Homework of Principal Components Analysis

Computational Linear Algebra for Large Scale Problems

Politecnico di Torino

A.Y. 2019/2020

Abstract

In this homework the students will apply to a dataset of football players' characteristics the PCA algorithms written during the laboratories of the course. Then, they will give an interpretation of the results obtained.

1 Introduction

In this homework, we consider a dataset characterizing the skills of (thousands of) football players; this skills are the official ones selected and quantified by FIFA for the year 2019 and are used to compute an overall evaluation of the athletes (but also videogame statistics etc.).

1.1 Brief Description of the Dataset

Each row of the dataset represents a football player, characterized by the following attributes (one for each column):

1. **Index:** row index (integer starting from 0);
2. **ID:** Identity number of the athlete (positive integer);
3. **Overall:** Overall evaluation of the player (integer in $[0, 99]$);
4. **Position:** Abbreviation of the main field-position of the football player;
5. **GeneralPosition:** Abbreviation of the *general* main field-position of the football player. More precisely: GK=*Goal Keeper*, DF=*Defender*, MF=*Midfielder*; FW=*Forward*;

$\geq$ 6. **Skills:** quantification (float in $[0, 99]$) of 34 football skills characterizing the athlete (see Figure 1).

1.2 How to prepare and present the homework:

This homework can be done singularly or, at most, in groups of two. The programming language can be either *python* or *matlab*. The presentation of the homework must consist in the following items:

- a report (*.pdf* file) with a description and explanation of the operations done and results obtained for each part of the homework;
- an archive (e.g. a *.zip* file) containing the *.py* or *.m* files used for the homework execution and an *INSTRUCTIONS.txt* file explaining how to run the scripts and replicate the results illustrated in the report;
- an archive containing the *.csv* files that you have obtained running your scripts.

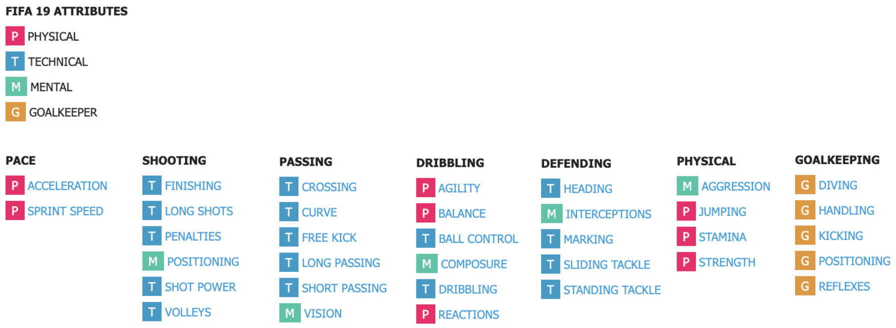

Figure 1: Skills characterizing the football players, divided in categories (with respect to both typology and utility).

These files must be uploaded on the web page of the course or on the *exercise* platform. More details and information will be provided by the professor in future lessons.

2 Homework: PCA applied to the Fifa 2019 Dataset

This homework is divided in 4 parts.

The first part consists in the importation of the dataset and the extraction of a sub-dataset of 10 000 players (i.e. rows). In the second part the students apply the PCA algorithms they wrote during the course for the computation of the *principal components* (PC) of the matrix X identified by the *skills* columns of the dataset. The third part describes and give a qualitative interpretation of the "*main*" PC, with the help of Figure 1 and (eventually) of the values in the other dataset's columns. In the fourth part a new representation of the dataset is created (and saved) with respect to the "*main*" PC.

Part 1: Extraction of the Working Dataset

In this part, the working dataset is extracted. More specifically:

- load the original dataset[1] *fifa19datastats.csv* as pandas DataFrame or Table if you are using *python* (function *pandas.read_csv*) or *matlab* (function *readtable*), respectively.
- set a random seed s (*numpy.random.seed* in python, *rng* in matlab), then extract a sub-dataset selecting $N = 10\,000$ random rows (*numpy.random.random_integers* in python, *randperm* in matlab); for the extraction use the commands

$$\begin{aligned} \texttt{dataset.loc[array\_of\_rowindexes, :]} \quad & \text{(python)} \\ \texttt{dataset(array\_of\_rowindexes, :)} \quad & \text{(matlab)} \end{aligned} \tag{1}$$

  on the DataFrame/Table *dataset* representing *fifa19datastats.csv*. The sub-dataset extracted will be your *working dataset*: use it for this homework.

  Attention: the seed s must be your university registration number (or one of the two numbers of the group members).
- save the sub-dataset with the same name of the original one, followed by "*_seednumber*", e.g.: *fifa19datastats_123456.csv*.

[1] uploaded on the web page of the course.

Part 2: Application of the PCA

Using the PCA algorithms implemented in previous lessons, compute the PC of the skills, i.e.:

- Let $X \in \mathbb{R}^{N \times n}$ be the matrix corresponding to the skill columns of the working dataset. The commands

$$\begin{aligned} \texttt{dataset[['column\_name\_1', ..., 'column\_name\_n']].values} &\quad \text{(python)} \\ \texttt{dataset\{:,\{'column\_name\_1', ..., 'column\_name\_n'\}\}} &\quad \text{(matlab)} \end{aligned} \tag{2}$$

  let to extract an array of values from the selected columns of the DataFrame/Table *dataset* (use them for the computation of X).

- Choose which PCA algorithm apply to X (and do it).

Part 3: PC Interpretation

Once you have applied the PCA algorithms to X, comment the results as described:

- select the first $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$ PC such that they explain $\sim 90\%$ of the total variance of X.
- show as horizontal bar graphs the PC $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$ (*matplotlib.pyplot.barh* in python, *barh* in matlab).
- with the help of Figure 1, the horizontal bar graphs of the PC and (eventually) of the values in the other dataset's columns, give an interpretation and a *name* to $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$.
- looking at the code of the example with the *iris flowers* (PCA laboratory), show (and comment) an arbitrary number of scatter plots with respect to couples of PC $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$, where dots are colored differently for each *GeneralPosition* value (e.g. Goal Keepers are red, Defenders are blue, etc.).

Part 4: Other Representations of the Dataset

- Create a new DataFrame/Table "*wdataset*" equal to the working one but such that the skill columns are removed and replaced by m new columns.

  The i-th new column must have as title the name of $\boldsymbol{u}_i$ (see Part 3) and as value for the j-th row the element w_{ij} of W (matrix of PC coordinates).

  Suggestion: for the creation of *wdataset* extract the "common columns" with the commands

$$\begin{aligned} \texttt{dataset[['commcolumn\_name\_1', ..., 'commcolumn\_name\_m']]} &\quad \text{(python)} \\ \texttt{dataset(:, \{'commcolumn\_name\_1', ..., 'commcolumn\_name\_m'\})} &\quad \text{(matlab)} \end{aligned} \tag{3}$$

  and then add (one by one) the new columns with the commands

$$\begin{aligned} \texttt{dataset['newcolumn\_name'] = array} &\quad \text{(python)} \\ \texttt{dataset.newcolumn\_name = array} &\quad \text{(matlab)} \end{aligned} \tag{4}$$

- save *wdataset* as a *.csv* file (*wdataset.to_csv* in python, *writetable* in matlab) with name "*fifa19datastats_w_seednumber*".
- Create a new DataFrame/Table "*approxdataset*" equal to the working one but such that the values under the skill columns are not X but its approximation $\tilde{X}$ computed with respect to $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$. Then, compute the mean relative error of the approximation, i.e.

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\parallel \boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i \parallel}{\parallel \boldsymbol{x}_i \parallel}, \tag{5}$$

  and comment the result ($\boldsymbol{x}_i^\top, \tilde{\boldsymbol{x}}_i^\top$ are rows of X and $\tilde{X}$, respectively).

- save *approxdataset* as a *.csv* file with name "*fifa19datastats_approx_seednumber*".