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

Lecturers: Dixon Devasia and Marijn Jongerden
Deadline: April 8th 2022

Prediction of drive failures

In the first part of the assignment we use the dataset `drive_diagnosis_log_reg.csv`. This dataset is based on a dataset on sensorless drive diagnosis, provided by Martyna Bator et al. through the UCI Machine Learning Repository [1]. This data can be used to create a model that predicts whether an electric drive in production plant may be faulty or not.

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 (three phases per drive). 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 [2]. This process yields the 48 features recorded in the provided data set. In the dataset the samples without failure are labeled with “0”, and with failure are labeled with “1”.

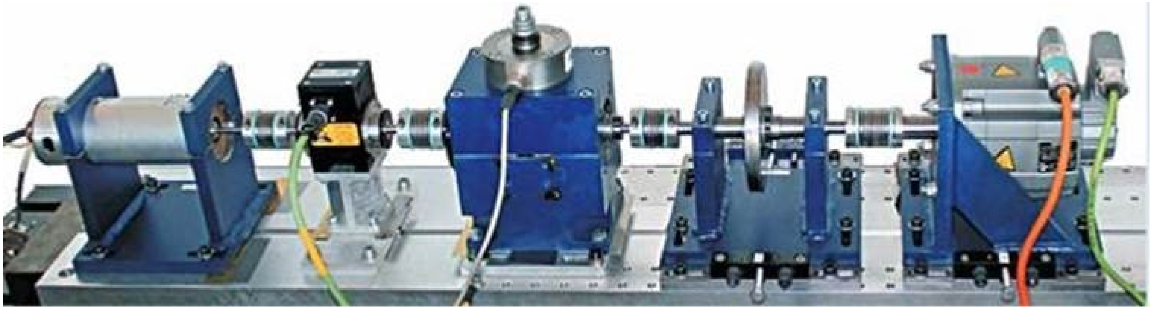

Figure 1: Mechanical structure of the test stand with test motor, measuring shaft, bearing module, flywheel, load motor (from left to right) [3].

Part 1 logistic regression with regularization (40%)

In the first assignment we want to create a model using logistic regression to predict whether a drive is faulty or not.

Question 1: data preparation logistic regression

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?
- Is there any missing values in the dataset? If so, how will you deal with it?
- Selection of a validation scheme and division of data for training and validation sets.
- You can implement other pre-processing steps that you find important.

In the report describe the steps you have taken and give a motivation for your choices.

Question 2: implementation logistic regression with regularization

a. Implement a gradient decent algorithm for logistic regression with regularization parameter and create a model. Make decision on the selection of the learning rate and regularization parameter. Motivate your choices.
b. Why is regularization used? Is it really required to be implemented for this group of input parameters?
c. Validate the model using the validation set and evaluate the performance of the model. How well does the model predict?

Part 2 artificial neural networks (60%)

In the zip-folder "`Example_code.zip`" you can find an implementation of a functional OCR (Optical Character Recognition) neural network learning algorithm. This implementation is based on the exercise given in the Mache Learning course by Andrew Ng on Coursera.

For this part of the assignment you can study this solution and use, adapt or modify the example code.

In the second dataset, `drive_diagnosis_NN.csv`, details are given on the type of failure of the drive. In the dataset 11 different classes are present. Each class represents a different mode of the system. Class 1 represents the case of no failure. Classes 2 to 11 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 [3].

ClassID	BF	AI	SM
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	0	1	1
6	1	0	1
7	1	1	0
8	1	1	1
9	1	0	1
10	1	1	0
11	1	1	1

Note that some classes have the same combination of failures, the difference is that in class 5, 9 and 11, the failure is more severe.

Question 3: data preparation neural network

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: basic 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 your implementation(s).

References

[1] 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. [Accessed January 2022].

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

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

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.