INFO284 Machine Learning Exam, spring 2024

Delivery date: April 30th 2024, 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.

Machine learning on fisheries data

At MittUiB you will find a compressed (zip) file that contains a tabular data set that contains information about fishery operations in Norwegian waters. They are collected by the Norwegian Directorate of Fisheries.

The dataset is reduced from a set that contains data about all the activities of a fishing vessel. This data set only reports the fishing operations data. Several of the original columns have also been removed from the dataset. The names of the columns and textual data are in Norwegian. Documentation about the features can be found in two files from the Directorate of Fisheries, also to be found in the zip file. They are in Norwegian, but assuming you have access to current translation software, you should be able to handle that if you do not know Norwegian. In the zip file you will also find a spread sheet containing a translation of terms in the data to English.

The data you will get mainly consists of all reported fisheries operations by Norwegian fishing vessels longer than 15 meter in 2018. A report is called a daily catch activity (DCA) and contains the following information:

- MessageID
- Data about when the activity was reported.
- Data about the start of the fishing activity.
- Data about the stop of the fishing activity.
- Data about the fishing gear.
- Data about the fish caught.
- Data about the fishing vessel.

There is a lot of redundant information in the data. Each DCA report consists of one or (normally) more rows in the data. The rows belonging to the same **catch operation** have the same message ID AND the same start and stop times. The only variation in the rows for one DCA is the start and stop times and the data about the fish caught. In the Directorate of Fisheries one such row is often called a line. Note that numbers are in Norwegian format, i.e., using comma as decimal separator.

Your tasks are:

a) **Preprocessing:** Get to understand the data and remove columns and rows that you do not find useful for your machine learning models. It is possible to for example focus on one type of gear, one or a few species, group categories of species, etc. etc. This will help to reduce

the data set you are working on. Understanding data and preparing them for model building is a main task of machine learning.

b) Supervised learning**: Build at least three machine learning** models to predict or classify **catch** data related to a fishery operation, i.e. species caught, amounts of one fish species (sums of amounts is also possible), or other features related to fish catches. **One of the models needs to be a deep learning model**.

c) Unsupervised learning: **Build a clustering model** for the data set. You may use a different preprocessing for this task than for the data in task b).

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 chosen models. You need to explain

- Important and relevant properties of the data
- how you preprocessed data like which features you selected, did you do dimension reduction, how you reformatted data, etc.
- how you decided on parameters for your machine learning models,
- if you used any regularization techniques? In case how.
- how the methods were measured and compared

Please inform about any special Python libraries that need to be installed to make your code runnable.

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. The problem may in fact be of such nature that it is not possible to get really good results on these data.

Final note: These data are prepared for this course and are shared with you in confidence that you do not share them in any way but use them only for the purpose of this exam.