DD2434/FDD3434 Machine Learning, Advanced Course
Assignment 2A, 2023

Aristides Gionis

Deadline, see Canvas

Read before starting

Please read the assignment questions carefully before starting working on the solutions.

You will present the assignment by a written report in PDF format, submitted before the deadline using Canvas. You may solve the assignment individually or in groups of two, and it will automatically be checked for similarities to other students' solutions as well as documents on the web in general. Although you are allowed to discuss the problem formulations with other groups, you are not allowed to discuss solutions, and any discussions concerning the problem formulations must be described in the solutions you hand in (including which group you discussed with).

From the report it should be clear what you have done and you need to support your claims with results. You are supposed to write down the answers to the specific questions detailed for each task. This report should clearly show how you have drawn your conclusions and explain your derivations. Your assumptions, if any, should be stated clearly. redShow the results of your experiments using images and graphs together with your analysis and add your code as an appendix.

Being able to communicate results and conclusions is a key aspect of scientific as well as corporate activities. It is up to you as a author to make sure that the report clearly shows what you have done. Based on this, and only this, we will decide if you pass the task. No detective work should be required on our side. In particular, neat and tidy reports please!

The grading of the assignment 1A and 2A will be as follows,

E 30-44 points, where at least 20 points are from Assignment 1A and 10 points from 2A.

D 45-60 points, where at least 20 points are from Assignment 1A and 10 points from 2A.

All points over 30 will be counted as bonus points for assignments 1B and 2B.

Good Luck!

2A/1 Principal Component Analysis (10 points)

While developing the PCA method, we required that the data are "centered." This step is performed by subtracting the mean from each data point. Essentially, with this step, we translate the center of mass of the data to the origin of the coordinate space.

Question 2A.1: *Explain why this data-centering step is required while performing PCA. What could be an undesirable effect if we perform PCA on non-centered data?*

Consider a data matrix of dimension $m \times n$. In some applications the role of points and dimensions can be interchanged. For example, given a document corpus represented as a matrix of type "documents × words", we may want to analyze documents based on which words occur in them, or we may want to analyze words based on which documents they appear in. So it is meaningful to perform PCA both with respect to the rows of a matrix and with respect to its columns.

As we discussed in the lectures, PCA relies on SVD. Moreover, since $(\mathbf{U\Sigma V}^T)^T = \mathbf{V\Sigma}^T\mathbf{U}^T = \mathbf{V\Sigma'U}^T$, where $\mathbf{\Sigma'}$ differs from $\mathbf{\Sigma}$ only in terms of size, performing SVD on a matrix gives also the SVD on its transpose.

Question 2A.2: *Does the previous argument imply that a* **single** *SVD operation is sufficient to perform PCA both on the rows and the columns of a data matrix?*

Justify your answer.

Consider a dataset $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ with n points of dimension d, i.e., $\mathbf{y}_i \in \mathbb{R}^d$. Assume that $d < n$. The *variance* of the dataset $\mathcal{Y}$ is

$$Var(\mathcal{Y}) = \sum_{\mathbf{y}\in\mathcal{Y}} \|\mathbf{y} - \bar{\mathbf{y}}\|_2^2, \tag{1}$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_{\mathbf{y}\in\mathcal{Y}} \mathbf{y}$ is the mean point. Further, assume that the data are zero-centered, so $\bar{\mathbf{y}} = \mathbf{0}$

The dataset $\mathcal{Y}$ can be represented as a $d \times n$ matrix $\mathbf{Y}$, and consider the SVD of $\mathbf{Y} = \mathbf{U\Sigma V}^T$.

Question 2A.3: *Show that the variance of the dataset $\mathcal{Y}$, as defined in Equation (1), can be expressed as a function of the singular values of $\mathbf{Y}$, and in particular*

$$Var(\mathcal{Y}) = \sum_{i=1}^{d} \sigma_i^2.$$

We perform PCA on $\mathcal{Y}$. Let $\mathbf{W}$ be the $d \times k$ matrix whose columns are the k first principal components of $\mathcal{Y}$, for $k < d$. Projecting $\mathcal{Y}$ on the space spanned by the columns of $\mathbf{W}$ gives the projected data points $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \{\mathbf{W}^T\mathbf{y}_1, \ldots, \mathbf{W}^T\mathbf{y}_n\}$, represented by the $k \times n$ matrix $\mathbf{X} = \mathbf{W}^T\mathbf{Y}$.

Question 2A.4: *Show that the variance of the projected data $\mathcal{X}$ is given by*

$$Var(\mathcal{X}) = \sum_{i=1}^{k} \sigma_i^2.$$

Finally, we consider the residual data points $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} = \{\mathbf{y}_1 - \mathbf{x}_1, \ldots, \mathbf{y}_n - \mathbf{x}_n\}$, where $\mathbf{z}_i = \mathbf{y}_i - \mathbf{x}_i = \mathbf{y}_i - \mathbf{W}^T\mathbf{y}_i$.

Question 2A.5: *Show that the variance of the residual data $\mathcal{Z}$ is given by*

$$Var(\mathcal{Z}) = \sum_{i=k+1}^{d} \sigma_i^2.$$

Conclude that

variance of original data = variance explained by PCA + variance of residual data.

2A/2 PCA vs. Johnson-Lindenstrauss random projections (2 points)

Both the PCA and the Johnson-Lindenstrauss random-projections methods are linear maps. Given data $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in \mathbb{R}^{d \times n}$, both methods find a matrix $\mathbf{A} \in \mathbb{R}^{k \times d}$ and reduce the dimension of the data from d to k by the projection $\mathbf{X} = \mathbf{AY}$.

Question 2A.6: *Provide a qualitative comparison (short discussion) between the two methods, PCA vs. Johnson-Lindenstrauss random projections, in terms of (i) projection error; (ii) computational efficiency; and (iii) target usecases.*

2A/3 Programming task — MDS (8 points)

Question 2A.7: (Data acquisition and processing)

In the website https://cadmus.eui.eu/handle/1814/74918 you can find a dataset providing voting information for members of the European Parliament (MEPs) on different issues.

We will conceptualize the data as a set of points, where each data point contains information about the votes of one MEP. In addition, for each MEP we have information about (i) their country and (ii) the European Parliament political group (EPG) they belong. This additional information, Country and EPG, can be seen as labels (colors) associated with the MEPs.

We want to estimate the degree to which two MEPs vote in a similar manner. Thus, we need to define a function that assigns similarity values to pairs of MEPs based on their votes. Your first task is to define two different similarity functions for this problem. You should aim for definitions of similarity that are meaningful for this domain and this dataset. Present the two functions that you came up with, and explain why you chose them. Please also discuss the transformation steps required for your similarity computation, e.g., mapping categorical to numerical values, handling missing values, etc.

Your second task is to preprocess the data and compute the pairwise similarity matrix between MEPs for the two functions you defined. Depending on the efficiency of your implementation, it is possible that computing such a matrix is computationally challenging. To reduce the computational cost of the exercise, it is OK to consider only a subset of MEPs, however, make sure that you take a large enough subset. Explain your reasoning for selecting that particular subset.

Question 2A.8: (MDS)

Apply MDS to compute an (x, y) coordinate for each MEP in your dataset, given the two similarity matrices you computed in the previous step.

Plot the MEPs on a plane using the coordinates you computed. Annotate the data points using the "colors" we mentioned above, Country and EPG.

Discuss the maps your created using the MDS method. For instance, which of the two similarity measures gives more intuitive results, and which of the two color annotations explains the data better?

For this task, you should implement the classical MDS methods yourself, by relying only on a package for eigenvector decomposition, that is, do not try to find an MDS function to use as a black box.