Introduction to Data Science WiSe 2017/18
Project: Classification of Chest X-Rays

Supervisor: Christian Holme
E-Mail: christian.holme@med.uni-goettingen.de
Tel.: 0551 39 20873
Office: MRT-Forschungsgebäude am Osteingang, 0.B5 601

Description:

In this project, your task is analyzing the most common form of medical images: X-ray images of the chest. Normally, these images are read by a trained radiologist. Your task is to train a classifier to "read" these images as well.

Data:

The dataset contains 121120 X-ray images of 30805 unique patients. All images are in `png`-format and scaled to 1024×1024 pixels. An additional file contains image labels for 14 different pathologies (Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural thickening, Cardiomegaly, Nodule, Mass and Hernia) or the label "No Finding". Of course, multiple pathologies can apply to each image.

Furthermore, there are lists of images which have been used previously for training and for testing a classifier.

Tasks:

1) Download the dataset from https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345 and read the accompanying paper (`ARXIV_V5_CHESTXRAY.pdf`)
2) Using the provided labels (`Data_Entry_2017.csv`), collect some statistics which describe and characterize the dataset. For example, how many images are there showing each of the pathologies? How many images show multiple pathologies? Which pathologies tend to occur together?
   i. Read up on the different pathologies and briefly describe how they appear in the image.s
   ii. Show example images for each pathology.
3) Think about a suitable data format to present the images in memory. Which challenges does this dataset pose there?
4) Train a classifier of your choice to automatically label the images with the pathologies that are visible. You can use the provided lists of training and testing data. Note that this is a multi-label classification problem.
   i. Describe how you structure the training. Did you train on the entire training data set at once? If not, how did you partition the dataset?
   ii. Validate your classifier. How accurate is it?
5) Explore how the accuracy varies with different training strategies. Try, for example, a training data set which includes each pathology an equal number of times.
6) Let's assume that your classifier always labels the images correctly. Is this enough to replace a radiologist's reading of the X-ray images? Why or why not?