

Machine Learning Project 2  
*Supervised Learning: Classification and Regression*

October 7, 2025

Group 94

Name	Student number	Task	Vincent	Diego	Albert
Vincent Van Schependom	s251739	Section 1	0%	0%	0%
Diego Armando Mijares Ledezma	s251777	Section 2	0%	0%	0%
Albert Joe Jensen	s204601	Section 3	0%	0%	0%
		L <sup>A</sup> T <sub>E</sub> X	0%	0%	0%

(a) Group members.

(b) Contributions & responsibilities table.

Table 1: Group information & work distribution.

Introduction

The objective of this report is to apply the methods that were discussed during the second section of the course *Machine Learning* [1] to a chosen dataset. The aim is to perform relevant regression and classification to the data.

The particular dataset that is being investigated is – just like in Project 1 – the *Glass Identification* dataset from 1987 by B. German [2]. Table 1a lists our full names and student numbers, while Table 1b shows an overview of the contribution of each team member.

Contents

1	Regression	2
1.1	Linear regression . . . . .	2
1.1.1	Aim . . . . .	2
1.1.2	Regularization . . . . .	2
1.1.3	Model interpretation . . . . .	2
1.2	Regularized linear regression vs an Artificial Neural Network . . . . .	2
2	Classification	2
2.1	Introduction . . . . .	3
2.2	Logistic regression vs [...method 2...] . . . . .	3
2.3	Interpretation of the LR model . . . . .	3
A	Repository and supplementary materials	3

# 1 Regression

## 1.1 Linear regression

### 1.1.1 Aim

[TODO: Explain what variable is predicted based on which other variables and what you hope to accomplish by the regression. Mention your feature transformation choices such as one-of-  $K$  coding. Since we will use regularization momentarily, apply a feature transformation to your data matrix  $X$  such that each column has mean 0 and standard deviation 1.]

[TODO: ...]

### 1.1.2 Regularization

[TODO: Introduce a regularization parameter  $\lambda$  as discussed in 14 of the lecture notes, and estimate the generalization error for different values of  $\lambda$ . Specifically, choose a reasonable range of values of  $\lambda$  (ideally one where the generalization error first drop and then increases), and for each value use  $K = 10$  fold cross-validation (algorithm 5) to estimate the generalization error. Include a figure of the estimated generalization error as a function of  $\lambda$  in the report and briefly discuss the result.]

[TODO: Figure]

[TODO: Reference figure]

[TODO: ...]

### 1.1.3 Model interpretation

[TODO: Explain how the output,  $y$ , of the linear model with the lowest generalization error (as determined in the previous question) is computed for a given input  $x$ . What is the effect of an individual attribute in  $x$  on the output,  $y$ , of the linear model? Does the effect of individual attributes make sense based on your understanding of the problem?]

[TODO: Final equation]

[TODO: ...]

## 1.2 Regularized linear regression vs an Artificial Neural Network

[TODO: Rewrite this (concisely!) so that it isn't an exact copy of the assignment]

In this section, we will compare three models: the regularized linear regression model from the previous section, an artificial neural network (ANN) and a baseline. We are interested in two questions: Is one model better than the other? Is either model better than a trivial baseline?. We will attempt to answer these questions with two-level cross-validation.

[TODO: Create the table as in the assignment (Vincent)]

[TODO: Write the accompanying text on how we retrieved the data in the table.]

[TODO: Write out the statistical comparisons using data from the table.]

[TODO: TABLE: Include p-values and confidence intervals for the three pairwise tests in your report.]

[TODO: Conclude on the results from the values in the table and reference the table.]

# 2 Classification

[TODO: Choose method 2: ANN, CT, KNN, NB]

## 2.1 Introduction

[TODO: Explain which classification problem you have chosen to solve. Is it a multi-class or binary classification problem?]

## 2.2 Logistic regression vs [...method 2...]

[TODO: Rewrite the assignment below such that it (consisely!) states what we will do in this section.]

We will compare logistic regression, method 2 and a baseline. For logistic regression, we will once more use  $\lambda$  as a complexity-controlling parameter, and for method 2 a relevant complexity controlling parameter and range of values. We recommend this choice is made based on a trial run, which you do not need to report. Describe which parameter you have chosen and the possible values of the parameters you will examine. The baseline will be a model which compute the largest class on the training data, and predict everything in the test-data as belonging to that class (corresponding to the optimal prediction by a logistic regression model with a bias term and no features).

[TODO: Perform a statistical evaluation of your three models similar to the previous section. That is, compare the three models pairwise.]

[TODO: TABLE: Include p-values and confidence intervals for the three pairwise tests in your report.]

[TODO: Conclude on the results from the values in the table and reference the table.]

## 2.3 Interpretation of the LR model

[TODO: Train a logistic regression model using a suitable value of  $\lambda$  (see previous exercise). Explain how the logistic regression model make a prediction. Are the same features deemed relevant as for the regression part of the report?]

## Use of GenAI

...

## References

- [1] Tue Herlau, Mikkel N. Schmidt, and Morten Mørup. *Introduction to Machine Learning and Data Mining*. Technical University of Denmark (DTU), Lyngby, Denmark, 2023. Lecture notes, Fall 2023, version 1.0. This document may not be redistributed. All rights belong to the authors and DTU.
- [2] B. German. Glass Identification. UCI Machine Learning Repository, 1987. DOI: <https://doi.org/10.24432/C5WW2P>.

## Appendix

### A Repository and supplementary materials

The full notebook, scripts, and generated figures for this project are available in the project repository:

[https://github.com/schependom/DTU\\_machine-learning-projects/tree/main](https://github.com/schependom/DTU_machine-learning-projects/tree/main)

This repository contains the data-loading and analysis code that produced the tables and figures cited above (see the `figures/` folder for the PDF outputs referenced in the report).