Data Normalisation

Description:

Quite often, when working on a given dataset, you need to normalise your data as a part of data pre-processing. That is to scale all features of the dataset to either in a range with the same magnitude or with the standard normal distribution.

In this competition, each group needs to finish the following three tasks:

1. write your own Python function in Jupyter Notebook to normalise a dataset in a matrix format to the range of [-1, 1].
2. Download *hcvdat0.csv*, which can be accessed from the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/00571/
   i) remove rows containing NA values
   ii) exclude the first four columns from the data
   iii) test your code on the dataset
3. write a short paragraph (no more than 200 words) to explain why normalisation may be important for many machine learning applications.

Timeline

Noon 19/11/2020 final submission deadline

Submission instruction

1. Each group prepares one and only one submission.
2. The submission including solutions/answers to those three tasks should be a Jupyter Notebook submission.
3. The submission should be done via Canvas.

Evaluation

Submissions will be judged by the panel consisting of 3 tutors: Emil Dmitruk, Chloe Zhuge and Felix Riegler, with Felix Riegler being the chair, based on the following criteria:

1. Group collaboration: Do group members participate actively? Are discussions within the group supportive and efficient? Panel members will be monitoring discussions within each group in our module site on Canvas.
2. Efficiency of the code
3. Clarity of the writing