INFO284 Machine Learning
Group Exam Spring 2023

Final delivery date: May 3rd 2022, 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.

Voluntary deliveries: Feb 17th, Mar 10th, and Mar 31st at 14:00

It necessarily that a high score for the machine learning models will give a good grade on your report, or vice versa, that low scores will give a bad grade. What counts, is a well-argued, well described and smart machine learning investigation from start to end. It may in fact be so that it is not possible to get good results on these data.

To get a good grade, you need to explain

a) Important and relevant properties of the data .
b) How you preprocessed data like which features you selected from the data set, did you leave any out, did you do dimension reduction, how you reformatted data, etc.
c) How you decided on the parameters of your machine learning models, did you use any regularisation techniques.
d) 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.

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.

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.

Task 1: Machine learning on tabular mushrooms

In the folder at MittUiB you will find a tabular data set that contains information about 8124 mushrooms. These data can also be found here https://archive.ics.uci.edu/ml/datasets/Mushroom.

You are supposed to

a) Build at least 3 different machine learning models to predict the edibility of the mushrooms in the UCI data set. You may take inspiration from example code from the web when building models for these data, but if you do you need to refer to the sources that you used. You also need to explain how you used and extended from the examples in your own solution. You should follow best practices like splitting the data into train, test, and dev.
b) Explain which feature are the most indicative of edible mushrooms.
c) Discuss the results of your best performing model. Would you trust it to classify a mushroom for you to eat? Why do you come to that conclusion? Do you think the data set is sufficient to classify the mushroom as edible or poisonous? Why do you believe your best performing model is the best model? Why do the other models perform worse?

Task 2: Sentiment analysis

In the folder at MittUiB you will find a data set that contains around 10000 sentences categorized to positive, neutral, and negative. The data set can also be found here: https://github.com/ltgoslo/norec_sentence

You are supposed to

a) Build a model that can classify the sentences with their sentiment.
b) Describe how you prepared the data for the model.
c) Discuss your results.

Task 3: Convolutional neural networks

In the folder at MittUiB you will find a data set that contains the 60000 images from the CIFAR-10 data set (https://www.cs.toronto.edu/~kriz/cifar.html).

You are supposed to

a) Train a convolutional neural network as a binary classifier of one category (by your choice) in the data set. In other words, the model should classify if an image is of that category or not. To do this you are can use a pre-trained convolutional neural network (of your own choice), but if you have available computational power, you may of course try to build you own complete CNN.
b) Find a new image (or take one yourself) of the category you chose, show how you would use your model to classify the new image.

Some links to information about pre-trained CNN:

- https://towardsdatascience.com/4-pre-trained-cnn-models-to-use-for-computer-vision-with-transfer-learning-885cb1b2dfc
- https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751
- https://medium.com/@mikhaillenko/instructions-for-transfer-learning-with-pre-trained-cnns-203ddaefc01
- BOOK: F. Chollet. Deep Learning with Python. Ch. 5.3 Using a pretrained convnet. p.143-159