DD2434/FDD3434 Machine Learning, Advanced Course
Assignment 2AD, 2025

Aristides Gionis, Theofanis Georgakopoulos

Deadline, see Canvas

Read this before starting

You will present the assignment by a written report and code, submitted before the deadline using Canvas. Furthermore, there will be an oral exam after the deadline where you have to show understanding of your solutions in order to keep your passing score. You may use AI tools to assist you in the writing (See Canvas for use of AI in the course), but you must ensure you understand any solution you provide. You must solve the assignment individually 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 others, you are not allowed to discuss solutions.

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. Show 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 an 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!

Grading system

This assignment is divided into grade levels D, C, B and A. In order to receive a particular grade, you must pass the criteria for that level and all levels below it. Each subproblem contains one or more questions. Each subproblem is graded Pass/Fail. Passing criteria for each level:

D Passing 2/3 subproblems of 1AD and the D-level (programming) task in 2AD.

C Passing 3/4 subproblems of 1AD and the C-level subproblem in 2AD.

B Passing 2/3 subproblems of 1AD and 2AD (combined).

A Passing 2/3 subproblems of 1AD and 2AD (combined).

These grades are valid for assignments submitted before the deadline, late assignments can at most receive the grade E, which makes it meaningless to hand in late solutions for this assignment.

Bonus points from 1E and 2E work as "save" points for levels D and C. If you fail one of the subproblems, your bonus points can still make you pass the level.

- ≥ 27 points: 1 save point for D and C level.
- ≥ 35 points: 2 save points for D and C level.

Oral exams

In order to retain your passing grade, you must be able to explain each problem and solution in front of a Teacher or TA. Therefore, if you use AI-tools to help solve the questions, make sure that you understand the solutions!

Good Luck!

D level

2AD.D Data exploration with dimensionality reduction

Question 2AD.D.1: (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 a similarity function for this task. You should aim for a definition of similarity that is meaningful for this domain and this dataset. Present the similarity function that you came up with, and explain why you chose it. 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 function 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, you should make sure that you take a large enough subset. Explain your reasoning for selecting that particular subset.

Question 2AD.D.2: (MDS)

Apply MDS to compute an (x, y) coordinate for each MEP in your dataset, given the similarity matrix 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 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.

Question 2AD.D.3: (Explorative data analysis)

Make a more nuanced analysis of the MEP voting dataset. The objective is to obtain more interesting insights about this data. You could consider using other embedding methods, using additional attributes (e.g., types of votes, or time information), and so on.

This is not a fixed task, so you can be creative!

Discuss your hypothesis, present your methodology and the results you obtained, and explain in which ways your findings are more interesting than the ones you had obtained in Question 2AD.D.2.

C level

2AD.C Principal component analysis

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

The dataset $\mathcal{Y}$ can be represented as a $d \times n$ matrix $\mathbf{Y}$, and consider the SVD of $\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. Define the similarity (Gram) matrix $\mathbf{S} = \mathbf{Y}^T\,\mathbf{Y} \in \mathbb{R}^{n\times n}$.

Question 2AD.C.1: *Show that* $\mathbf{S} = \mathbf{V}\,\boldsymbol{\Sigma}^2\,\mathbf{V}^T$*.*

The classical MDS constructs a k-dimensional embedding using the top k eigenpairs of $\mathbf{S}$, so let

$$\mathbf{X}_{\text{MDS}} \in \mathbb{R}^{k\times n}\,,$$

and PCA obtains a k-dimensional embedding using the top k left singular vectors of $\mathbf{Y}$, so let

$$\mathbf{X}_{\text{PCA}} \in \mathbb{R}^{k\times n}\,.$$

Question 2AD.C.2: *Prove that the two embeddings differ only by an orthogonal transform. In particular, show there exists an orthogonal matrix* $\mathbf{R} \in \mathbb{R}^{k\times k}$ *such that* $\mathbf{X}_{\text{PCA}} = \mathbf{R}\,\mathbf{X}_{\text{MDS}}$*.*

Conclude that for Euclidean data, classical MDS yields the same embedding as PCA (up to rotation).

Next, we define the *variance* of the dataset $\mathcal{Y}$ to be

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

where $\bar{\mathbf{y}}$ is as defined above.

Question 2AD.C.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 2AD.C.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 $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, where $\mathbf{z}_i = \mathbf{y}_i - \mathbf{W}\mathbf{W}^T\mathbf{y}_i$.

Question 2AD.C.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.

B level

2AD.B Node similarity for representation learning

Let $G = (V, E)$ be an undirected and connected graph and let $\mathbf{A}$ be the adjacency matrix of G, that is, $\mathbf{A}_{ij} = 1$ if $(i, j) \in E$ and $\mathbf{A}_{ij} = 0$ otherwise.

Let $\mathbf{D}$ be a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$, and let $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$.

In graph representation learning, our goal is to learn vector representations (embeddings) for the nodes of the graph. The main idea is to define an appropriate similarity measure between the graph nodes, and then learn vector representations for the graph nodes, so that the similarity between pairs of learned vectors approximates the similarity between the corresponding graph nodes.

Assume now that for a similarity measure between graph nodes, we define

$$\mathbf{S}_{ij} = \sum_{k=1}^{\infty} \alpha^k \, \mathbf{P}_{ij}^k,$$

for each pair of nodes $i, j \in V$, and for some real $0 < \alpha < 1$.

Question 2AD.B.1: *Explain the intuition for the definition of the similarity measure $\mathbf{S}$.*

Question 2AD.B.2: *Show that for all $i, j \in V$ it is $\mathbf{P}_{ij} \leq 1$ and $\mathbf{S}_{ij} < +\infty$.*

Question 2AD.B.3: *Show that $\mathbf{S}$ can be computed efficiently using a matrix inversion operation.*

A level

2AD.A Metric MDS

Consider the metric MDS setting. We are given a distance matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ containing all pairwise distances between n data points, where $[\mathbf{D}]_{ij} = d_{ij}$. Our objective is to find a k-dimensional vector $\mathbf{x}_i \in \mathbb{R}^k$, for each point i, or equivalently a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{k\times n}$, so as to minimize the MDS "stress" function

$$E(\mathbf{X}) = \sum_{i<j} w_{ij} \left(d_{ij} - \|\mathbf{x}_i - \mathbf{x}_j\|\right)^2 , \tag{2}$$

where the weights w_{ij} are assumed to be known — either they are given as input or they can be computed as a function of the distances d_{ij}.

We discussed in class that Equation (2) does not have a closed-form solution and thus we resort to optimization by gradient-descend methods.

Question 2AD.A.1: *Show that the gradient of E with respect to a point $\mathbf{x}_i$ is*

$$\nabla_{\mathbf{x}_i} E = 2 \sum_{j\neq i} w_{ij} \left(1 - \frac{d_{ij}}{\|\mathbf{x}_i - \mathbf{x}_j\|}\right) (\mathbf{x}_i - \mathbf{x}_j) \, .$$

(Assume that $\mathbf{x}_i \neq \mathbf{x}_j$ for all pairs of points.)

Question 2AD.A.2: *Let $\mathbf{A} \in \mathbb{R}^{k\times k}$ be an orthonormal matrix, $\mathbf{b} \in \mathbb{R}^k$, and $\mathbf{1}_n \in \mathbb{R}^n$ a vector of all 1's. Show that*

$$E(\mathbf{A}\,\mathbf{X} + \mathbf{b}\mathbf{1}_n^T) = E(\mathbf{X}) \, .$$

Conclude the invariance of the stress function under rigid transformations and the non-uniqueness of the local optima solutions.

Question 2AD.A.3: *Let $k_1, k_2 \in \mathbb{N}$ such that $0 < k_1 < k_2$. Define the optimal stress in dimension k as*

$$\mathrm{OPT}(k) = \inf_{\mathbf{X}\in\mathbb{R}^{k\times n}} E(\mathbf{X}),$$

where $E(\mathbf{X})$ is defined as in Equation (2). Prove that

$$\mathrm{OPT}(k_2) \leq \mathrm{OPT}(k_1).$$

Conclude that the minimum achievable stress decreases as the embedding dimension grows.

Question 2AD.A.4: *Explain why a local optima of the stress function satisfy*

$$\sum_{j\neq i} w_{ij} \left(1 - \frac{d_{ij}}{\|\mathbf{x}_i - \mathbf{x}_j\|}\right) (\mathbf{x}_i - \mathbf{x}_j) = \mathbf{0} \, .$$

Question 2AD.A.5: *Interpret a local-optima solution as different forces acting on each point $\mathbf{x}_i$. Distinguish the cases when (i) $\|\mathbf{x}_i - \mathbf{x}_j\| > d_{ij}$ and (ii) $\|\mathbf{x}_i - \mathbf{x}_j\| < d_{ij}$.*