - brand name of the motorcycle
2. Model - model name of the motorcycle
3. Year - year the motorcycle was built
4. Category - sub-class the motorcycle belongs to in the market (style of motorcycle)
5. Rating - review average out of 5 stars

A.2 Features

1. Displacement (ccm) - engine size of the motorcycle in cubic centimeters (ccm)
2. Power (hp) - max power output in horsepower (hp) along with peak power rpm
3. Torque (Nm) - max torque in newton-meters (Nm) along with peak torque rpm
4. Engine cylinder - number of cylinders in the engine as well as configuration
5. Engine stroke - number of stages to complete one power stroke of the engine
6. Gearbox - number of gears in transmission
7. Bore (mm) - diameter of each cylinder in millimeters (mm)
8. Stroke (mm) - distance within the cylinder a piston travels in millimeters (mm)
9. Transmission type - type of transmission of the motorcycle
10. Front brakes - type of front brake
11. Rear brakes - type of rear brake
12. Front tire - front tire size
13. Rear tire - rear tire size
14. Front suspension - front suspension type and configuration
15. Rear suspension - rear suspension type and configuration
16. Dry weight (kg) - weight of the motorcycle, without any fluids, in kilograms (kg)
17. Wheelbase (mm) - distance between the points where the front and rear wheels touch the ground in millimeters (mm)
18. Fuel capacity (lts) - maximum capacity of fuel tank in liters (lts)
19. Fuel system - fuel delivery system into engine
20. Fuel control - valve configuration fo the engine
21. Seat height (mm) - height from bottom of seat to the ground in millimeters (mm)
22. Cooling system - engine cooling system