A black background with red and white triangles

Description automatically generated

02450 - Introduction to Machine Learning and Data Mining

Supervised learning: Classification and regression

Group 157

Even Johan Pereira Haslerud – s231473

Jørgen Finsveen – s231469

November 09, 2023

Table of Contents

[Author Agreement 3](#_Toc150798627)

[Appendix A: Exam Questions 3](#_Toc150798628)

[Question 4. Spring 2019 Question 20: 3](#_Toc150798629)

[Attributes in Table 1: 4](#_Toc150798630)

[Question 5. Spring 2019 Question 22: 4](#_Toc150798631)

[Question 6. Spring 2019 Question 26: 4](#_Toc150798632)

# Author Agreement

This document is the report for Project 2 in the course "Introduction to Machine Learning and Data Mining" during the Autumn Term of 2023 at Denmark's Technical University. The authors, Even Johan Pereira Haslerud and Jørgen Finsveen have contributed equally to the project's work.

|  |  |  |
| --- | --- | --- |
| **What type of work?** | **Even Johan** | **Jørgen** |
| *Regression, part a* |  |  |
| *Regression, part b* |  |  |
| *Classification* |  |  |
| *Discussion* |  |  |
| *Exam problem 1* |  | *X* |
| *Exam problem 2* |  | *X* |
| *Exam problem 3* |  | *X* |
| *Exam problem 4* | *X* |  |
| *Exam problem 5* | *X* |  |
| *Exam problem 6* | *X* |  |
| *Tables* |  |  |
| *Figures* |  |  |
| *Report* |  |  |

# 1 Regression Part A

The dataset

# Appendix A: Mandatory Exam Questions

## Question 4. Spring 2019 Question 20:

**Option D:**

We can easily identify the splitting rules after setting the decision tree combinations into the classification boundary. The nodes A and C are determined by , while the other nodes B and D are determined by . This information is shown in Figure 1 below:

**D:**

**B:**

|  |  |  |
| --- | --- | --- |
| 2 | 1 | 4 |
| 1 | 3 | 4  **C:** |

**A:**

*Figure 1: Combination of the classification boundaries and decision tree.*

### Attributes in Table 1:

* **Number 1:** Congestion level 1
* **Number 2:** Congestion level 2
* **Number 3:** Congestion level 3
* **Number 4:** Congestion level 4
* **Nodes – A, B, D, and C:** These are the nodes from the structure of the given decision tree.
* **:** Is the x-coordinate for PCA 1 in Figure 4 of the Classification boundary.
* **:** Is the y-coordinate for PCA 2 in Figure 4 of the Classification boundary.

## Question 5. Spring 2019 Question 22:

**Option C:**

Both artificial neural network (ANN) and logistic regression models require performing training and testing on the same outer and inner folds. As a result, λ, and have a total of 5 possible values for both training and testing. This requirement affects the total time taken for the process, which can be calculated as follows:

This timing calculation is essential and must be carried out once for each of the five outer folds to identify the optimal model (which is the one with the lowest generalization error, ). Furthermore, it is necessary to select the best model from among the four inner folds. This selection process must be repeated five times, leading to the equation:

This ensures that each fold contributes to the determination of the most efficacious model, considering the time required for both training and testing.

## Question 6. Spring 2019 Question 26:

**Option B:**

The first step of resolving this question is to put the value of b and into the given equation to compute the , for respectively. Then we need to put the value of , for into the given softmax function to calculate the , for . According to the calculation of the option B, the approximate probabilities of the 4 classes are 0.05, 0.06, 0.15, 0.73. The class has the largest probability, so the option should be B.