# Data science 8 – introduction machine learning – final assignment

In this final assignment, the student shows that they are able to apply the knowledge from past weeks to a practical assignment

## Introduction

A driving shaft is driven by a motor that makes the shaft rotate at a constant speed of 2000 rpm. In addition, the shaft carries weight of more than 2700 kg (6000lbs). Data was collected during a run-to-failure experiment by two accelerometers for the x and y directions. Every 10 minutes during the experiment, 1 second of data is collected at a frequency of 20480 Hz. This means that each 1 second sample contains 20480 data points. Figure 1a shows a picture of this setup and Figure 1b shows a schematic illustration of the setup.

Afbeelding met tekst

Automatisch gegenereerde beschrijving Afbeelding met diagram, schets, Technische tekening, Plan

Automatisch gegenereerde beschrijving

Figure 1a: Setup of the run-to-failure experiment Figure 1b: Schematic representation of the setup

Roller bearings are used so that the shaft can rotate smoothly. These bearings wear out and often break down in industrial machines, often bringing the entire process to a standstill. To avoid downtime, it is often predicted how long the bearings can last or how much wear there is.

As shown in Figure 1b, there are 4 bearings for which data is collected. About bearing 4, we know what stage the bearing is at per sample. These stages are defined as follows:

1. Early
2. Normal
3. Suspect
4. Roll element failure (figure 2a)
5. Stage 2 failure (figure 2b)

For the other bearings, the stage at which they were per sample was not determined.

Afbeelding met cilinder, grond, metaal, zwart-wit

Automatisch gegenereerde beschrijving Afbeelding met tafelgerei, kom

Automatisch gegenereerde beschrijving

Figure 2a Roll element failure Figure 2b Stage 2 failure

## Data format

Each sample is in a separate csv file and each file contains 20480 data points about the acceleration of 2 bearings in both the x and y directions (bearing 1 and bearing 4). In addition, there is a bearing\_conditions.csv file showing what the stage of bearing 4 is for each file.

The files are sorted by time. File 0.csv is the first sample and file 1723.csv is the last.

Afbeelding met tekst, software, Multimediasoftware, schermopname

Automatisch gegenereerde beschrijving

## Report and code

At the end, you will hand in a report that answers all the questions (if asked). Make sure this report (written in LaTeX) sis structured well, all graphs have labels and titles, and any bit of code is explained well. Do not forget to add an introduction and a conclusion.

Hand in your code/notebook that you used to make these exercises. You are allowed to work with multiple files, but make sure that it is clear which file belongs to which exercise. Do not forget that your code is also graded on efficiency and readability. So make sure your code is well organized.

## Exercise 1

The main goal in this exercise is to make a Machine Learning model that predicts the degradation stage of bearing 4. We will guide you through the process of making different features and making models. However, the exact implementation and choices about the model are up to you. *Note: you are not allowed to use the index of the sample as feature.*

**a.** Load the first sample (0.csv) and visualize the acceleration in the x direction of bearing 4. Compare this visualization to 3 other samples from different degradation stages.

**b.** Make a function that, given an index, loads the corresponding sample and returns mean, standard deviation, and root mean squared of both x and y acceleration. The formula for the root mean squared is

Furthermore, add two features of your choice. One of the features you could add is the spectral flatness. This feature tells something about the flatness of the power spectrum, which could indicate something about the amount of impurities in the bearings.

Clearly state which features you added in your report and explain why you think these features might be helpful to predict the degradation stage of bearing 4. Make sure you support your explanation with visualizations.

**c.** Apply the function from **b.** to the entire dataset and store your results in a dataframe. Perform some explanatory data analysis to get some insight on which classification models might be suitable for predicting the degradation stage.

**d.** We have used the standard deviation and the rms. However, these features are highly correlated. Explain why these features are highly correlated (either in words or mathematically).

**e. (1)** Make one or several machine learning models to predict the degradation stage based on the features calculated in the previous exercises. Explain why you chose these models and their parameters.

**e. (2)** Pick your best model based on your evaluation criteria.

**f.** Try to improve your model from **d.** Explain in your report why you tried these improvements, what their results were, and if this was what you expected.

**g. (discussion)** Evaluate your model from **f.** and answer the following questions:

**g. (1)** Why is the accuracy not a good metric for this model? What would be a good solution for this problem?

**g. (2)** How did you make sure that your model did not overfit?

**g. (3)** Would you use this model in a production setting at a company?

**g. (4)** What would you do to improve the model (for example, if you had more time)?

## Exercise 2.

The drive shaft shows wear despite bearing and lubrication. In order to study the degree of wear more closely, the radius of the drive shaft is recorded once an hour at a certain spot during 1000 hours of operation. It turns out that the radius of the shaft wears down at certain times and then again shows little or no wear for a certain period of time.

In the file shaft\_radius.csv you will find the measured radius (m) of 1000 measuring moments.

**a.** We want to make a regression model to predict the wear of the drive shaft. First, split the dataset in a train and test. Use the train set to plot the data in a suitable manner.

**b.** Make a regression model to predict radius of the drive shaft based on the time that the drive shaft had been running. Explain why you chose this regression model.

**c.** Plot the predictions of the regression model with the test data.

**d. (discussion)** Evaluate your model from **b.** and answer the following questions:

**d. (1)** What metric did you use for evaluating your model? Explain the pro’s and cons.

**d. (2)** Explain why you think this model is reliable.

**d. (3)** What would you do to improve the model (for example, if you had more time)?

## Exercise 3

The stages of degradation of bearing 4 are determined after the experiment was done. After this experiment, they analyzed the data and determined for each sample what the stage was. We know that bearing 4 failed completely, however this does not have to be the case for bearing 1, 2, and 3. For example, bearing 1, 2, and 3 did not break because the scientists stopped the experiment after bearing 4 broke down.

The same can be said about all the stages of bearing 4. There is a high chance that the other bearings follow different stages than bearing 4. In this exercise, we try to determine the number of stages that bearing 1 went through.

**a.**  Make several features that could tell something about the degradation stage of bearing 1. Explain why you chose these features *Hint: use/modify steps from exercise 1.*

**b.** Perform some explanatory data analysis on these features. Make use of various data visualizations and use the time component given (i.e. the index of the sample). Carefully document your findings in your report.

**c.** From your conclusions of **b.** perform a cluster analysis where each cluster corresponds to a stage of degradation. In this exercise, you are allowed to use the index as a feature.

**d.** Interpret each stage/cluster and give a suitable name (just like the stages for bearing 4).

**e. (discussion)** Evaluate your cluster analysis from **c.** and answer the following questions:

**e. (1)** Do you expect that the bearing will actually go through the definition you defined in **d.**

**e. (2)** Do you think cluster analysis is a suitable method to attack this problem?

**e. (3)** What would you do to improve the model (for example, if you had more time)?

# Appendix – feature examples

|  |  |  |
| --- | --- | --- |
| Name | Formula | Remarks |
| (arithmetic) mean |  |  |
| Standard deviation |  |  |
| Root mean squared |  |  |
| Fast Fourier transform | From the frequency domain, we can calculate the power spectrum as follows: | Numpy has an FFT algorithm:  [numpy.fft.fft — NumPy v1.26 Manual](https://numpy.org/doc/stable/reference/generated/numpy.fft.fft.html) |
| Spectral Flatness |  | <https://en.wikipedia.org/wiki/Spectral_flatness> |