Group project 2

GRA4157

October 20, 2023

Machine learning

In this project, you will analyze a dataset of choice using machine learning. You are free to choose the dataset, and it does not have to be the same data set you used for visualization. If you are uncertain whether your data set suits machine learning or if you need supervision or assistance, you can contact me at steven@simula.no. Define one or more questions that you want to answer, which should be answered during your presentation. You will work in groups of 1-3 students and may use the same groups as in the previous project if you want to. Your efforts will be presented in a 10-20 minute presentation, depending on the group size, just like the last project. Every group member needs to contribute equally to the presentation, but it is okay if each group member has responsibility for different tasks within the project. You will receive feedback on the following:

- The question you are trying to answer.
- Did you find a data set suitable for machine learning?
- The raw data should be presented in an easily readable format. Table, map, statistics. Keep this section short in the presentation.
- Do you perform proper evaluation of the algorithm's performance?
- The essence of machine learning (ML) is to predict output given some input data. For example: If you have data from the housing market, you may have available the listed price, the size, number of rooms, size of the property, renovation year and coordinates/position of the house. Given enough entries in the dataset, a ML algorithm can find the most important factors (features) for the house price, and be able to predict a price if you give it all the other parameters. This type of ML can be done on most datasets. You can thus train the data on around 60–70% of your data, and test if the algorithm can predict prize on the with the rest of the data.
- Examples: Predict price of a wine bottle given large amounts of other data. Predict the usage of a bike station given it's coordinates (or lots of ohter data). Predict the number of vacant seats on a flight given all other data. Predict the number of private, for-profit and public universities in a state given a states GDP and population. Predict university rank based on e.g. number of students, number of professors, budget, citations per year (and so on).

The group presentations will be held Friday 3rd of November. The presentations are not graded, but you will use material from the presentation to write part of a report for the final assignment/exam in the course.