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

Exercise Sheet 4

Machine Learning Basics

Deadline: 20.11.2024 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 any 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**. These instructions are mandatory. If you are not following them, tutors can decide not to correct your exercise.
Please note:

- For written assignments, please submit a pdf (written using Latex) with the **names, matriculation IDs and emails** of all team members for this part. In case you are not familiar with Latex, clearly written handwritten submissions are also accepted, but we strongly encourage pdfs written using Latex.
- There is no programming assignment in this exercise sheet.
- Submit the pdfs and notebooks together in a zip file in CMS. No need to submit any datasets.

Exercise 4.1 - Bias and Variance (6 points)

In this exercise, you will explore the relationship between bias and variance of an estimator, and how these concepts relate to model capacity and overfitting in practice.

a) Describe bias and variance of an estimator, and the tradeoff between these two. Relate bias and variance to overfitting and underfitting (*0.25 + 0.25 + 0.5 point*)

b) Describe the impact each of the following operations will have on your model's bias and variance. Indicate each operation's impact in the table with:

(i) increase (↑)
(ii) decrease (↓)
(iii) no change (–)

(iv) not enough information (NEI)

Answers without justification will receive 0 point. A well-justified answer that takes a creative detour from our solution will still earn full points – so go ahead, let your imagination run free (within reason)! (*3 points*)

	Bias	Variance	Justification
Increase the width of hidden layers in a neural network			
Increase the polynomial degree n in estimator $\hat{y} = b + \sum_{i=1}^{n} w_i x^i$			
Train a classification model on less data			
Add regularisation to the training objective of a classifier			
Remove only some non-support vectors in a SVM			
Decrease hypothesis space of the model			

c) Given a training set $S = \{(x_i, y_i)\}_{i=1}^{n}$, we assume that $y = f(x) + \epsilon$, where $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \tau^2$. We then train a $\hat{f}_n$ on training data S. Now, consider a test example (x_{test}, y_{test}), prove the following:

$$MSE(y_{test}, \hat{f}_n) = Var(\hat{f}(x_{test})) + Bias^2(\hat{f}(x_{test})) + \tau^2$$

Write down every step of your proof clearly and state <u>all</u> assumptions (*2 points)*

Exercise 4.2 - Interpretations of MLE (3 points)

This exercise is based on Chapter 5, Deep Learning book, Ian Goodfellow and Yoshua Bengio and Aaron Courville, and extends your knowledge beyond what was covered in the lecture. Please read section 5.5 and 5.6 on MLE and Bayesian Statistics.
In this exercise, we consider a set of m training examples $\mathbb{X} = \{x^{(1)}, \cdots, x^{(m)}\}$ i.i.d. from an unknown distribution $p_{data}(\boldsymbol{x})$. Denote $\hat{p}_{data}(\boldsymbol{x})$ as the empirical distribution of the data $\mathbb{X}$. Let $p_{model}(\boldsymbol{x}; \boldsymbol{\theta})$ be a parametric family of probability distributions over the same space indexed by $\boldsymbol{\theta}$. Our goal is to optimise for $\boldsymbol{\theta}$ such that probability of the real data is maximised, or formally:

$$\boldsymbol{\theta}_{ML} = argmax_{\boldsymbol{\theta}}\, p_{model}(\mathbb{X}; \boldsymbol{\theta}) = argmax_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{model}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \tag{1}$$

Write down every step of your proof clearly and state <u>all</u> assumptions in each of the following questions:

a) *MLE as minimising Kullback–Leibler (KL) divergence*: The goal of MLE is to "try our best" to make the model distribution p_{model} match the empirical distribution $\hat{p}_{data}$. This can also be understood as minimising the dissimilarity between these distributions, measured by KL divergence. To see this, prove the following:

(i) State a potential issue with the product $\prod_{i=1}^{m} p_{model}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$in equation (1). Convert this product into a sum using the logarithm, and argue why this transformation preserves the same argmax.

(ii) Formally, KL divergence is defined as $D_{KL}(P||Q) = \sum_{i=1}^{m} P(x^{(i)}) \log\left(\frac{P(x^{(i)})}{Q(x^{(i)})}\right)$. Prove that MLE of $\boldsymbol{\theta}$ is equivalent to minimising KL divergence between $\hat{p}_{data}$ and p_{model}.

b) *Linear Regression as MLE*: We reexamine linear regression through the lens of MLE. For a given input $\boldsymbol{x}$, our model now yields a conditional distribution $p(y|\boldsymbol{x})$ instead of a single prediction $\hat{y}$. Demonstrate, step by step, how linear regression with mean squared error can be derived as an MLE procedure, specifying all assumptions made.

c) *MLE and Training data:* Discuss how m affects MLE of $\boldsymbol{\theta}$, i.e., if m is small and if $m \to \infty$, and consider how the amount of training data influences overfitting/underfitting, bias, and variance in the estimator. Provide detailed reasoning to support your answer.

Exercise 4.3 - Maximum A Posteriori (MAP) Estimation (1 points)

This exercise is based on Chapter 5, Deep Learning book, Ian Goodfellow and Yoshua Bengio and Aaron Courville, and extends your knowledge beyond what was covered in the lecture. Please read section 5.6.

Given a training set $S = \{(x_i, y_i)\}_{i=1}^{n}$, we consider the least square loss with L2 regularisation defined as follows:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T x_i)^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

Show that minimising $\mathcal{L}(\boldsymbol{\theta})$ can be justified as a MAP procedure by assuming a a Gaussian prior on the weights, i.e., $\mathcal{N}(0, \frac{1}{\lambda}\boldsymbol{I}^2)$.