# Neural Networks Homework 5

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

November 27, 2024

## Exercise 5.1

*Proof.* a) The distance d from a point $x_i$ to the hyperplane is given by the formula:

$$d = \frac{|w^T x_i + b|}{||w||_2}$$

In the context of SVM, we also consider the label $y_i$ to ensure that the margin aligns with the classification

$$y_i(w^T x_i + b) > 0, \forall i$$

So the signed distance to the hyperplane can then be written as:

$$d_i = \frac{y_i(w^T x_i + b)}{\|w\|_2}$$

To determine the margin of the hyperplane, we find the minimum distance among all points:

$$\text{margin} = \min_i (\frac{y_i(w^T x_i + b)}{\|w\|_2})$$

b) First we need to find derivatives of *Lagrangian* w.r.t $\mathbf{w}$ and $b$ and set them to zero:

$$L(\mathbf{w}, b, \lambda) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{N} \lambda_i (1 - y_i(\mathbf{w}^T x_i + b))$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \lambda_i y_i x_i = 0$$

This gives us the equation:

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i x_i$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i = 0$$

This gives us:

$$\sum_{i=1}^{N} \lambda_i y_i = 0$$

Now that we have $\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i x_i$, substitute this into the Lagrangian $L(\mathbf{w}, b, \lambda)$ to eliminate $\mathbf{w}$. The Lagrangian becomes:

$$L(\lambda) = \frac{1}{2} \left( \sum_{i=1}^{N} \lambda_i y_i x_i \right)^T \left( \sum_{i=1}^{N} \lambda_i y_i x_i \right) + \sum_{i=1}^{N} \lambda_i \left( 1 - y_i(\mathbf{w}^T x_i + b) \right)$$

$$L(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j$$

Now, we solve the dual problem by maximizing $L(\lambda)$ with respect to the Lagrange multipliers $\lambda_i$:

$$\text{maximize} L(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j$$

subject to the constraints:

$$\lambda_i \geq 0, \quad \forall i, \quad \sum_{i=1}^{N} \lambda_i y_i = 0$$

This gives the *dual problem* of the SVM, and solving it provides the values for $\lambda_i$. Once the $\lambda_i$ are determined, we can compute $w$ and $b$ using the equations derived earlier.
c) It determines how much influence each point has on the margin between the two classes. Higher values of $\lambda_i$ indicate that the corresponding data point is a support vector and has a greater impact on the decision boundary

$\square$

## Exercise 5.2

*Proof.* a)
To calculate the gradient of $f$, we need to find its partial derivatives with respect to $x_1$ and $x_2$:

$$\frac{\partial f}{\partial x_1} = 2x_1 - 3 - x_2, \quad \frac{\partial f}{\partial x_2} = 2x_2 - x_1$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 - 3 - x_2 \\ 2x_2 - x_1 \end{bmatrix}$$

The Hessian matrix is the matrix of second-order partial derivatives:

$$\frac{\partial^2 f}{\partial x_1^2} = 2, \quad \frac{\partial^2 f}{\partial x_1 \partial x_2} = -1, \quad \frac{\partial^2 f}{\partial x_2^2} = 2, \quad \frac{\partial^2 f}{\partial x_2 \partial x_1} = -1$$

$$H_f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

Now, to minimize $f(x)$, we set $\nabla f(x) = 0$:

$$\begin{bmatrix} 2x_1 - 3 - x_2 \\ 2x_2 - x_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

From the first equation:

$$2x_1 - 3 - x_2 = 0 \implies x_2 = 2x_1 - 3$$

From the second equation:

$$2x_2 - x_1 = 0 \implies x_2 = \frac{x_1}{2}$$

Equating the both of them:

$$2x_1 - 3 = \frac{x_1}{2} \implies 4x_1 - 6 = x_1 \implies 3x_1 = 6 \implies x_1 = 2, x_2 = 2x_1 - 3 = 2(2) - 3 = 1$$

Thus, the critical point is:

$$\hat{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

A function is convex if its Hessian matrix is positive semidefinite. To check this, we need to find the eigenvalues of $H_f(x)$ by solving $\det(H_f - \lambda I) = 0$.

$$H_f(x) = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

$$\det \begin{bmatrix} 2 - \lambda & -1 \\ -1 & 2 - \lambda \end{bmatrix} = 0$$

$$(2 - \lambda)^2 - (-1)(-1) = 0 \implies (2 - \lambda)^2 - 1 = 0$$

$$(2-\lambda-1)(2-\lambda+1)=0 \implies \lambda = 1 \text{ or } 3$$

Since both eigenvalues (1 and 3) are positive, then $H_f(x)$ is positive definite, and $f(x)$ is convex. So, we can guarantee the critical point $\hat{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ is the global minimum.

b)

Using the gradient descent update rule $x_{k+1} = x_k - \varepsilon \nabla f(x_k)$, the iterations are as follows:

**Iteration 1**:

$$x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$f(x_0) = 1^2 - 3(1) + 1^2 - 1(1) = 1 - 3 + 1 - 1 = -2$$

$$\nabla f(x_0) = \begin{bmatrix} 2(1) - 3 - 1 \\ 2(1) - 1 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

$$x_1 = x_0 - \varepsilon \nabla f(x_0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.5 \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1+1 \\ 1-0.5 \end{bmatrix} = \begin{bmatrix} 2 \\ 0.5 \end{bmatrix}$$

**Iteration 2**:

$$x_1 = \begin{bmatrix} 2 \\ 0.5 \end{bmatrix}$$

$$f(x_1) = 2^2 - 3(2) + 0.5^2 - 2(0.5) = 4 - 6 + 0.25 - 1 = -2.75$$

$$\nabla f(x_1) = \begin{bmatrix} 2(2) - 3 - 0.5 \\ 2(0.5) - 2 \end{bmatrix} = \begin{bmatrix} 1.5 \\ -1 \end{bmatrix}$$

$$x_2 = x_1 - \varepsilon \nabla f(x_1) = \begin{bmatrix} 2 \\ 0.5 \end{bmatrix} - 0.5 \begin{bmatrix} 1.5 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 - 0.75 \\ 0.5 + 0.5 \end{bmatrix} = \begin{bmatrix} 1.25 \\ 1 \end{bmatrix}$$

At first we had $f(x_0) = -2$, after 2 iterations we have $f(x_2) = -2.4375$ and the global minimum $f(\hat{x})$ at $\hat{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ is $f(\hat{x}) = 2^2 - 3(2) + 1^2 - 2(1) = 4 - 6 + 1 - 2 = -3$. This shows the gradient descent is working and with each iteration we are getting closer to the global minimum.

c)

Using:

$$f(x) = x_1^2 - 3x_1 + x_2^2 - x_1x_2, \nabla f(x) = \begin{bmatrix} 2x_1 - 3 - x_2 \\ 2x_2 - x_1 \end{bmatrix}, \quad H_f(x) = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

And the Newton's update rule $x_{t+1} = x_t - H_f(x_t)^{-1} \nabla f(x_t)$, the iterations are as follows:

**Iteration 1**:

$$\nabla f(x_0) = \begin{bmatrix} 2(1) - 3 - 1 \\ 2(1) - 1 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

$$H_f(x_0) = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

$$H_f(x_0)^{-1} = \frac{1}{\det(H_f)} \text{adj}(H_f) = \frac{1}{(2)(2) - (-1)(-1)} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Now, we use Newton's update rule:

$$x_1 = x_0 - H_f(x_0)^{-1} \nabla f(x_0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} -3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1+1 \\ 1+0 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

The updated point is $x_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ where $\nabla f(x_1) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Thus, $x_1$ is a critical point with $f(x_1) = -3$ and we have found the global minimum and cannot continue.

□

## Exercise 5.3

*Proof.* a) Change of sign of input and output weights for a single neuron should be considered. There are m neurons that each of them can have two different transformations. Hence we have $2^M$ transformations from this side.
Also we have to consider transformations between two neurons. By choosing each two neurons, we can have a transformation. There are $\frac{M(M-1)}{2}$ transofmations with this type.
So we conclude that totally there are $2^M \cdot \frac{M(M-1)}{2}$ transformations.
b)
The transformations in each layer is independent of the other layers. Hence we can say the total number of transformations is:

$$\prod_{i=1}^{N} (2^{M_i} \times \frac{M_i(M_i-1)}{2})$$

□