Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNTI Winter Term 2024/2025

Exercise Sheet 5

Feedforward Neural Networks

Deadline: 27.11.2024 23:59

Guidelines: You are expected to work in a group of 2-3 students. While submitting the assignments, please make sure to include the following information for all our teammates in your PDF/python script:

Name:

Student ID (matriculation number):

Email:

Your submissions should be zipped as **Name1_id1_Name2_id2_Name3_id3.zip** when you have multiple files. For assignments where you are submitting a single file, use the **same naming convention** without creating a zip. For any clarification, please reach out to us on the **CMS Forum**.
Note that the above instructions are mandatory. If you are not following them, tutors can decide not to correct your exercise.

Exercise 5.1 - SVMs and Kernels (0.5+2+0.5 points)

In this exercise, we will use SVMs (refer to the book Pattern Recognition and Machine Learning in the chapter on Sparse Kernel Machines for a more detailed understanding.) to solve a classification problem between two classes.

Assume that the data pairs in the training set are $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ where the vector $x_i \in \mathbb{R}^d$ represents the input of a data point and y_i is the label of that data point. Here, d is the dimensionality of the data, and N is the number of data points. Assume that the label of each data point is determined by $y_i = 1$ or $y_i = -1$ and the training data set is linearly separable.

For any input x_i, consider w (the weight vector normal to the classifier hyperplane) and the bias b such that the sign of $w^T x_i + b$ determines the class of the input. We can assume that with a classifier dividing the feature space into two halves, the positive half will contain the points with label $y_i = 1$, and vice versa.

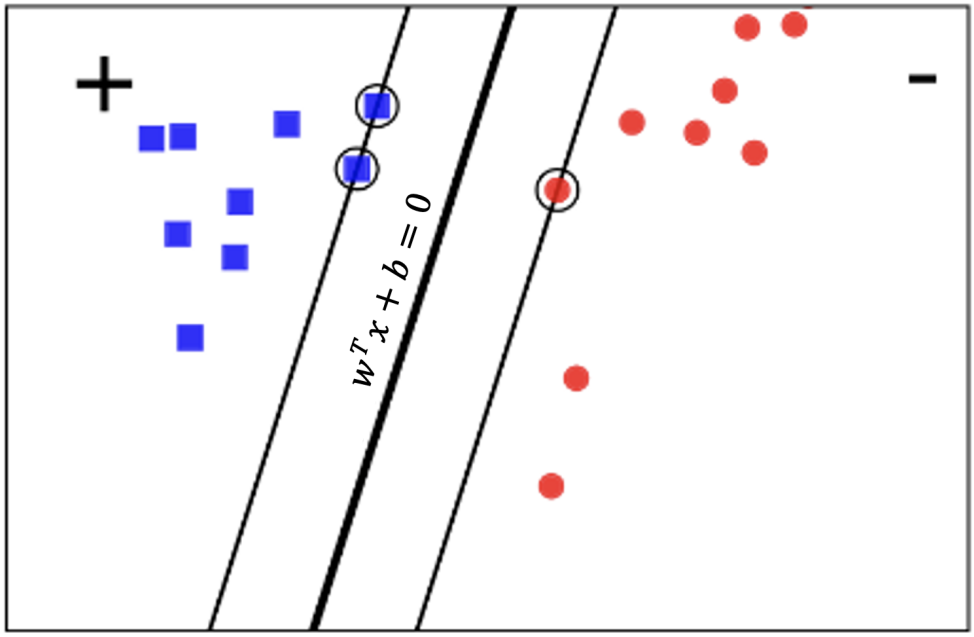

Figure 1: Example of the problem on 2-dimensional data.

a) Given the classifier as hyperplane defined by w and b, the margin is calculated as the shortest distance from a point to that plane (regardless of which class the point belongs to). Derive that:

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

b) We note that if we make the rescaling $w \rightarrow kw$ and $b \rightarrow kb$ then the distance from any point x_i to the decision surface is unchanged which means that the magin is unchanged. Based on this property, we can assume:

$$y_i(w^T x_i + b) = 1 \tag{1}$$

for the points that are closest to the separating hyperplane.

The fundamental idea of SVM is to find the hyperplane that maximizes the margin between the two classes now becomes maximizes $\frac{1}{||w||_2}$. This is equivalent to to an optimization problem that minimizes $||w||_2^2$ with a constraint (1).

Solve the following constraint-based optimization problem using the *Lagrangian* to find the *extremum*:

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

where λ_i is the Lagrange multiplier.

Hint: For this you have to first find w and b by setting the derivatives of $\mathcal{L}(\mathbf{w}, b, \lambda)$ w.r.t w and b to zero and then replace them in $\mathcal{L}(\mathbf{w}, b, \lambda)$.

c) What is the significance of the term λ_i in SVMs? *(max 2 sentences)*

Exercise 5.2 - Gradient Descent and Newton's Method (1+1+1 points)

In the optimization setting, Gradient Descent and Newton's Method are two commonly used iterative methods for finding a solution. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the following real function:

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

a) Compute the gradient $\nabla f(x)$ and the Hessian matrix $H_f(x)$. Find an $\hat{x}$ that minimizes f. Discuss the convexity of the function f and and determine whether $\hat{x}$ is guaranteed to be the global minimum of f.

b) Perform gradient descent on f in order to minimize it. Use a learning rate $\epsilon = 0.5$ and the starting point x_0:

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

Demonstrate 2 iterations, and on each iteration, make sure to show the following:

- The value of x
- The value of $\nabla f(x)$
- The value of $f(x)$

Once finished, compare the value of $f(x)$ to $f(\hat{x})$ and $f(x_0)$

c) Derive the update rule for Newton's method:

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

where $H_f(x_t)$ is the Hessian of function f at x_t. Starting from the same point x_0, use Newton's Method to find a solution.

Exercise 5.3 - Weight Space Symmetry (0.5 + 0.5 points)

Consider a Neural Network with a single hidden layer consisting of M neurons and *tanh* activation function. For any neuron in the hidden layer, simultaneous change of sign of input and output weights from the neuron leads to no change in the output layer therefore producing an equivalent transformation. Similarly, for any pair of neurons interchange of input weights between the neurons and simultaneous interchange of output weights produces an equivalent transformation:

a) Find the total number of equivalent transformation for the hidden layer.

b) Consider a deep neural network with N hidden layers. Each hidden layer consists of M_i neurons where $i \in 1, 2, ..., N$ and *tanh* activation function. Find the total number of equivalent transformations for the network

Exercise 5.4 - Feedforward Neural Network (3 points)

See the accompanying jupyter notebook.