NNTI Assignment 04

Name 1: Md Mobashir Rahman
Student ID 1: 7059086
Email 1: mdra00001@stud.uni-saarland.de

Name 2: Ratnadeep Chakraborty
Student ID 2: 7022859
Email 2: rach00002@stud.uni-saarland.de

Exercise 5.1: SVMs and Kernels

(a)

Given the classifier as a hyperplane defined by $\mathbf{w}$ and b, the equation of the hyperplane is:

$$\mathbf{w}^T\mathbf{x} + b = 0$$

The condition for a correctly classified point $\mathbf{x}_i$ is:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0$$

The margin is defined by the closest point from the hyperplane:

$$\text{margin} = \frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{\|\mathbf{w}\|_2}$$

To find the closest point, we have to consider all the points and find the one with the minimum distance:

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

(b)

The given Lagrangian is:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{N} \lambda_i \left(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right)$$

First, differentiating $\mathcal{L}$ with respect to $\mathbf{w}$ and setting it to 0, we get:

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

So:

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

Now, we differentiate $\mathcal{L}$ with respect to b:

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

Therefore:

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

Now, substituting $\mathbf{w}$ into the Lagrangian:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{N} \lambda_i \left(1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)\right)$$

Calculating $\|\mathbf{w}\|_2^2$:

$$\|\mathbf{w}\|_2^2 = \mathbf{w}^T \mathbf{w}$$

Substituting $\mathbf{w}$:

$$\|\mathbf{w}\|_2^2 = \left(\sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i\right)^T \left(\sum_{j=1}^{N} \lambda_j y_j \mathbf{x}_j\right)$$

Expanding:

$$\|\mathbf{w}\|_2^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

After substitution, we get:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{N} \lambda_i \left[1 - y_i \left(\sum_{j=1}^{N} \lambda_j y_j \mathbf{x}_j^T \mathbf{x}_i + b\right)\right]$$

Differentiating with respect to b gives us:

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

Therefore, the final equation is:

$$\mathcal{L}(\boldsymbol{\lambda}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^{N} \lambda_i + b \sum_{i=1}^{N} \lambda_i y_i$$

(c)

λ_i is the Lagrange multiplier. When $\lambda_i > 0$, it means the point x_i lies on the margin and influences the decision boundary. When $\lambda_i = 0$, it means the point x_i is not a support vector and does not affect the boundary.

Exercise 5.2: Gradient Descent and Newton's Method

(b)

Given the function:

$$f(x) = x_1^2 - 3x_1 + x_2^2 - x_1x_2$$

As we know, the gradient is the vector of partial derivatives with respect to x_1 and x_2.

The gradient of $f(x)$ is given by:

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

Taking the partial derivative with respect to x_1:

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

The partial derivative with respect to x_2:

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

So:

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

Calculating the Hessian matrix:

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

The second derivatives are:

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

So, the Hessian matrix is:

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

To find the critical points, we set:

$$\nabla f(x) = 0$$

This gives:

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

From the first equation:

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

From the second equation:

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

Substituting $x_2 = \frac{x_1}{2}$ into $x_2 = 2x_1 - 3$:

$$\frac{x_1}{2} = 2x_1 - 3$$

Solving for x_1:

$$x_1 = 4x_1 - 6 \implies x_1 = 2$$

So, $x_1 = 2$ and $x_2 = 1$.
This gives us:

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

which minimizes f.

To find out if $\hat{x}$ is guaranteed to be the global minimum, we need to look at the eigenvalues of $H_f(x)$.

The eigenvalues of $H_f(x)$ are 1 and 3, which are both positive. This means $f(x)$ is convex, and $\hat{x}$ is guaranteed to be the global minimum.

(b)

Starting point:

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

Learning rate: $\epsilon = 0.5$.

Iteration 1:

Calculating $\nabla f(x_0)$:

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

Updating x:

$$x_1 = x_0 - 0.5 \cdot \nabla f(x_0)$$

$$x_1 = \begin{bmatrix}1\\1\end{bmatrix} - 0.5 \cdot \begin{bmatrix}-2\\1\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}-1\\0.5\end{bmatrix} = \begin{bmatrix}2\\0.5\end{bmatrix}$$

Computing $f(x_1)$:

$$f(x_1) = (2)^2 - 3(2) + (0.5)^2 - (2)(0.5)$$

$$f(x_1) = 4 - 6 + 0.25 - 1 = -2.75$$

Iteration 2:
Calculating $\nabla f(x_1)$:

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

Updating x:

$$x_2 = x_1 - 0.5 \cdot \begin{bmatrix}0.5\\-1\end{bmatrix}$$

$$x_2 = \begin{bmatrix}2\\0.5\end{bmatrix} - \begin{bmatrix}0.25\\-0.5\end{bmatrix} = \begin{bmatrix}1.75\\1\end{bmatrix}$$

Compute $f(x_2)$:

$$f(x_2) = (1.75)^2 - 3(1.75) + (1)^2 - (1.75)(1)$$

$$f(x_2) = -2.9375$$

(c)

Newton's method is a second-order optimization technique used to find critical points of a function. It is based on the Taylor expansion of $f(x)$ at x_t:

$$f(x) \approx f(x_t) + \nabla f(x_t)^T (x - x_t) + \frac{1}{2}(x - x_t)^T H_f(x_t)(x - x_t)$$

If we set the gradient of the above equation to 0:

$$\nabla f(x_t) + H_f(x_t)(x - x_t) = 0$$

This gives:

$$x = x_t - H_f(x_t)^{-1} \nabla f(x_t)$$

So, the Newton's update rule is:

$$x_{t+1} = x_t - H_f(x_t)^{-1} \nabla f(x_t)$$

Exercise 5.3: Weight Space Symmetry

(a)

If each neuron from M neurons keeps its weights unchanged or flips their signs, then there are 2^M possible configurations.

Similarly, for M neurons, there remain $M!$ ways of swapping neurons in the hidden layer.

So, the total number of equivalent transformations:

$$2^M \cdot M!$$

(b)

In the previous question, we found that there remain $2^M \cdot M!$ ways of transformation in one hidden layer. When we have N hidden layers, then:

$$\text{Total transformations} = \prod_{i=1}^{N} \left(2^{M_i} \cdot M_i!\right)$$