DD2434/FDD3434 Machine Learning, Advanced Course
Assignment 2B, 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 grade thresholds of assignments 1B and 2B are given below. Note that you can have 30 bonus points from assignments 1A and 2A.

D 30 points.

C 50 points.

B 70 points.

A 90 points.

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.

Good Luck!

2B/1 Multidimensional Scaling (MDS) and Isomap (10 points)

In the derivation of classical MDS with distance matrix, our goal is to derive the Gram matrix (similarity matrix) $\mathbf{S} = \mathbf{Y}^T\mathbf{Y}$ from the distance matrix $\mathbf{D}$, while $\mathbf{Y}$ is unknown.

We get $s_{ij} = -\frac{1}{2}(d_{ij}^2 - s_{ii} - s_{jj})$. In the lectures we mention the "double centering" trick, and how this can be used to solve for matrix $\mathbf{S}$ given $\mathbf{D}$. The mathematical derivation for the "double centering" trick is given in the textbook of Lee and Verleysen, Section 4.2.2.

Question 2B.1: *Explain in English what is the intuitive reason that the "double centering" trick works, and allow us to solve for* $\mathbf{S}$ *given* $\mathbf{D}$*.*

Question 2B.2: *Use the same reasoning as in the previous question to argue that* s_{ij} *can be computed as* $s_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{1i}^2 - d_{1j}^2)$*, where* d_{1i} *and* d_{1j} *are the distances from the first point in the dataset to points* i *and* j*, respectively.*

In particularly, argue that although the solution obtained by the "first point" trick will be different than the solution obtained by the "double centering" trick, both solutions are correct.

Consider the classical MDS algorithm when $\mathbf{Y}$ is known. In that case, we form $\mathbf{S} = \mathbf{Y}^T\mathbf{Y}$ and obtain the MDS embedding by the eigen-decomposition of $\mathbf{S}$. Observe that PCA involves a singular-value decomposition (SVD) operation, while classical MDS involved an eigenvector decomposition (EVD) operation.

Question 2B.3: *Show that the two methods, that is, classical MDS when* $\mathbf{Y}$ *is known and PCA on* $\mathbf{Y}$*, are equivalent.*

Which of the two methods is more efficient? (Hint: Your answer may involve a case analysis.)

Consider the Isomap method used to reduce the dimensionality of a given dataset. Isomap requires constructing a neiborhood graph G, as discussed in the lectures.

Question 2B.4: *Argue that the process to obtain the neiborhood graph* G *in the Isomap method may yield a disconnected graph. Provide an example.*

Question 2B.5: *Propose a heuristic to patch this problem. Explain the intuition of your heuristic and argue why it will be expected to work well in practice. How does it behave in the example you provided in the previous question?*

2B/2 Success probability in the Johnson-Lindenstrauss lemma (5 points)

In the proof of Johnson-Lindenstrauss lemma we first bounded the probability that a single projection maintains all pairwise distances with distortion that is between $(1-\epsilon)$ and $(1+\epsilon)$. In particular, we showed that the probability of achieving such a distortion for all pairs of points is at least $1/n$. Assume now, that we want to boost the probability of success to be at least 95%.

Question 2B.6: *Show that $\mathcal{O}(n)$ independent trials are sufficient for the probability of success to be at least 95%.*

An independent trial here refers to generating a new projection of the data points with a newly-generated projection matrix.

2B/3 Node similarity for representation learning (5 points)

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 2B.7: *Explain the intuition for the definition of the similarity measure* $\mathbf{S}$.

Question 2B.8: *Show that* $\mathbf{S}$ *can be computed efficiently using a matrix inversion operation.*

2B/4 Spectral graph analysis (5 points)

Question 2B.9: *Let $G = (V, E)$ be an undirected d-regular graph, let A be the adjacency matrix of G, and let $L = I - \frac{1}{d}A$ be the normalized Laplacian of G. Prove that for any vector $\mathbf{x} \in \mathbb{R}^{|V|}$ it is*

$$\mathbf{x}^T L\, \mathbf{x} = \frac{1}{d} \sum_{(u,v)\in E} (x_u - x_v)^2. \tag{1}$$

Question 2B.10: *Show that the normalized Laplacian is a positive semidefinite matrix.*

Question 2B.11: *Assume that we find a non-trivial vector $\mathbf{x}_*$ that minimizes the expression $\mathbf{x}^T L\, \mathbf{x}$. First explain what non-trivial means. Second explain how $\mathbf{x}_*$ can be used as an embedding of the vertices of the graph into the real line. Use Equation (1) to justify the claim that $\mathbf{x}_*$ provides a meaningful embedding.*

2B/5 Programming task (5 points)

This part is continuation to the programming task 2A/3.

Question 2B.12:

Make a more nuanced analysis of the dataset in 2A/3. The objective is to obtain more interesting insights about the MEP voting 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 2A/3.