Machine Learning Engineer Nanodegree

Capstone Proposal

Manuel Maqueda Vinas July 10th, 2020

Proposal

Pulmonary Fibrosis Progression (Kaggle Competition)

Domain Background

Pulmonary fibrosis, a disorder with no known cause and no known cure, created by scarring of the lung. Outcomes of the disease can vary from long-term stability to rapid deterioration, but doctors are not easily able to understand the exact outcome in advance.

In addition some of the current methods implies long times and substantial effort in order to produce an accurate prognose. This increases patients anxiety and delays the application of the right mitigation action.

Problem Statement

The goal of this exercise is to help doctors with predictions of pulmonary fibrosis progressions using CT scan images, metadata and FVC (force vital capacity) data as an input.

Datasets and Inputs

The datasets are provided by the Open Source Imaging Consortium.

The input data corresponds with a baseline CT scan on Week 0 and a series of FVC measurements that corresponds to follow-up visits of the patients in the next 1-2 years. Note also that some of the visits can be also negative in the case of patients data acquired before the baseline CT scan.

Input Data fields

- Patient- a unique Id for each patient (also the name of the patient's DICOM folder)
- Weeks- the relative number of weeks pre/post the baseline CT (may be negative)
- 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

CT scan

A baseline CT scan obtained in week 0. This image is stored in a DICOM subfolder with the Patient unique ID.

Solution Statement

The solution aims to predict the pulmonary fibrosis progression for the last 3 weeks of the patients' visit. In particular we will attempt to predict the FVC measurements for those visits but also a confidence metric (standard deviation), which should reflect both accuracy and certainty of the predictions.

Benchmark Model

There are other existing solutions that attempts to resolve this problem as part of the Kaggle competition. The ambition the ML model built as part of this project is contribute and improve other models of the competition.

Evaluation Metrics

The objective of the competition is to obtain good predictions for the follow-up patient's FVC measurements. In particular, the evaluation metric for those measurements is as follows:

$$\Delta = \min\left(\| FVC_{ture} - FVC_{predicted} \| , 1000 \right)$$

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

Project Design

First I will do some exploration in the given data in order to understand better the nature of it, to detect any imbalance or particularity.

Then to address the proposed solution I will start by implementing a linear regression neural network using the input data from each of the patient's visits and using

the suggested metric to optimise the network.

At the end I will dedicate some time to improve the first solution by analyzing the images and to include a CNN looking at some image properties.

I will also do some research in some of the public notebooks in the competition to re-use and improve suggested techniques.

Reference

- Kaggle Competition
- Scikit Linear Models