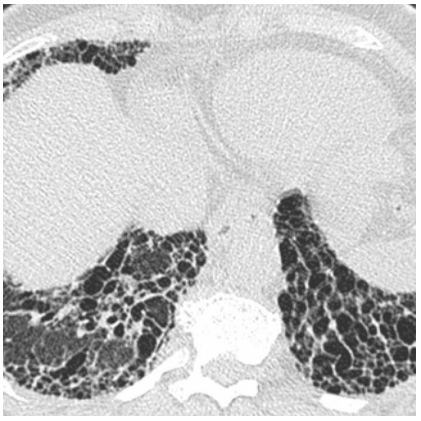
Machine Learning Engineer Nanodegree Capstone Proposal

Andreas Quauke

August 19th, 2020

Domain Background

The topic of my capstone project is based on a recent Kaggle competition that deals with a lung disease called *pulmonary fibrosis*. "Pulmonary fibrosis is a condition in which the lungs become scarred over time. Symptoms include shortness of breath, a dry cough, feeling tired, weight loss, and nail clubbing. Complications may include pulmonary hypertension, respiratory failure, pneumothorax, and lung cancer."[1]



Pulmonary fibrosis is a disorder with no known cause and no known cure, created by scarring of the lungs. People who get this disease would want to know their prognosis. And that is where a troubling disease becomes frightening for the patient: "Outcomes can range from long-term stability to rapid deterioration, but doctors aren't easily able to tell where an individual may fall on that spectrum. Current methods make fibrotic lung diseases difficult to treat, even with access to a chest CT scan. In addition, the wide range of varied prognoses create issues organizing clinical trials. Finally, patients suffer extreme anxiety—in addition to fibrosis-related symptoms—from the disease's opaque path of progression."[2]

In this capstone project - like in the respective competition on Kaggle - I would like to predict a patient's severity of decline in lung function based on a CT scan of their lungs and determine lung function based on output from a spirometer, which measures the volume of air inhaled and exhaled. The goal is to predict the patient's

[1] Wikipedia contributors. "Pulmonary Fibrosis." *Wikipedia*, 12 Aug. 2020, en.wikipedia.org/wiki/Pulmonary_fibrosis.

[2] OSIC Pulmonary Fibrosis Progression. (n.d.). Kaggle. Retrieved August 18, 2020, from https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression

Forced Vital Capacity (FVC) by classifying her CT image and analysing her prevailing FVC measurements. The FVC measurement shows the amount of air a person can forcefully and quickly exhale after taking a deep breath.[3]

My personal motivation is to help patients and their families better understand their prognosis when they are first diagnosed with this disease. Indeed, I have never worked in the medical field before, but this is exactly why I find it interesting and challenging to understand how machine learning techniques can help to make people's life better.

Problem Statement

The problem to solve is to predict a patient's FVC in weekly steps over a given period of time in the future. By assessing the patient's lung CT and her given FVC measurements of her past examinations, I get indicators if the disease is progressing, stabilizing or even getting better. The CT images can be classified in terms of progression of the disease, and the FVC history also gives indications of its recent development.
The FVC values are predicted in milliliters (ml) and therefore are clearly quantifiable and the quality of the predictions are measurable as well. As there are many patients affected by this disease, the problem is also replicable.

Datasets and Inputs

For this project I will use the dataset provided by Kaggle for their competition. It is publicly accessible on their website, where the structure of the data is also described:[4]

"In the dataset, you are provided with a baseline chest CT scan and associated clinical information for a set of patients. A patient has an image acquired at time Week = 0 and has numerous follow up visits over the course of approximately 1-2 years, at which time their FVC is measured.

- In the training set, you are provided with an anonymized, baseline CT scan and the entire history of FVC measurements.
- In the test set, you are provided with a baseline CT scan and only the initial FVC measurement. You are asked to predict the final three FVC measurements for each patient, as well as a confidence value in your prediction.

---

[3] FEV1 and FVC: What Do They Mean for You? (2018, February 11). Lung Health Instititute. https://lunginstitute.com/blog/fev1-and-fvc/

[4] Dataset OSIC Pulmonary Fibrosis Progression. (n.d.). Kaggle. Retrieved August 18, 2020, from https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression/data

Since this is real medical data, you will notice the relative timing of FVC measurements varies widely. The timing of the initial measurement relative to the CT scan and the duration to the forecasted time points may be different for each patient."

Description of the features

- `Patient`: a unique Id for each patient
- `Weeks`: the relative number of weeks pre/post the baseline CT
- `FVC`: the recorded lung capacity in ml
- `Percent`: a computed field which approximates the patient's FVC as a percent of the typical FVC for a person of similar characteristics
- `Age`
- `Sex`
- `SmokingStatus`

Solution Statement

The proposed solution to this problem consists of two steps:
Firstly, I apply an image classification algorithm that gives me an indication what the current status of the disease is, e.g. by measuring the degree of honeycombing[5] on the lung tissue or the lung volume. I can then add these newly discovered features to the existing dataset.
Secondly, I build a forecasting algorithm like XGBoost or a LSTM deep neural net to predict the time series for future FVC values based on the existing and newly added features.

Benchmark Model

For this particular problem, there is no available benchmark. However, I will define an evaluation metric, a modified version of Laplace Log Likelihood (cf. "Evaluation Metrics" below).

The minimum benchmark for my model to beat is to provide a better prediction than you would get by predicting the mean and standard deviation for each patient.

[5] Wikipedia contributors. (2017, November 4). Honeycombing. Wikipedia. https://en.wikipedia.org/wiki/Honeycombing

"-8.023 is the default score to beat while cross-validating models on train data. Any model scoring worse than this is not useful. You can get this default score to beat for each (fold of) validation data as well."[6]

Evaluation Metrics

This capstone project is evaluated on a modified version of the *Laplace Log Likelihood*. "In medical applications, it is useful to evaluate a model's confidence in its decisions. Accordingly, the metric is designed to reflect both the accuracy and certainty of each prediction. For each true FVC measurement, you will predict both an FVC and a confidence measure (standard deviation σ). The metric is computed as:

$$\sigma_{clipped} = max(\sigma, 70),$$

$$\Delta = min(|FVC_{true} - FVC_{predicted}|, 1000),$$

$$metric = -\frac{\sqrt{2}\Delta}{\sigma_{clipped}} - \ln(\sqrt{2}\sigma_{clipped})$$

The error is thresholded at 1000 ml to avoid large errors adversely penalizing results, while the confidence values are clipped at 70 ml to reflect the approximate measurement uncertainty in FVC. The final score is calculated by averaging the metric across all test set Patient_Weeks (three per patient). Note that metric values will be negative and higher is better."[7]

Project Design

Basic Workflow

The workflow of this capstone project is planned as follows:

1. Perform an exploratory data analysis (EDA) on the given dataset to get insights about which features could be interesting for the prediction.
2. Clean and preprocess the data if necessary.

---

[6] Vopani. (2020, July 17). OSIC: Understanding Laplace Log Likelihood. Kaggle. https://www.kaggle.com/rohanrao/osic-understanding-laplace-log-likelihood

[7] Dataset OSIC Pulmonary Fibrosis Progression. (n.d.). Kaggle. Retrieved August 18, 2020, from https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression/data

3. Train a base forecasting model based on the given feature data without the CT scans to establish a benchmark.
4. Build a lung CT classifier to get additional features from the lung CT scans. I would like to test different classifiers and masks to compare performance. Possible new features could be the progression of the disease or the lung volume of the patients.
5. Merge the additional data into the existing dataset and perform an EDA again.
6. Re-train the model from 3. with the additional features and compare the outcome
7. Fine tune the model's hyperparameters

Technologies & Libraries

The following technologies and libraries might be used in solving the problem:

- Programming language: Python 3.8
- Pandas
- Numpy
- Scikit Learn
- Pytorch
- JoHof/lungmask library[8]

[8] Hofmanninger, J. (n.d.). JoHof/lungmask. GitHub. Retrieved August 19, 2020, from https://github.com/JoHof/lungmask