INFO284 Machine Learning Exam, spring 2021

Delivery date: May 11th 2021, 14:00

Format: Jupyter notebook (ipynb-file) containing runnable Python code, documentation and reflections on process and result.

Word limits: The total text parts should not be more than 3000 words. There are no limits on Python code size.

Task

At kaggle.com you will find a data set that contains data about property sales in New York. The link to the data is https://www.kaggle.com/new-york-city/nyc-property-sales.

You are supposed to **build at least five machine learning models from these data** to predict or classify one relevant target feature for new data points. You can choose target feature yourself, but sales price is perhaps the most suitable. You may also reduce the number of data points somewhat by focusing on only specific meaningful parts of the data. Or perhaps you will try dimension reduction.

You may take inspiration from the example code at kaggle.com or other web sites when building models for these data, but **if you do you need to refer to these examples**, and you need to explain how you used and extended these approaches in your own solution.

In particular**, you shall build one or two neural network/deep learning model for these data**. If you build two neural network models they need to be significantly different in approach.

You shall deliver code in the form of a **well commented Jupyter notebook**. This code needs to run on the original data set, so any preprocessing you choose to do needs to be programmed in Python and included in the notebook. The code shall in the end return the results of your experiments with your five chosen models and data sets. You need to explain

- Important and relevant properties of the data
- how you preprocessed data like which features you selected, which data points you dropped, did you go for a subset of the data, did you do dimension reduction, how you reformatted data, etc.
- how you decided on parameters for your machine learning models, did you use any regulation techniques
- how the five methods were measured and compared to each other

Finally, as a concluding comment in the Jupyter notebook, you need to write a summary of your results, and discuss consequences of such results.

It is not necessarily so that high scores for machine learning models will give a good grade on your report, or vice versa, low scores a bad grade. What counts, is a well-argued, well described and smart machine learning investigation from start to end.