Machine Learning Engineer Nanodegree

Capstone Proposal

Aditya Mehta June 28th, 2019

Proposal

Domain Background

Domain is Healthcare and focussed on detecting blindness beforehand using eye images. This is provided as a competition on Kaggle by Aravind Eye Hospital in India- https://www.kaggle.com/c/aptos2019-blindness-detection/overview.

I am motivated to solve this problem statement as it would help in improving the lives of people by enabling proactive actions to prevent blindness.

Problem Statement

The objective of this competition is to detect diabetic retinopathy using images which Aravind technicians capture from rural areas of India.

Current State- Currently highly trained doctors review these images and provide diagnosis

Desired State- Scale the doctors efforts through technology by automatically screening the images for disease and provide information on how sever the condition may be.

Datasets and Inputs

We are provided with a large set of retina images taken using fundus photography under a variety of imaging conditions.

A clinician has rated each image for the severity of diabetic retinopathy on a scale of 0 to 4:

0 - No DR

1 - Mild

2 - Moderate

3 - Severe

4 - Proliferative DR

Like any real-world data set, we will encounter noise in both the images and labels. Images may contain artifacts, be out of focus, underexposed, or overexposed. The images were gathered from multiple clinics using a variety of cameras over an extended period of time, which will introduce further variation.

Solution Statement

My approach for solving this problem is build a Machine Learning model which would accurately predict the severity of the disease based on the training data and have high quadratic weighted kapp score.

Benchmark Model

The benchmark for me would be the human ratings which I would try to replicate with the machine learning model. I would also be comparing my results with some of the other participants' results to get an understanding of how accurate my model is.

Evaluation Metrics

Scoring metric is based on the quadratic weighted kappa, which measures the agreement between two ratings. This metric typically varies from 0 (random agreement between raters) to 1 (complete agreement between raters). In the event that there is less agreement between the raters than expected by chance, this metric may go below 0. The quadratic weighted kappa is calculated between the scores assigned by the human rater and the predicted scores.

Images have five possible ratings, 0,1,2,3,4. Each image is characterized by a tuple (e,e), which corresponds to its scores by Rater A (human) and Rater B (predicted). The quadratic weighted kappa is calculated as follows. First, an N x N histogram matrix O is constructed, such that O corresponds to the number of images that

received a rating i by A and a rating j by B. An N-by-N matrix of weights, w, is calculated based on the difference between raters' scores:

An N-by-N histogram matrix of expected ratings, E, is calculated, assuming that there is no correlation between rating scores. This is calculated as the outer product between each rater's histogram vector of ratings, normalized such that E and O have the same sum.

Project Design

I will be following the below steps-

- Going through the data and sample images to understand visual pattenrs
- Understanding the quadratic weighted kappa measure thoroughtly
- Exploratory data analysis to identify
    - Class Distribution between the five classes
    - In case if class imbalance is there, apply techniques like SMOTE
    - Identify faulty or errornous images to remove them from training set
- Try various algorithms on the data. Some of my initial thoughts are to try-
    - LogisticRegression
    - RandomForestClassifier
    - ExtraTreeClassifier
    - GradientBoosting Methods
    - Neural Networks like CNNs since they work well on image data
- Once I finalize the algorithm I would tune hyper-parameters using techniques like k-fold cross validation with grid search
- Will put checks in place to make sure that the model does not have high-bias and high-variance
- Final output from this Capstone would be a model which captures most of the patterns in the training data and predicts accurately on eye images on which the status of the disease is not known
- Although, the accuracy metric is already defined, my focus would be to build a high recall model so that it does not miss on any actual case of disease even though if there are a couple of false alarms.