Technical University of Denmark

02452 Machine Learning
DTU Compute
Fall 2025

DTU

Machine Learning Project 2

Supervised Learning: Classification and Regression

October 21, 2025

Group 94

Name	Student number
Vincent Van Schependom	s251739
Diego Armando Mijares Ledezma	s251777
Albert Joe Jensen	s204601

(a) Group members.

Task	Vincent	Diego	Albert
Training & test loops	0%	0%	0%
Coding visualisations	0%	0%	0%
Section 3	0%	0%	0%
LaTeX	0%	0%	0%

(b) Contributions & responsabilities 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

Outer fold	ANN		Linear regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	500	2.607×10^{-2}	0.797	1.600×10^{-6}	1.970×10^{-5}
2	480	1.327×10^{-2}	2.84	4.249×10^{-7}	1.228×10^{-5}
3	230	3.239×10^{-2}	0.325	4.821×10^{-7}	4.482×10^{-6}
4	190	1.592×10^{-2}	0.167	9.633×10^{-7}	5.047×10^{-6}
5	120	1.030×10^{-2}	1.43	1.500×10^{-6}	3.080×10^{-6}
6	500	3.226×10^{-3}	0.01	2.142×10^{-6}	5.982×10^{-6}
7	120	1.603×10^{-2}	0.325	1.284×10^{-6}	8.186×10^{-6}
8	240	3.639×10^{-2}	0.01	1.018×10^{-6}	7.210×10^{-6}
9	290	6.221×10^{-2}	1.27	1.637×10^{-7}	1.946×10^{-5}
10	500	2.641×10^{-2}	3	5.225×10^{-7}	7.245×10^{-6}

Table 2: Summary of two-level cross validation for predicting [TODO: see Python script]. Hyperparameters and test errors E_i^{test} on $\mathcal{D}_i^{\text{test}}$ per outer fold i, for each of the three considered models.

1 Regression

[TODO: Make visualisations of the results in Table 2]

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 λ as discussed in 14 of the lecture notes, and estimate the generalization error for different values of λ. Specifically, choose a reasonable range of values of λ (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 λ 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]

Outer fold	[TODO: Decision Tree?]		Logistic regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	3	0.4091	0.336	0.2273	0.5455
2	6	0.2273	0.0785	0.3182	0.5000
3	15	0.3182	0.008 86	0.4091	0.8182
4	3	0.3636	0.0379	0.4545	0.9091
5	5	0.3333	0.008 86	0.2857	0.5714
6	6	0.4762	0.004 28	0.4286	0.8095
7	5	0.2857	0.004 28	0.3333	0.6667
8	10	0.2857	0.336	0.3333	0.7619
9	4	0.3333	0.004 28	0.3810	0.6667
10	7	0.3333	0.695	0.4762	0.8095

Table 3: Summary of two-level cross validation for predicting the glass type. Hyperparameters and test errors E_i^{test} on $\mathcal{D}_i^{\text{test}}$ per outer fold i, for each of the three considered classificationmodels. [TODO: Change λ to C?]

[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 re- gression 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 retreived 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]

[TODO: Make visualisations of the results in 3]

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 λ 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 λ (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`

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).