Module Big data & Small data
Data Collection and Machine Learning
HTE1

Lecturers: Dixon Devasia and Marijn Jongerden
Deadline: April 5th, 2024

Part 1 linear regression with regularization(40%)

Prediction of heating demand of a house

The dataset to be used for this part is given in the file `energy_efficiency_data_heating_load.csv`. The data set is based on the data obtained from [1]. This dataset has been generated with the Ecotect simulation tool. With this tool, the heating demand of a building can be analyzed. A dataset has been created that gives the heating load of a building for a set of key building properties, such as wall area, roof area and glazing area.

In the this assignment we want to create a model using linear regression with regularization to predict the heating load of a house.

Question 1: Data preparation

Go through the data set. Make decisions and implement steps for pre-processing of the data in preparation for the implementation of the logistic regression algorithm. The following points may help you:

- Is normalization of the features necessary?
- Is there a problem of outliers in the dataset? If so, how will you solve this problem?
- Can quadratic terms help with the accuracy of the algorithm in this problem?
- Selection and reasoning of a validation scheme and division of data for training and validation sets.
- You can implement other pre-processing steps that you find important.

Question 2: Implementation linear regression with regularization

a. Implement a gradient decent algorithm for linear regression with regularization parameter and create a model for predicting the heating load. Make decision on the selection of the learning rate and regularization parameter, the features. Motivate your choices.
b. Why is regularization used? Is it really required to be implemented for this group of features for this project? Why or why not?
c. Validate the model using the validation set and evaluate the performance of the model. How well does the model predict? Can you make some comments on the usability of the model based on the observed performance?

Part 2 artificial neural network(60%)

Prediction of drive failures

In this assignment we use the dataset `drive_diagnosis_NN.csv`. This dataset is based on a dataset on sensorless drive diagnosis, provided by Martyna Bator et al. through the UCI Machine Learning Repository [2]. This data can be used to create a model that predicts whether an electric drive in a production plant may be faulty or not, and which failure mode is present.

The features in this data set are based on the electrical current signals measured at the drive. No additional sensors, such as vibration sensors, have been used. In a test bench, see Figure 1, the current signals for various faulty and properly functioning drives have been recorded. From the time series data the features in the dataset have been extracted by using a so-termed Hilbert-Huang transform, and computing the statistical properties of the obtained intrinsic mode functions and the residuals [3]. This process yields the 48 features recorded in the provided data set.

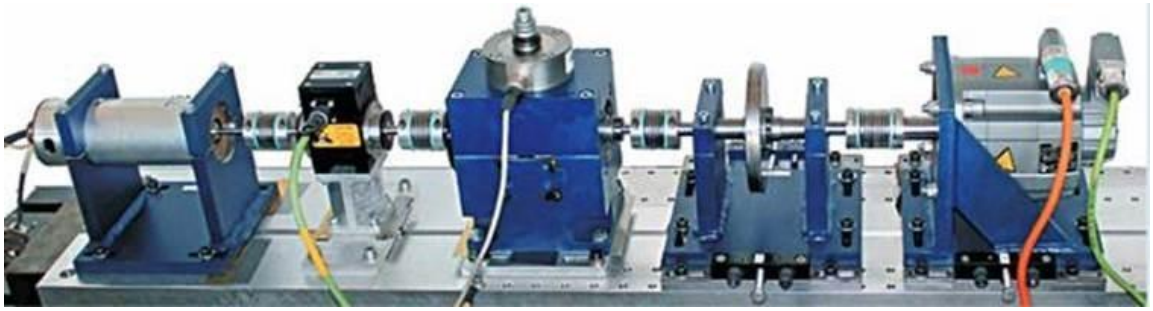

Figure 1: Mechanical structure of the test stand with test motor, measuring shaft, bearing module, flywheel, load motor (from left to right) [4].

The output for each sample is the Class-ID, which represents the failure mode of the system. In the dataset, 7 different classes are present. Each class represents a different mode of the system. Class 1 represents the case of no failure. Classes 2 to 7 represent different failure modes, where there is a shaft misalignment (SM) , axle inclination (AI), bearing failure (BF) or a combination of these failures. The table below gives an overview of the different classes.

Table 1: Indication of the failures present for each classID [4].

Class-ID	BF	AI	SM
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	1	0	1
6	1	1	0
7	1	1	1

Question 3: data preparation

Check the new data set and apply the appropriate preprocessing steps for creating and optimizing a neural network model that can identify the specific failure class. Document the steps you have taken and give your motivation for your choices.

Question 4: neural network

Based on the example code from “Example_code.zip”, implement a neural network learning algorithm, with the network configuration of 3 layers (input, output, and one hidden layer that comprises of 2 neurons.) Evaluate its corresponding learning and prediction performance.

Question 5: network optimization

Carefully consider the content regarding ‘Bias vs. Variance’. Argue whether the network configuration from (question 4) is 'underfitting', 'overfitting', or 'just right' for this particular dataset. If an improvement should be possible, modify the network’s configuration and/or the learning process to improve its prediction performance. Describe and justify your choice of implementation(s). Evaluate the quality of your final model.

Hand in

Use the “*Handin-app*” to submit a Word or pdf document where you have discussed your results and a zip-file that contains your MATLAB code, including the provided script and data files. Please share the submission to both or either: Marijn Jongerden and Dixon Devasia.

References

[1] A. Tsanas en A. Xifara, „Energy Efficiency,” 2012. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/energy+efficiency.

[2] M. Bator, „UCI Machine Learning Repository, Dataset for Sensorless Drive Diagnosis,” 24 February 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Dataset+for+Sensorless+Drive+Diagnosis. [Geopend January 2022].

[3] M. Bator, A. Dicks, U. Mönks en V. Lohweg, „Feature Extraction and Reduction Applied to Sensorless Drive Diagnosis,” in *22nd Workshop on Computational Intelligence*, Dortmund, 2012.

[4] M. Bator, *Private communication,* October 2021.