DD2434/FDD3434 Machine Learning, Advanced Course
Assignment 3E, 2025

Aristides Gionis

Deadline, see Canvas

Read this before starting

You will present the assignment by a written report in PDF format, submitted before the deadline using Canvas. The assignment should be done 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. 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!

The grading of the assignment 1E, 2E and 3E (20 points each) will be as follows,

E 40 points, with least 10 points from each assignment.

- All points over 40 will be counted as bonus points for assignments 1AD and 2AD.

Good Luck!

3E.1 Principal component analysis (10 points)

While developing the PCA method, we consider that each data point $\mathbf{y} \in \mathbb{R}^d$ is generated by a latent vector $\mathbf{x} \in \mathbb{R}^k$, with $k < d$, through a linear transformation

$$\mathbf{y} = \mathbf{W}\,\mathbf{x}\,, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times k}$ is a matrix with orthonormal columns. We then deduced that the inverse map is obtained by

$$\mathbf{x} = \mathbf{W}^{+}\,\mathbf{y}\,, \tag{2}$$

where $\mathbf{W}^{+}$ is the pseudo-inverse of $\mathbf{W}$, that is,

$$\mathbf{W}^{+} = \mathbf{V}\,\mathbf{\Sigma}^{+}\,\mathbf{U}^{T}\,, \tag{3}$$

where $\mathbf{W} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^{T}$ is the SVD decomposition of $\mathbf{W}$, and $\mathbf{\Sigma}^{+}$ is formed from $\mathbf{\Sigma}$ by taking the reciprocal of all the non-zero elements, leaving all the zeros in place, and ensuring that $\mathbf{\Sigma}^{+}$ has the correct dimensions.

Question 3E.1.1: *Explain why we can assume that the matrix* $\mathbf{W}$ *has orthonormal columns.*

Question 3E.1.2: *Show how we derive Equation (2) from Equation (1) where the pseudo-inverse* $\mathbf{W}^{+}$ *is given by Equation (3).*

Question 3E.1.3: *Justify why in this particular case of deriving PCA, Equation (3) simplifies to* $\mathbf{W}^{+} = \mathbf{W}^{T}$.

For the PCA method, we also 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 3E.1.4: *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 $d \times n$. 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 $\times$ 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 $(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{T})^{T} = \mathbf{V}\mathbf{\Sigma}^{T}\mathbf{U}^{T} = \mathbf{V}\mathbf{\Sigma}'\mathbf{U}^{T}$, 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 3E.1.5: *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.

In one exercise session we discussed whether we should normalize the features before performing PCA. We will now look in a bit more detail about the consequence of normalizing the data features.

As usual, we consider a data matrix $\mathbf{Y} \in \mathbb{R}^{d \times n}$, i.e., the data points are represented as columns of $\mathbf{Y}$, and assume that the data points in $\mathbf{Y}$ are centered. We consider a $d \times d$ diagonal matrix $\mathbf{D} = \text{diag}(c_1, \ldots, c_d)$, where the constant c_j is scaling the j-th feature. The scaled data are then represented by $\mathbf{Y}' = \mathbf{D}\,\mathbf{Y}$.

Question 3E.1.6: *Consider the SVD of $\mathbf{Y}$ and the SVD of $\mathbf{Y}'$. Are the SVDs of these two matrices related or not?*

Using your insights from the previous question (SVD on $\mathbf{Y}$ vs. SVD on $\mathbf{Y}'$) argue whether there is a connection between the PCA on $\mathbf{Y}$ and the PCA on $\mathbf{Y}'$.

Conclude by reasoning whether PCA is sensitive to feature scaling or not.

3E.2 Isomap (3 points)

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

Question 3E.2.1: *Argue that the process to obtain the neiborhood graph G in the Isomap method may yield a disconnected graph.*

Provide an example.

Question 3E.2.2: *What would be the effect of obtaining a disconnected graph and why is it undesirable?*

Question 3E.2.3: *Propose a heuristic to address this issue and avoid obtaining a disconnected graph. 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?

3E.3 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 $\mathbf{X} = \mathbf{A}\mathbf{Y}$.

Question 3E.3.1: *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.*

3E.4 Spectral graph analysis (5 points)

Question 3E.4.1: *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*

$$\mathbf{x}^T L\, \mathbf{x} = \frac{1}{d} \sum_{(u,v)\in E} (x_u - x_v)^2. \tag{4}$$

Question 3E.4.2: *Show that the normalized Laplacian is a positive semidefinite matrix.*

Question 3E.4.3: *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 (4) to justify the claim that $\mathbf{x}_*$ provides a meaningful embedding.*