PhUSE US Connect 2019

Paper ML02

Getting REAL with Supervised and Unsupervised Machine Learning Algorithms in the Real World

Himanshu Pandya, Biogen, RTP, USA
Bob Engle, Biogen, Cambridge, USA

ABSTRACT

Machine learning is playing an increasing role in today's world and is rapidly being deployed in the health care industry and for sound health care decisions in the REAL world. Currently, multiple Real World Data (RWD)/Real World Evidence (RWE) initiatives are being sponsored by Biogen. Machine learning concepts are being applied across one such initiative.

This paper highlights features of the two key Machine learning algorithms - Supervised and Unsupervised learning, their implementation on the RWD. We explore how the algorithms can be developed and deployed to gain better insights, understand and predict patient's disease severity, disease progression and some key clinical outcomes. The solutions are implemented using the R statistical software.

DISCLAIMER

The scope of this paper is to present the opinions and suggestions of the authors. The interpretations of standards and procedures contained in this paper are those of the authors and may not be necessarily correct. Any views and recommendations stated within this paper are those of the authors, and they do not represent the positions of their employer.

INTRODUCTION

Machine learning (ML) is THE buzz word of this century. Hardly anyone might not have heard the name "Machine Learning", irrespective of their industry domain. This paper presents a brief overview on what ML is about, it's concepts and why it's gaining traction (now), some of the popular ML algorithms being used and how the ML is shaping the industry. Further we discuss and present ML applications in an on-going project based on real world data (RWD) at Biogen – ML model implementations, challenges faced, lessons learned and conclusions.

This paper is broken into the following sections:

- What is Machine Learning (ML) and why future of ML is NOW
- Potential applications of ML in the Life Science industry
- Some popular ML learning styles
- ML application and implementation at Biogen
- Conclusions (so far)

WHAT IS MACHINE LEARNING (ML)

Machine learning (ML) is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence (AI) based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. It is the scientific study of algorithms and statistical models that computer systems use to progressively improve their performance on a specific task. ML is an application of artificial intelligence (AI) that provides systems or machines the ability to automatically learn and improve from experience without being explicitly programmed. While AI is the broad science of mimicking human abilities, ML is subset of AI that trains machines how to learn.

The basic fundamental concept is that machines will train algorithms with input test data, and the trained algorithms will predict the results with the real data. This provides a level of fine tuning and repetition that is very difficult to impossible to achieve manually.

Different definitions of ML, five definitions from reputable sources:

- "Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world." – Nvidia
- "Machine learning is the science of getting computers to act without being explicitly programmed." – Stanford
- "Machine learning is based on algorithms that can learn from data without relying on rules-based programming."- McKinsey & Co.
- "Machine learning algorithms can figure out how to perform important tasks by generalizing from examples." – University of Washington
- "The field of Machine Learning seeks to answer the question "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" – Carnegie Mellon University

FUTURE OF ML IS NOW

ML concepts have existed for a long time but because of new computing technologies such as faster processors etc., machine learning today is not like machine learning of the past. It's much more refined, advanced, efficient and very cost effective.

Growing volumes and varieties of available data, computational processing (especially cloud computation) that is cheaper and more powerful, and affordable data storage.

All of these mean it's possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results. By building precise models, organizations have better data insights and have a better chance of identifying opportunities.

POTENTIAL APPLICATIONS OF ML IN THE LIFE SCIENCE INDUSTRY

- Improve Medical Diagnosis
- Better Patient Care
- Drug discovery
- Medical image recognition
- Site selection
- Data anomality detection
- Medical coding
- Biomarker Detection in terms of both prognosis and diagnosis

POPULAR ML LEARNING STYLES

- **Supervised Learning:**
  Dependent/outcome variable is predicted from a given set of independent/predictors variables. Using these set of variables, a function is generated that map inputs to desired outputs. A model is prepared through a training process in which it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data.
  Supervised learning algorithms include classification and regression.
  - **Classification**
    Algorithms are used when the outputs are restricted to a limited set of values
    Program learns from the data input given to it and then uses this learning to classify new observation.
    Examples - Logistic Regression, Naive Bayes Classifier, Support Vector Machines, Decision Trees, Random Forests, Neural Networks
  - **Regression**
    Algorithms are used when the outputs may have any numerical value within a range.
    Examples - Linear, Stepwise, Polynomial, Ridge, Lasso, ElasticNet

- **Unsupervised Learning:**
  Infer patterns from a dataset without reference to known, or labeled, outcomes. Unsupervised learning can be used for discovering the underlying structure of the data. Unlike supervised machine learning, this method cannot be directly applied to a regression or
  a classification problem because you have no idea what the values for the output data might be. Instead unsupervised learning algorithms identify commonalities in the data.
  - **Clustering**
    Clustering allows you to automatically split the dataset into groups according to similarity. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters. Data points inside a cluster are homogeneous, but heterogeneous to other clusters. K-Means is one of the most popular algorithms use for clustering application.
  - **Association mining**
    Identifies sets of items that frequently occur together in a dataset.
  - **Anomaly Detection**
    Detection can automatically discover unusual data points in a dataset.

- **Reinforcement Learning:**
  It's learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences. The idea is to let machine learn from past experiences and to capture the best possible knowledge to make accurate decisions
  Examples – Markov Decision Process (MDP), Monte Carlo, Q-Learning, Deep Q Network (DQN)
  .

ML APPLICATION AND IMPLEMENTATION AT BIOGEN

Currently, multiple Real World Data (RWD)/Real World Evidence (RWE) initiatives are being sponsored by Biogen. In this paper we discuss and present ML application/implementation in an on-going project based on real world data (RWD), for Multiple Sclerosis (MS) indication with more than 10K patient population. RWD were collected across North America and Europe.

Supervised and unsupervised ML learning concepts were applied for this project.

- **Unsupervised Learning**
  The goal is to discover interesting things about the data. Are there an informative way to visualize the data, other than the regular pre-define subgroups based on gender/age/ethnicity/sites etc? Basically, to discover unknown new sub-groups among the data based on certain Multiple Sclerosis (MS) disease characteristics i.e disease duration, EDSS, PDDS, PST, WST.
  This can be efficiently achieved by means of "Clustering". Clustering allows us to identify which observations are alike, and potentially categorize them.

  K-Means clustering is the simplest and the most commonly used clustering method for splitting a dataset into a set of k groups.

  Steps followed for the K-Means clustering analysis -
  - **Data Preparation**
    Clean, non-missing data is the key/base for accuracy of all analysis. Missing data were imputed (using SAS, Proc MI procedure) and the outliners (10% either side) were dropped from the analysis. Thereafter the data was standardized/scale i.e. z-scores were computed for all the variables of interest.
  - **Pre-Clustering Preparation**
    Used R Statistical Software for cluster algorithm.
    Packages used – Tidyverse (for data manipulation), Cluster (for clustering algorithm), Factoextra (for clustering algorithm and cluster visualizations)
  - **Applying the K-Means Clustering**
    The goal is to partition a given data set into a set of *k* groups (i.e. *k* clusters), where *k* represents the number of groups pre-specified by the analyst. It classifies objects in multiple groups (i.e., clusters). The basic idea behind k-means clustering consists of defining clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized.
    Major steps of this process –
    a) Specify the number of clusters (K) to be created.
    b) The algorithm selects randomly k objects from the data as the initial cluster centers
    c) Next, it assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid.
    d) For each of the k clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster.
    e) Iterate steps c and d until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations.

Computing K-Means

K-Means can be computed in R by using the kmeans function -

Kmeans(datain, centers =xx, nstart = xxx).

Wherein datain is the input data, centers is the arbitrary user assigned number of clusters to be created and nstart is the number of initial configurations.

Kmeans output

Some of the key output variables -

i. *cluster: Cluster to which each point is allocated.*
ii. *centers: A matrix of cluster centers.*
iii. *totss: Total sum of squares.*
iv. *withinss: Vector of within-cluster sum of squares.*
v. *tot.withinss: Total within-cluster sum of squares, i.e. sum(withinss).*
vi. *betweenss: The between-cluster sum of squares, i.e. $totss-tot.withinss$.*
vii. *size: The number of points in each cluster.*

Cluster number optimization

Selecting right number of clusters can be very trick and subjective. Fortunately, R has made it easy for the users. Clusters can be visualized by using the R function – **fviz_cluster.** It provides very good understanding of different clusters and helps in deciding the optimum number of clusters.

Apart from it there's an "Elbow method" for a better precision in optimizing the clusters. In this method "Total within-cluster sum of square (wss)" is plotted against number of clusters, which gives a knee-shaped curve. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

R has made it very simple by wrapping the entire process in a single function **fviz_nbclust**

Both the methods can be utilized to determine ideal number of clusters for your data.

- **Supervised Learning**

  Under the supervised learnings, Linear Regression was used to understand relationship among the variables of interests. Regression is a supervised learning algorithm which helps in determining how does one variable influence another variable. The algorithm consists of a dependent /outcome variable which is to be predicted from a given set of predictors/independent variables. Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data.

  Steps implemented -

  a. **Data Preparation**
     It's of paramount importance to perform the data prep work, just like as explained in the Unsupervised Learning part (above).

  b. **Splitting into Train/Test Data**
     Prior to building any model on the data, we are supposed to split the data into "train" and "test" sets. The model will be built on the "train" set and it's accuracy will be checked on the "test" set. Used R package "caTools" to split the data into two sets. This package provides a function "sample.split()" which helps in splitting the data-
     sample.split(datain$var1,SplitRatio = 0.xx)->split_index
     Basically xx% of the observations from var1 column will be assigned to the Training Dataset and the rest (100-xx)% will be assigned under the Test Dataset.

  c. **Building the model**
     Using LM function in R, the built model on Training data. Used one of the MSFC4 component (Multiple Sclerosis Functional Composite) as the dependent/outcome variable. The independents variables of interest were - demographics, patient characteristics, disease duration, certain neurological score variables.

  d. **Predictions**
     Used the PREDICT function in R for predictions based on the test data.
     predict(model,test)->result
     The predicted results are stored under the result object.

  e. **Error/RMSE Computation**
     Computed the error by subtracting the predicted values from the actual values.
     Further calculated the Root Mean Square Error (RMSE) which gives an aggregate error for all the predictions. The RMSE is the square root of the variance of the residuals/errors.
     It indicates the absolute fit of the model to the data close the observed data points are to the models predicted values.

  f. **Determining model fitness**
     RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is one of the most (but not the ONLY) important criterion for fit if the main purpose of the model is prediction.
     Multiple models can be built by adjusting the data split ratio, predictors (adding/dropping predictors) and compared based on the RMSE. Based on it users can determine and select the best fit model.

Conclusion:

The time to adopt Machine Learning is NOW. -
For years the Life Science industry has been accumulating immense amounts of data (BIG DATA) Advances in cloud technologies have provided a way to store all of that data, for a fraction of cost. Excellent statistical software like R/SAS/Python can perform the heavy lifting statistical jobs. ML can crunch vast amounts of data and identifying patterns that is simply very to difficult to impossible to achieve manually.
While AI/ML enables better mining of big data, it still requires user supervision to set up, train and use the system in an intelligent way. It will thus facilitate faster, more informed and more accurate research than ever before, but it will complement and assist, not replace, researchers.
So let's not fear but EMBRACE Machines and educate them!

After implementation of ML (still on-going) at Biogen, we have received great insights about the collected patient RWD, especially understanding the hidden cohorts in our data based on the Kmeans clustering algorithms. It's a great mechanism to understand the patient population distribution classified on certain patient characteristics, disease duration/characteristics, demographics etc.
Supervised Learning has helped us to develop and train our models for better predictions of clinical variables of interests. We are learning something new and getting better every day.
It's an iterative and evolving process!

ACKNOWLEDGMENTS
We would like to thank Chris Kania (Director, Statistical Programming, Biogen) for his invaluable content planning and editing assistance.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Authors can be contacted at:

Himanshu Pandya, Biogen, RTP, USA
Himanshu.Pandya@Biogen.com

Bob Engle, Biogen, Cambridge, USA
Bob.Engle@Biogen.com