Machine Learning Engineer Nanodegree

Capstone Proposal

Guillermo Aure
April 9, 2017

Domain Backgroud

Doctors in the health industry consider obesity a risk factor for multiple types of illness. Conditions like diabetes type 2, hypertension and cardiovascular diseases and various types of tumors, in overweight patients [1] are associated with higher rates of mortality. It seems clear that obesity has a negative impact on humans.

In the medical field, there is a hypothesis know as The obesity paradox [2]. The obesity paradox hypothesis state that overweight elderly patients have a better chance to survive illness like the ones describes in the previous paragraph than underweight patients. Multiples studies try to proof or invalidate such hypothesis [3, 4, 8]. Some of the studies have found evidence that there is indeed some time of protection granted by a high BMI (body mass index) and others argue that confounding variables and lack of follow up invalidate such proofs [5]. One thing is sure, that despite which study is the correct all of them prove one way or another that there is a correlation between the: BMI, the age, the medical condition and the smoking habits; and that this association influence negative or positive the resilient capacity of a human being.

My proposal is based on the background knowledge of the correlation of the features described in the previous paragraph; it should be possible to train a learner to predict how much time a patient will stay hospitalized (including ICU time) due to surgery, based in the afford mentioned features.

Problem Statement

Preoperative evaluation purpose is not to explain patients for elective surgery, but rather to put in place controls that minimize the risk of potential surgery complications [6]. As was describe in the domain background on this document obesity as a general rule is consider a surgery risk factor, and as consequence hospital tries to prepare patients for surgery, so they dont come to the procedure overweight. The problem with this approach is that considers overweight as a factor risk a not as a potential benefit. The obesity paradox hypothesis opens the door to a more custom preoperative evaluation and a possible reduction of the time a patient that gets complicated during a high-risk surgery will spend on the ICU unit. The key indicator here is the time spent hospitalized for surgery, the elapse time indicator, represents the resilient condition on a patient body based on the features describe before, in other words, the label of our predictor.

The elapsed time is a multivariable function; it depends on several features including the BMI. The objective is to predict such time using the patient's electronic medical record (EHR). The problem is we humans have problems visualizing beyond three variables. An EHR has more than three variables; this represents a challenge for a person and an opportunity for the utilization of computers and a Machine Learning algorithms. The EHR is an excellent source of information, but the different types of data (continue and categorical) will require to encode and transform some of its features. Because the elapsed time is in nature continues, the model needs to predict also a no discrete output, because of this our problem requires a regression learner.
Details of the type of input and model that we will use are explained in the following sections of this proposal.

Datasets and Inputs

In order to train and test our predictor we will use some of the feature of the MIMIC-III (MIT Lab for Computational Physiology) dataset. The MIMIC-III dataset contains more than 40,000 critical patients records. Information and details related to it can be found in the web site `https://mimic.physionet.org`, this is not an open source data a access must be requested. I will select the sample of the dataset by answering the questions documented in the solution statement of this documents. The resulting sample size after filtering the dataset is of 6862 patients. I selected the sample of the dataset by answering the questions documented in the solution statement of this documents. The resulting sample size after filtering the dataset is of 6862 patients. Along with the project files, there is an ECLIPSE BIRT workbench report design file; this file can be imported back to an ECLIPSE BIRT installation to recreate the dataset used in this project. The only modification to the BIRT report design file is the "data source" object that has to be pointed to an instance of the MIMICIII database. The file can be opened in any text editor to see the SQL queries used to create the dataset. It is importance to note that the final data set was created using BIRT joins, so I extracted three separated data sets using PostgreSQL queries and joined them using BIRT; again the dataset can be recreated using the companion BIRT report design file.
The sampled dataset will be reduce (some features will be calculate from some of the dataset values) to the following features:

FEATURES

- Body Mass Index: CHARTEVENTS table
- Sub Set of Medical Conditions: DIAGNOSESICD table
- Age: PATIENT table
- Sex: PATIENT table

LABEL

- Admission Time Discharge Time (MIMIC admissions table): ADMISSION table

NOTE: The "Body Mass Index"needs to be calculated from the heigh and weight values of the patients data. The formula we will use is:

$$BMI = WeightKg/(Heightmtrs)^2 \tag{1}$$

The features were selected based on the medical studies related to the Obesity Paradox.' The selected label is continue,' as most of the features values; some of them like the conditions (ICD9) and sex is categorical data and will be encoding or converted into dummy variables.

Solution Statement

As was described at the beginning of this document, obesity is considered a health factor risk in some cases and a protection feature for some others. Previous medical studies are proof one versus the other with no clear conclusion. My study proposes not agreed with any of the two positions but to use the existing information to measure how resilient the body is to surgery using the features of those previous studies. The metric is the time a patient spent hospitalized when it subjected to a medical operation. A potential more complete metric will be the time to recovery, but unfortunately, we do not pose such data set, so we propose to use only the hospitalize time. We are assuming that no doctor will discharge a patient which condition is inadequate for release. After assembling the dataset and scale the features we will identify the unit of the label time (days, hours, minutes).

I will propose the use of a linear regression machine learning algorithm to predict the elapse time. The main reasons for choosing regression are the constraints in time and computing power (less time to train), and second, I expect a linear correlation (positive or negative) based on the previous studies about the BMI and its implications. I will verify the two previous assumptions during the data exploration phase. A proposed benchmark model is the nearest neighbors model, this type of models also requires less time to train. I will consider a regression decision tree model but to find the optimal size of our training data. The MIMICIII data set contains 40,000 patientes records which I filtered based on the answer to the below questions:
1. Did the patient survive, yes or no? 2. Does the patient has pre-existing conditions, other than the one selected in the obesity paradox studies? 3. Does the patient has both metrics recorded, height and weight. 4. The MIMICIII database contains information on each visit of the patient to the hospital, for the purpose of this project we only selected the first visit information.
After applying the below filter the 40,000 records, I ended up with a sample of 6862 patients. I will use to an unsupervised clustering algorithm (k-mean) to classify it. From the resulting classes, I will select 1000 records randomly, representing each class. The resulting dataset will be used to plot the training accuracy versus the test accuracy and determine the optimal size of our training data set.

Benchmark Model

We will use a supervise learning method to predict the time a patient spend in the hospital. Because our vector label continues we will use a regression predictor, and we will compare its prediction with an information based learner (K-nearest neighbor), this will allows to benchmark one model versus the other. We will use cross-validation to make sure with dont overfit the model and the R2 Coefficient of Determination to measure the accuracy of our model or what is the same how well it explains the label based on the features. Our learner contains some bias introduced during the features selection; this is because we are selecting the medical conditions used in previous obesity paradox studies, selecting only patients that survive a surgery and finally only taking in consideration the first visit to a hospital. The added bias is based on the observable features related to the paradox, and the sampling process. Introducing new features will be only a guess that can create model complexity due to the additional dimensions.

Evaluation Metrics

To measure our model's accuracy, we will use the R2 or the coefficient of determination. This metric tells us how much our prediction is explained by the vector features. The highest good value for R2 is 1.0, and the worst can be negative. A model that predicts correctly not matter what features are input will always give an R2 of 0; this is a clear indication of an overfitted model. The Mathematically operation to calculate R2 is as follow: the square difference between the real label value and the predicted value,' divided by the square variance of the original label's distribution minus 1. I am dividing the actual error of my prediction between the spread of the values of the fitted data distribution, in and an ideal scenario where the prediction is close to the real value the different between one and the division is close to 1. R2 metric help to tune the model parameters so this error gets lower, and the R2 gets higher. The mathematical representation of the R2 is:

$$R^2(y, \hat{y}) = 1 - \Sigma(yi, \hat{yi})^2 / \Sigma(yi, \bar{y})^2 \tag{2}$$

Project Design

We will use Python to build the predictor; specifically, we will use the Python sk-learn framework, the modules: panda, numpy, and the python math lab API for visualization. What follows is the breakdown structure of the project.

1. Extract the dataset from the MIMIC database, and export the data to a CSV file, once the data is in CSV imported to a pandas dataframe.

2. Encode the non-numeric data in the pandas dataframe in preparation for the data exploration. Calculate the label by substracting the admission time vs. the discharge time.

3. Explore the dataset using numpy to understand its features distributions. Plot the correlation between the features and the time spend in the hospital to determine its relation.

4. Transform the features to scale them using the natural log scale, based on the distribution visualizations. If some of the distributions are skew, search for outliers and remove them using numpy quartiles function.

5. Train, test, and tune a sk-learn linear regressor model. Sklearn uses the 'Ordinary Least Squares' method to calculate the unknow parameters of the model. I will use also the sk-learn cross-validation and the sk-learn grid search function for optimization.

6. Perform and predict a sample several times using the same features to measure the model sensitivity. Calculate the result variance to see how spread the results are.

7. Plot the prediction of a subset of the dataset distribution. Compare the statistical metrics vs. the whole real values dataset. Determine if the metric matches.

8. Finally, benchmark the model against the KNN model.

References

[1] Flegal KM, Kit BK, Orpana H, Graubard BI. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. JAMA 2013;309:7182.

[2] Obesity paradox. `https://en.wikipedia.org/wiki/Obesity_paradox`

[3] Braun, N., Hoess, C., Kutz, A., Christ-Crain, M., Thomann, R.,Henzen, C., ... & Schuetz, P. (2017). Obesity paradox in patients with community-acquired pneumonia: Is inflammation the missing link?. Nutrition, 33, 304-310

[4] Carbone, S., Lavie, C. J., & Arena, R. (2017, January). Obesity and Heart Failure: Focus on the Obesity Paradox. In Mayo Clinic Proceedings. Elsevier.).

[5] BMI and all cause mortality: systematic review and non-linear dose-response meta-analysis of 230 cohort studies with 3.74 million deaths among 30.3 million participants

[6] Dagfinn Aune, PhD student and research associate1 2, Abhijit Sen, postdoctoral fellow1, Manya Prasad, resident3, Teresa Norat, principal research fellow2, Imre Janszky, professor1, Serena Tonstad, head physician3, Pl Romundstad, professor1, Lars J Vatten, professor1

[7] Preoperative Evaluation MITCHELL S. KING, M.D, Northwestern University Medical School, Chicago, Illinois Am Fam Physician. 2000 Jul 15;62(2):387-396. `http://www.aafp.org/afp/2000/0715/p387.html`

[8] TIENE IMPACTO LA HIPERGLUCEMIA Y LA OBESIDAD EN LA PROLONGACION DE LOS DIAS DE HOSPITALIZACION EN LOS PACIENTES EN EL CENTRO MEDICO DOCENTE LA TRINIDAD Author: Madeleyn Santaella,

Tutor: Gestne Aure Centro Medico Docente La Trinidad Direccion De Educacion e Investigaciones