Machine Learning Engineer Nanodegree (2019)

Capstone Proposal

Sameh Adel
June 4, 2019

Breast Cancer Detection Using Machine Learning

Domain Background

One of the most promising fields that pay time and money in Artificial Intelligence researches is Healthcare. Many areas in Healthcare uses Machine Learning to help Doctors and hospitals create perfect environment that aims for better health as well as for better life, starting from disease detection going through treatment recommendations.

Many companies and institutes developed researches and solutions in this field like IBM, Microsoft, Google, ETC. Google Inc have built Massively Multitask Networks for Drug Discovery. In order for Google to train these Networks they gathered large amounts of data from public sources to create a dataset of nearly 40 million measurements across more than 200 biological targets.

WHY this domain of problem should be solved?
Because this field is so sensitive to human errors and mispredictions, so Artificial Intelligence can provide this domain with variety of applications not to replace the Doctors but to create powerful tools between Doctors' hands in order to classify patients more accurately as well as the accurateness of treatment recommendation. Also, I can say mixing the power of a great Doctor with help of AI can lead to near ZERO errors and save a lot of money, time and most important human lives.

Problem Statement

The goal of this problem is to use the power of Machine Learning algorithms to take the dataset of past measurements and understand which most features that leads to Breast Cancer? Also, to predict the likelihood of future patients to be diagnosed as sick.

So given important measurements of a future patient we can train a ML algorithms to predict if he/she carries a Breast Cancer easily and accurately.

Datasets and Inputs

For this problem I have downloaded a dataset called "Breast Cancer Wisconsin (Diagnostic) " from Kaggle datasets which is originally loaded from UCI Machine Learning datasets.

This dataset contains 569 rows each row represents one observation, each observation has 32 features from which I dropped one unnecessary column from it (id column).

Features of this dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Solution Statement

For this problem I'm going to explore the dataset and extract insights from it, after that I'm going to use the best Machine Learning algorithms to classify the current and the future observations to see if a patient sick or not.

Benchmark Model

I planned to list some of the best Machine Learning models and see which ones are appropriate for this problem, and after that I will filter these models to pick the best one that give me the best accuracy.

These models are:

1) Support Vector Machine (SVM)
2) Logistic Regression
3) Random Forest Classifier
4) Multi-Layer Perceptron (MLP)

Finally, I will pick the best model of them and I'll try to apply an **ensemble** method to see if this could give better result or not.

Evaluation Metrics

I going to use some metrics and techniques to evaluate the selected models and the final model, these metrics are:

1) Receiver operating characteristic curve (ROC)
2) Learning Curve to help in overfitting and underfitting detection
3) F1 score based on Precision and Recall

Project Design

My Project for this problem will go through a lifecycle which following these steps:

1) **Importing the dataset.**
2) Work with some **Statistics** to get important insights:
   In order to get familiar with the dataset I going to apply some basic Statistics like median, mean, std, min, max, IQR, ETC.

3) **Apply Exploratory Data Analysis:**
   In this stage I'm going to plot histograms to see the distribution of the data and some scatter plots and heat map to detect the features dependences that will help in feature selection.

4) **Apply Feature Engineering and scaling:**
   Here I will include and exclude some features according to each feature importance. The important feature will included otherwise will excluded.

5) **Model Selection process to select the most appropriate ML Model for this problem:**
   At this stage I have some models (4 models as I mentioned in Benchmark Model section), so I'll plot learning curves to help me pick the most appropriate one that neither overfit nor underfit.

6) **Model Evaluation and Fine Tuning the model hyper-parameter:**
   In this final stage I'll use the mentioned evaluation metrics to evaluate the model performance, after this I'll apply grid search in order to fine tune the model's hyper-parameters. After applying grid search I'll use an ensemble method trying to get more accurate predictions. At the end of this stage the final model will be ready to go to life.