DD2434/FDD3434 Machine Learning, Advanced Course
Assignment 3A, 2024

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 1A, 2A and 3A will be as follows,

E 30-44 points, with least 10 points from each Assignment.

D 45-60 points, with least 10 points from each Assignment.

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

Good Luck!

3.1 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 3.1.1: *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 $\times$ 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 3.1.2: *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 $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

Question 3.1.3: *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 3.1.4: *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 3.1.5: *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.

3.2 Classifying non-linearly-separable data with a linear classifier (4 points)

Consider a set of points in 2 dimensions, where the data points have one of two labels, $\star$ or $\circ$. An example of such a dataset is shown in Figure 1. Assume that our task is to train a classifier that classifies the data points according to their label.

As depicted in the figure, the data points are not linearly separable according to their label. However, assume that for the purpose of this exercise we would like to use a linear classifier, such as, for example, the perceptron algorithm. A reason for insisting on using a linear classifier could be that we have access to an efficient and robust implementation of such a method.

For classifying the non-linearly-separable data using the linear classifier we plan to pursue the following idea:

1. Map the data in a two-dimensional space (different from the original two-dimensional representation) so that the data become linearly separable according to their label, and then
2. use the linear-classification algorithm (e.g., the perceptron) to train a classifier.

Question 3.2.6: *Explain in detail how to achieve step 1 above, i.e., how to map the data into a two-dimensional space so that they become linearly separable. Write down all the steps of your proposed method. Start with a two-dimensional representation of the data in the original (non-linearly-separable) space, and explain how to obtain a mapping into the two-dimensional (linearly-separable) space.*

Question 3.2.7: *Assume now that we have obtained a projection to the linearly-separable space and we have trained the linear classifier using the available data. A new (previously-unseen) data point is given (represented in the original two-dimensional space) and we want to apply the linear classifier on that point to obtain a prediction for its label. Explain in detail your method to classify such a previously-unseen data point.*

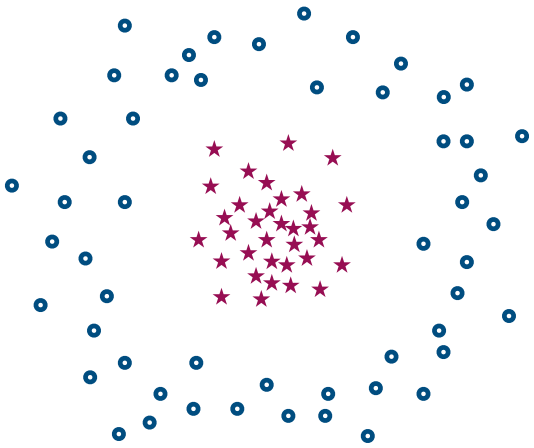

Figure 1: A dataset containing data points of two labels, which are not linearly separable.

3.3 Spectral graph analysis (6 points)

Question 3.3.8: *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*

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

Question 3.3.9: *Show that the normalized Laplacian is a positive semidefinite matrix.*

Question 3.3.10: *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 (2) to justify the claim that $\mathbf{x}_*$ provides a meaningful embedding.*