Machine Learning Engineer Nanodegree

Capstone Proposal

Vitor R. Machado

August 29st, 2017

Domain Background

Cancer is a group of diseases involving abnormal cell growth, being one of the most complex group of diseases in the world. There are several universities and research centers searching for the best treatment for cancer. In Brazil, the state of Sao Paulo has a group of hospitals and treatment centers called ICESP known as the best option for cancer treatment in Brazil (read more in this link - in portuguese). These care centers provide indicators of the quality of their treatment based on their records (read more in this link, page 6), however, this network of hospitals does not have a method to predict the likelihood of mortality or the lifetime expectancy for the cases in which the cure is remotely possible.

Problem Statement

The ICESP is a network of hospitals of a third world country, therefore it does not posses the most advanced equipments for cancer treatment. Moreover, the most advanced treatments for cancer are not available in this country yet. Based on this, the global average life expectancy for cancer patients and chances for surviving may not be applicable for this case. Furthermore, the ICESP does not have a method to predict the probability of survive for a given patient given a combination of treatments neither the lifetime expectancy in the cases in which death is more likely.

Datasets and Inputs

The ICESP has several records of their patients. Their dataset is very detailed with a lot of information and can be downloaded through this link. Their dataset has information about more than 700k patients from 2000 to 2017. Since this dataset has several columns, each column in this dataset is described in an external PDF file (columns_definition.pdf). The original description of each column is available in this link, in portuguese. This dataset contains data such as the age and sex of patients, the treatment they received (Immunotherapy, hormone therapy, chemotherapy, radiotherapy, surgery), the type and stage of the cancer, and if the patient was cured, died or left the hospital before the end of the treatment. This dataset is extremely relevant for the project since it is possible to make classifications and regressions using Machine Learning through this dataset. Learning more about this dataset would allow Machine Learning algorithms to make predictions about new patients and their chances of surviving.

Solution Statement

My main goal is to build a Machine Learning model to help the ICESP to make predictions of the likelihood of life or death for every future patient of this network of hospitals. Moreover, for the cases in each the first model predicts a high probability of death, I aim to build a model to predict the life expectancy in number of days for these patients given the same treatments. These models would help ICESP to increase the efficiency and rate of success of their treatments.

Benchmark Model

For making predictions of the likelihood of life and death, a deep neural network is applicable. I've chosen DNN because of its versatility and performance. DNN's reduce the need for feature engineering and their architecture is adaptable for several types of problems. The softmax function in the output layer would give the probability for both life and death of the future patients.

For making a linear regression to output the life expectancy, the algorithm SVM (more specifically the Support Vector Regression) is the one I've chosen as solution, since it is very effective in high dimensional spaces (dataset with many features).

Evaluation Metrics

For the deep neural network, the softmax function will output the probability for life and death. This result can be one-hot encoded to return life or death, the most probable of both. Based on this, it is possible to compare the output of the neural network in a test set to have the accuracy of the model. It is also necessary to compare the accuracy and loss of the neural network in different epochs to avoid under/overfitting and select the best number of epochs for training.

For the SVM, the coefficient of determination R^2 is the performance metric chosen to evaluate this model. The R^2 compares the predicted value to the actual value and gives an output between 0 and 1, being 1 the optimal value since in these cases the model is making a good job in predicting linear values.

Project Design

Firstly, prior to using the records of the dataset to train and test both models, some preprocessing is required:

- Categorical features will be transformed into dummy features;
- Numerical features such as age, days between first appointment and last information about the patient will be normalized to range between 0 and 1 to facilitate both training and testing;
- Features that are supposed to be irrelevant or redundant (such as the state of residence of the patient, age range and patient’s education degree) will be removed.
- Missing values in some columns will be replaced with 0’s.
- For the cases of death, to get the number of days that a given patient lived in the hospital after the diagnosis of cancer it is necessary to sum up two columns - the number of the days from the first appointment and the diagnosis and the number of days of treatment - into a new column.

The first model will be a deep neural networks with 2 hidden layers. Both hidden layers will have a ReLu activation, and the final layer will have a softmax activation to give probabilities for both life and death which will be one-hot encoded. Then the predicted values will be compared to the labels of the test set to give the final accuracy of the model. The input layer will receive the selected features from the preprocessed dataset.

The second model will be a Support Vector Regressor algorithm. This model will be fitted with the preprocessed dataset, but only with the records in which the patients died from cancer. The output of this model will be different; it will output the predicted number of days as lifetime expectancy, which will be compared to the new feature produced as the sum of the number of the days from the first appointment and the diagnosis and the number of days of treatment until death. This model will be tuned using GridSearchCV to find the best values for both C and Epsilon (which are parameters of the Support Vector Regressor) and cross-validation.

The following Python libraries will be used for this task:

- Pandas;
- Numpy;
- Keras;
- Tensorflow;
- SKlearn.