Neural Networks Homework 3

Sayeh Jarollahi (7073520, saja00006@stud.uni-saarland.de)
Mahsa Amani (7064006, maam00002@stud.uni-saarland.de)

November 10, 2024

Exercise 3.1

a)

Proof. Let's start with this equation:

$$f(W) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x} - \mathbf{y})^2 \tag{1}$$

To minimize $f(W)$, we take its derivative with respect to w and set it to 0, $\nabla_w f(W) = 0$.

$$\nabla_w f(W) = \nabla_w \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x} - \mathbf{y})^2 = \frac{2}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x} - \mathbf{y})\mathbf{x} = 0 \tag{2}$$

As $\frac{2}{N}$ is not 0, so the $\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x} - \mathbf{y})\mathbf{x} = 0$.

$$\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x} - \mathbf{y})\mathbf{x} = \sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x})\mathbf{x} - \sum_{i=1}^{N}\mathbf{y}\mathbf{x} = 0 \tag{3}$$

Based on the associative property of matrices and that both of $\mathbf{w}^T\mathbf{x}$ and $\mathbf{x}^T\mathbf{w}$ are scaler values, we can rewrite the above equation as follows:

$$\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x})\mathbf{x} - \sum_{i=1}^{N}\mathbf{y}\mathbf{x} = \sum_{i=1}^{N}\mathbf{x}(\mathbf{x}^T\mathbf{w}) - \sum_{i=1}^{N}\mathbf{y}\mathbf{x} = \sum_{i=1}^{N}(\mathbf{x}\mathbf{x}^T)\mathbf{w} - \sum_{i=1}^{N}\mathbf{y}\mathbf{x} = 0 \tag{4}$$

$$\sum_{i=1}^{N}(\mathbf{x}\mathbf{x}^T)\mathbf{w} = \sum_{i=1}^{N}\mathbf{y}\mathbf{x} \tag{5}$$

$$\sum_{i=1}^{N}(\mathbf{x}\mathbf{x}^T)\mathbf{w} = \sum_{i=1}^{N}\mathbf{y}\mathbf{x} \tag{6}$$

Now, if we design matrix X with rows of $\mathbf{x}$, and vector $\mathbf{y}$ with label values, we can simplify the equation as:

$$(X^TX)\mathbf{w} = X^T\mathbf{y} \tag{7}$$

$$\mathbf{w} = (X^TX)^{-1}X^T\mathbf{y} \tag{8}$$

i.

1. Time and Space complexity: Calculation of $(X^TX)^{-1}$, which requires the inverse calculation, has $O(n^3)$ complexity. This will become even more complex in terms of space complexity as the size of the dataset increases (we will need more space to save X).

2. Numerical instability: To calculate the inverse of (X^TX), it should be invertible, by adding a regularization term we can solve this problem.

3. Offline learning: As it requires all the data at once, it is not suitable for online learning.

ii. Yes, it will work even when X is not invertible because (X^TX) might still be invertible. In this case, when (X^TX) is not invertible, we can calculate its inverse by adding a regularization term to the equation:

$$\mathbf{w} = (X^TX + \lambda I)^{-1}X^T\mathbf{y} \tag{9}$$

This regularization term helps ensure that $(X^TX + \lambda I)$ is invertible.

□

b)

Proof. The matrix X and vector $\mathbf{y}$ are as follows:

$$X = \begin{bmatrix} 1 & 9.0 \\ 1 & 2.0 \\ 1 & 6.0 \\ 1 & 1.0 \\ 1 & 8.0 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1.0 \\ 0.0 \\ 3.0 \\ 0.0 \\ 1.0 \end{bmatrix} \tag{10}$$

$$X^TX = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 9.0 & 2.0 & 6.0 & 1.0 & 8.0 \end{bmatrix} \begin{bmatrix} 1 & 9.0 \\ 1 & 2.0 \\ 1 & 6.0 \\ 1 & 1.0 \\ 1 & 8.0 \end{bmatrix} = \begin{bmatrix} 5 & 26.0 \\ 26.0 & 186.0 \end{bmatrix} \tag{11}$$

$$(X^TX)^{-1} = \frac{1}{5 \times 186 - 26 \times 26} \begin{bmatrix} 186.0 & -26.0 \\ -26.0 & 5 \end{bmatrix} = \begin{bmatrix} 0.7323 & -0.1024 \\ -0.1024 & 0.0197 \end{bmatrix} \tag{12}$$

$$X^T\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 9.0 & 2.0 & 6.0 & 1.0 & 8.0 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.0 \\ 3.0 \\ 0.0 \\ 1.0 \end{bmatrix} = \begin{bmatrix} 5.0 \\ 35.0 \end{bmatrix} \tag{13}$$

$$\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y} = \begin{bmatrix} 0.7323 & -0.1024 \\ -0.1024 & 0.0197 \end{bmatrix} \begin{bmatrix} 5.0 \\ 35.0 \end{bmatrix} = \begin{bmatrix} 0.0775 \\ 0.1775 \end{bmatrix} \tag{14}$$

$$y = 0.0775 + 0.1775x \tag{15}$$

□

c)

Proof. For dataset D1 we have:

$$y = b_1 + w_1 x \tag{16}$$

and for D2 dataset we have:

$$y + \eta = b_2 + w_2(x + \gamma) = b_2 + w_2 x + w_2 \gamma \tag{17}$$

$$y = b_2 + w_2 x + w_2 \gamma - \eta \tag{18}$$

For these 2 models to be consistent, we need:

$$w_1 = w_2 \tag{19}$$

and

$$b_1 = b_2 + w_2 \gamma - \eta \tag{20}$$

which can be rewritten using equation (19):

$$b_1 = b_2 + w_1 \gamma - \eta \tag{21}$$

Given the constraint ($w_1\gamma \neq \eta$), b_1 and b_2 must be different, thus case (c) holds. This implies the slope w remains unchanged as the linear relationship between x and y in D1 and D2 datasets does not change; only the intercept b is affected due to the shifts by γ and η. The intercept b changes because $w_1\gamma \neq \eta$, so the adjustment in y does not match the adjustment in x scaled by the slope.

□

Exercise 3.2

Proof. a) Normalization makes sure that the scale of different variables does not affect the final result and variables are equally contributing. If normalization is not applied, PCA might determine that the direction of maximal variance corresponds highest with the variable that has the largest scale.

b) PCA fails to capture circular patterns, hence we can say a dataset that consists of:

$$x^2 + y^2 + \epsilon_1 = 1$$
$$x^2 + y^2 + \epsilon_2 = 4$$

The image of such dataset is as follows:

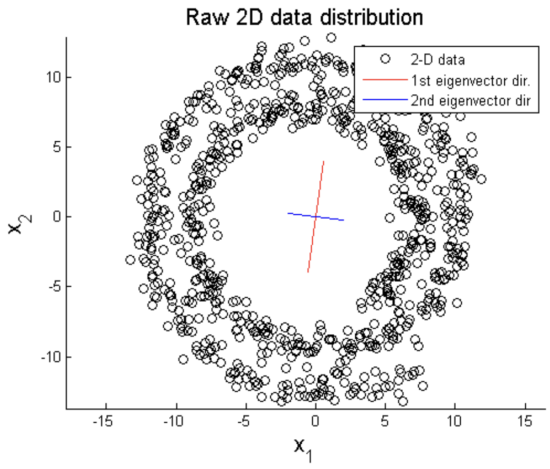

Figure 1: Sample where PCA fails with eigenvectors

Reference: link of reference

□

Exercise 3.3

Proof. a) As we know, in Principal Component Analysis, if we have samples $x_1, x_2, x_3, x_4, x_5, x_6$ in the R^4 space, the number of eigenvalues will be 4, since it equals to number of features.
b) To calculate how much of the variance in the data is preserved by the first two principal, we only need to sum up their eigenvalues and divide it to the sum of all eigenvalues. Because the eigenvalues show how much information of data(including variance) is in the corresponding eigenvector. Hence we have:

$$\frac{(0.739 + 0.685)}{(0.739 + 0.685 + 0.239 + 0.004)} = \frac{1.424}{1.667} = 0.85$$

c) As mentioned in the section b, we calculate this value for eigenvalues one and three:

$$\frac{0.739 + 0.239}{0.739 + 0.685 + 0.239 + 0.004} = \frac{0.978}{1.667} = 0.586$$

□