Machine Learning – Why we should know and How it works

Kevin Lee, Clindata Insight, Moraga, CA

ABSTRACT

The most popular buzz word nowadays in the technology world is "Machine Learning (ML)." Most economists and business experts foresee Machine Learning changing every aspect of our lives in the next 10 years through automating and optimizing processes such as: self-driving vehicles; online recommendation on Netflix and Amazon; fraud detection in banks; image and video recognition; natural language processing; question answering machines (e.g., IBM Watson); and many more. This is leading many organizations to seek experts who can implement Machine Learning into their businesses.

Statistical programmers and statisticians in the pharmaceutical industry are in very interesting positions. We have very similar backgrounds as Machine Learning experts, such as programming, statistics, and data expertise, thus embodying the essential technical skill sets needed. This similarity leads many individuals to ask us about Machine Learning. If you are the leaders of biometric groups, you get asked more often.

The paper is intended for statistical programmers and statisticians who are interested in learning and applying Machine Learning to lead innovation in the pharmaceutical industry. The paper will start with the introduction of basic concepts of Machine Learning - hypothesis and cost function and gradient descent. Then, paper will introduce Supervised ML (e.g., Support Vector Machine, Decision Trees, Logistic Regression), Unsupervised ML (e.g., clustering) and the most powerful ML algorithm, Artificial Neural Network (ANN). The paper will also show how programmers can use Python Scikit-Learn and Google TensorFlow. The paper will also introduce some of popular SAS ® ML procedures and SAS Visual Data Mining and Machine Learning. Finally, the paper will discuss the current ML implementation, its future implementation and how programmers and statisticians could lead this exciting and disruptive technology in pharmaceutical industry.

INTRODUCTION OF MACHINE LEARNING

Machine Learning 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. The key of the Machine Learning definition is "**without being explicitly programmed**". What it means is that the machines will learn new skills by themselves without programmers adding new rules for new skills. Let's explore how machines learn.

HOW MACHINES LEARN

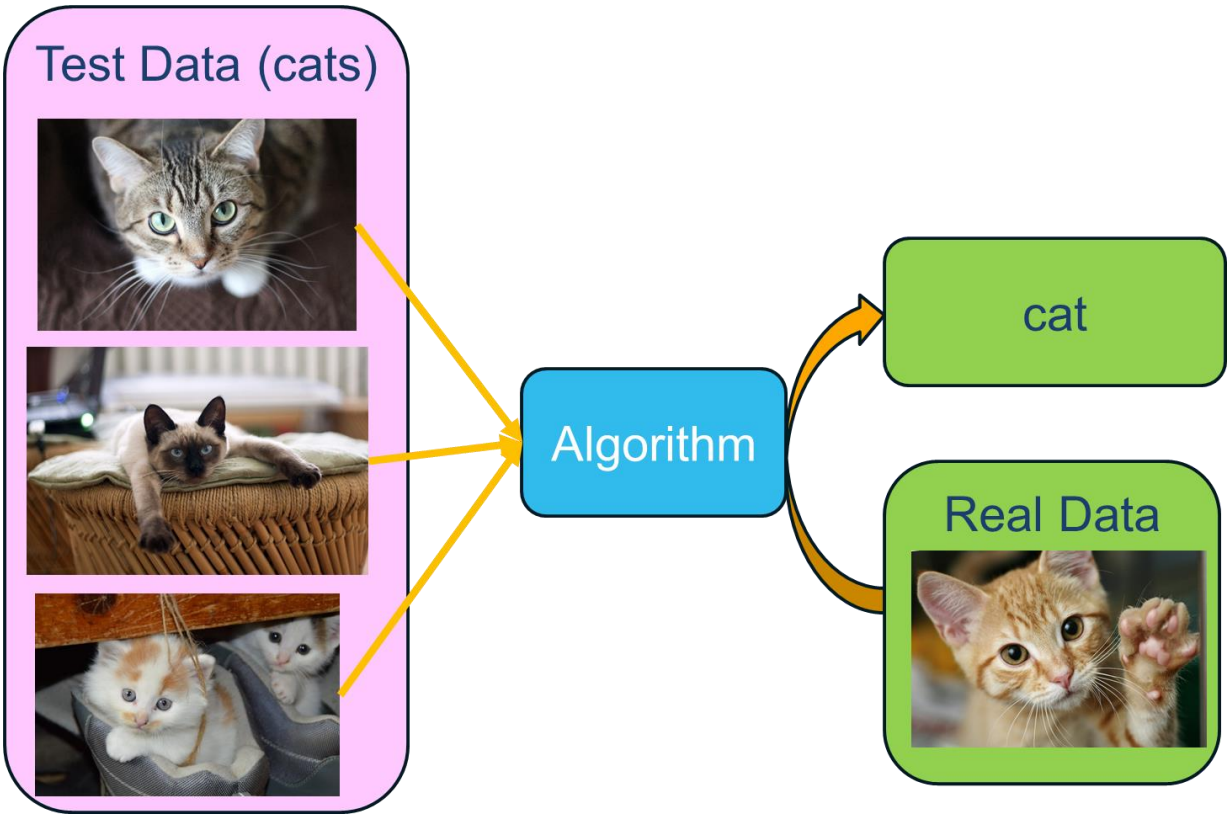
Machine Learning is very similar to human learning. Humans learn from experiences. We learn from seeing, hearing, tasting, touching and smelling. We know that a picture in Figure 1 is a cat because we have seen cat pictures before. So, when we see a cat picture like one in Figure 1, we can accurately predict it is a cat.

Figure 1. a picture of a cat

Machines learn very similar ways. Machines learn from data since machines do not have senses of humans. Basically, machines will train algorithms with input test data, and the trained algorithms will predict the results with the real data. Its use case is shown in Figure 2.

As seen in Figure 2, a lot of cat's images will train its algorithm and the trained algorithm can predict the cat accurately when real data is processed.

Figure 2. How machines learn using data and algorithms



TYPICAL MACHINE LEARNING PROJECT WORKFLOW

Machine Learning project is very complex. Like in Figure 2, the machine learning requires data and algorithms, and it also requires very thoughtful process. For successful Machine Learning project, the data scientists follow the following steps:

1. Identify the problems to solve.
2. Integrate data from multiple sources.
3. Ensure data quality and transform the data.
4. Prepare input train and test data.
5. Select the proper Machine Learning algorithm.
6. Train Machine Learning algorithm to build the best model.
7. Implement the trained Machine Learning model into the production.
8. Predict the results with real data using the trained model.

BASIC MACHINE LEARNING ALGORITHMS

Machine Learning algorithm requires three main functions – hypothesis function, cost function and gradient descent.

1. First, Machine Learning requires Hypothesis function. Hypothesis function is basically model for the data. For example, the hypothesis function of lineal model is $Y = \theta_0 X_0 + \theta_1 X_1 + \theta_2 X_2 +,,,,, + \theta_n X_n$.
2. Second, Machine Learning require the cost function. The cost function measures how well hypothesis function fits into data. It calculates the difference between actual data point and hypothesis data point.
3. Last, Machine Learning uses gradient descent to find the best model of the data. Gradient Descent will minimize the cost function and eventually find the best model for the data.

TYPE OF MACHINE LEARNING ALGORITHMS

Once data are prepared, data scientists should select proper algorithm. When data scientists select the algorithms for data, they need to consider the type of ML tasks – supervised machine learning and unsupervised machine learning tasks.

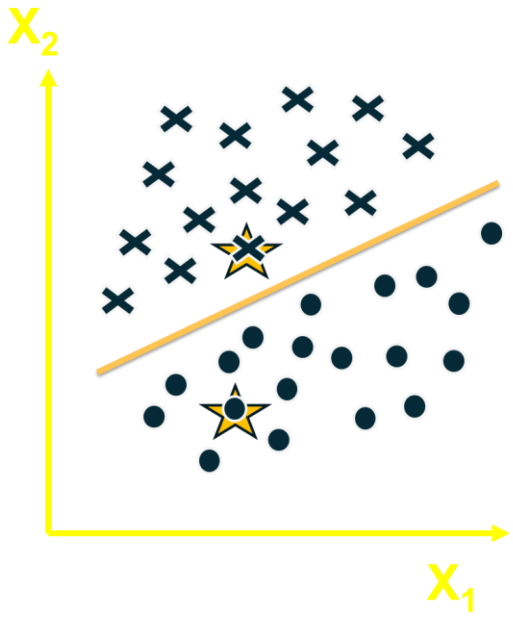
Supervised Machine Learning Tasks

Supervised Machine Learning tasks are used when input data is labeled. Basically, input data has results like "Yes"/"No", 0 to 9 or "cat"/"dog". Supervised Machine Learning tasks are very common machine learning projects. The followings are the most basic supervised machine learning tasks.

- Classification
- Regression

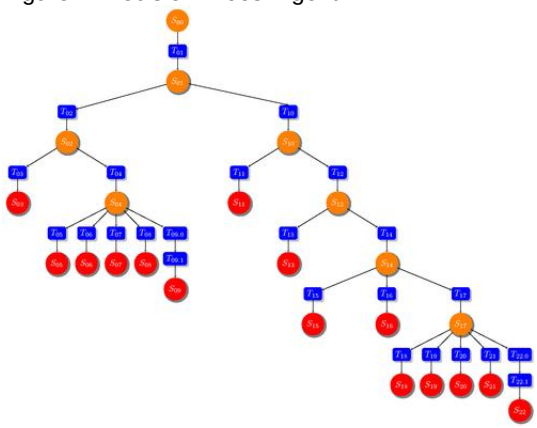
Classification tasks have the categorical target such as “Yes”/”No”, “Mild”/”Moderate”/”Severe”, “X”/”O”, or 0 to 9. Its basic ML algorithms are Logistic Regression, Support Vector Machine (SVM), Decision Trees and Random Forests. As shown in Figure 3, Machine Learning Classification algorithm will find the line or border between X and O, and it will predict whether the next data point will be X or O using the line.

Figure 3. Machine Learning Classification



Decision trees are one of the most popular Machine Learning algorithms, capable of fitting complex datasets. Decision trees algorithm can identify various ways of splitting datasets into branch-like segments. Its trees-like analysis is shown in Figure 4.

Figure 4. Decision Trees Algorithm



SAS has Decision Trees procedure and its simple codes are the followings.

```
PROC HPSPLIT data = ADAE maxleaves=100
        maxbranch = 4 leafsize=1 ;
        model Y(event='y') = x1 x2 x3 x4;
Run;
```

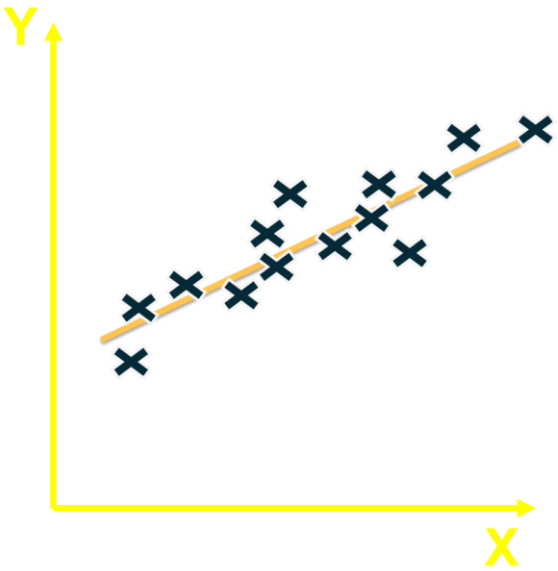
Python is one of the most popular Machine Learning language since the most popular and widely machine learning frameworks are implemented in Python. Python comes with a huge number of in-built libraries for Machine Learning including scikit-learn and TensorFlow. The paper will introduce Python Machine Learning libraries, scikit-learn and TensorFlow.

Below are the sample Python codes for Decision Tress.

```
#import ML algorithm
from sklearn.tree import DecisionClassifier

#prepare train and test datasets
x_train = …
y_train = ….
x_test = ….

#select the algorithm and train model
d_tree = DecisionClassifier(max_depth=4)
d_tree.fit(x_train, y_train)

#predict output
predicted = d_tree.predict(x_test)
```

Regression tasks have the numeric targets, continuous variables. Its basic ML algorithms are Simple Linear Regression and Polynomial Regression. Simple Machine Learning regression tasks are shown in Figure 5.

Figure 5. Machine Learning Regression



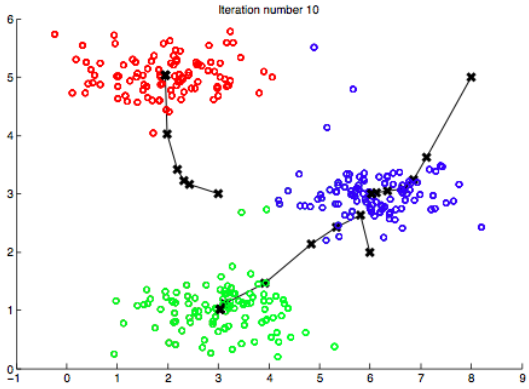
Linear Regression algorithm is probably one of the simplest model and widely used model. Below are the sample Python codes for Linear Regression.

```
#import ML algorithm
from sklearn import linear_model

#prepare train and test datasets
x_train = …
y_train = ….
x_test = ….

#select and train model
linear = linear_model.LinearRegression()
linear.fit(x_train, y_train)

#predict output and validate
predicted = linear.predict(x_test)
```

Unsupervised Machine Learning Tasks

Unsupervised Machine Learning tasks are used when input data is NOT labeled. Basically, data does not have results. Unlike supervised machine learning tasks, unsupervised machine learning tasks are used mainly for exploratory analysis. One of the most popular unsurprised machine learning is clustering. Clustering algorithm is used to assign the set of observations into subsets (clusters) as shown in Figure 6.

Figure 6. Clustering algorithm



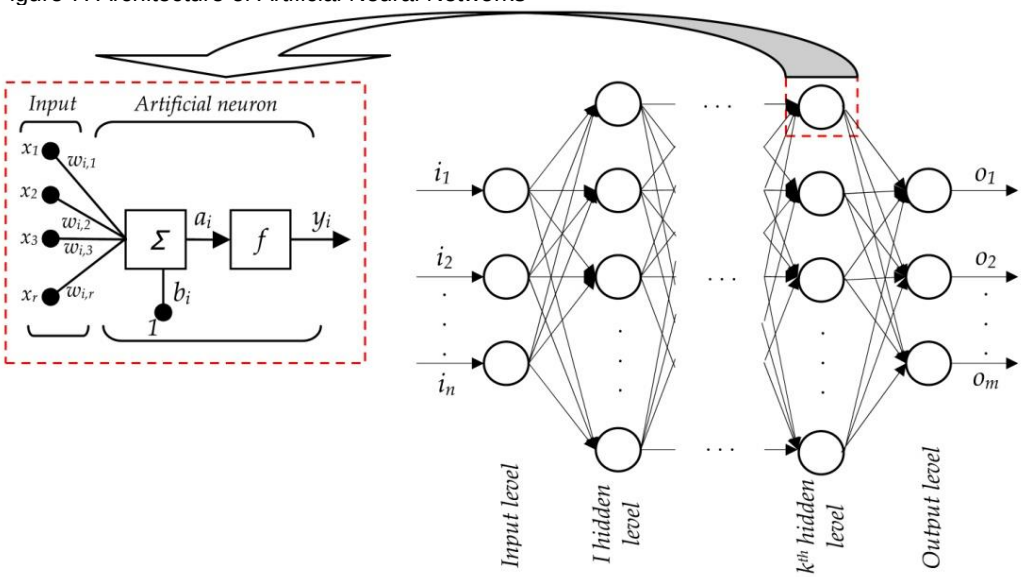
ARTIFICIAL NEURAL NETWORK (ANN)

Artificial Neural Network is the most powerful Machine Learning algorithm. It is versatile, scalable and powerful. It is ideal for very complex problems like Natural Language Processing, Voice Recognition Systems (e.g., Siri, Alexa), Autonomous Vehicles, Images Recognitions, Music or Video Recommendation and Face Recognitions, capable to beat the best GO player.

Artificial neural networks are very similar to human neural networks. Human neural network receives the electronic transmitters through its dendrites, process the electronic signal and send its own signal to next neurons.

Just like human neural neurons, ANNs receive data from previous ANNs, process received data and send the process data to next ANNs. The basic architecture of ANNs is shown in Figure 7.

Figure 7. Architecture of Artificial Neural Networks



DEEP NEURAL NETWORK (DNN)

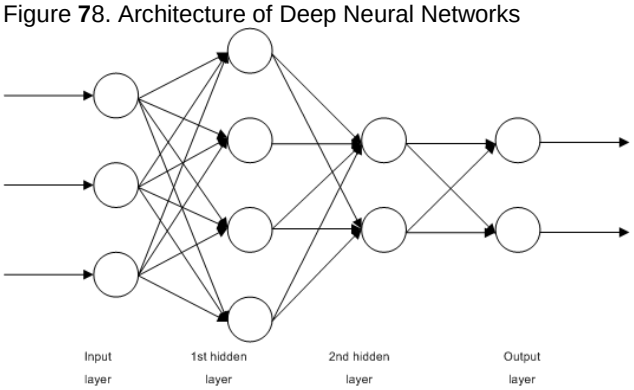
Deep Neural Networks are distinguished from single hidden layer artificial neural networks by their depth. Basic architecture of Deep Neural Networks contains the followings.

- Input layer – input data such as X_1 , X_2 , X_3 , X_4,,,,,,,,,X_n
- Hidden layers – the multiple layer of nodes
- Output layers – final results such as “Yes/No”, “Cat”/”Dog”, 0 to 9, or the price of house.

In Figure 8, the architecture of DNN consist of the followings.

- Input layer has 3 features (variables)
- Two hidden layers
  - Hidden layer 1 has 4 neurons
  - Hidden layer 2 has 2 neurons
- Output layer – 2 outputs

Figure 78. Architecture of Deep Neural Networks



SAS has procedure for Deep Neural Networks. Its SAS codes for **Figure 9** is followed.

```
PROC NNET data=Train;
        architecture mlp;
        hidden 4;
        hidden 2;
        input x1 x2 x3 ;
        target Y;
RUN;
```

Sample Python codes of DNNs for **Figure 10** are the followings.

```
# Import DNN - TensorFlow
Import tensorflow as tf

#prepare train and test datasets
x_train = …
y_train = ….
x_test = ….

# Prepare placeholder
X = tf.placeholder(..)
Y = tf.placeholder(..)

# DNN architecture
hidden1 = tf.layer.dense(X, 4, activation=tf.nn.relu)
hidden2 = tf.layer.dense(hidden1, 2, activation=tf.nn.relu)
output = neuron_layer(hidden2, 2)

# Create loss function and optimizer
cost_f = tf.nn.sparse_softmax_cross.entropy_with_logits(labels=y_train, logits=output)
loss = tf.reduce_mean(cost_f)
optimizer = tf.train.GradientDescentOptimezer(0.1)
traing_op = optimizer.minimizer(loss)

# Train DNN model
tf.Session.run(training_op, feed_dict={X:x_train, Y:y_train})
```

SAS VISUAL DATA MINING AND MACHINE LEARNING

SAS has developed easy-to-use Machine Learning platform, SAS Visual Data Mining and Machine Learning. It has very intuitive and easy to use tools for Machine Learning tasks such as Liner Regression, Logistic Regression, Support Vector Machine, Decision Tress and Deep Neural Networks. The snapshot of its platform is shown in Figure 11.

Figure 11. SAS Visual Data Mining and Machine Learning.

WHY MACHINE LEARNING IS SO POPULAR NOWADAY

The machine learnings or AI is predicted to revolutionize all the industries, especially healthcare industry. First, the Machine Learning can help us to solve a lot of complex business problems that we have not been able to solve before.

Secondly, Machine Learning can be also very cost effective since it can automate a lot of process. Andrew Ng, the founder of Coursera and previous chief data scientist of Baidu said, **"Pretty much anything that a normal person can do or think less than 1 second, we can now automate with AI"**. Machine Learning along with robotic is expected to automate a lot of human labors. According McKinsey, **as many as 375 million workers (14% of global workforce) may need to switch jobs due to automation by Machine Learning and AI**.

Due to its cost-effective potential and ability to solve complex problems, many companies and thought leaders are considering Machine Learning as the next industrial revolution. More and more businesses are implementing Machine Learning to innovate and lead the next industrial revolution. The companies like Google and Facebook target Machine Learning/AI as their priority, moving from mobile.

THE CURRENT MACHINE LEARNING IMPLEMENTATION

Machine Learning is being used more than we realized. The followings are Machine Learning implementation in our daily lives.

- Voice Recognition System – Siri, Alexa, Google Home
- Recommendation – Amazon, Netflix, Spotify
- Customer Service – Online chatting (e.g., Chatbots)
- Cashless store – Amazon GO
- Autonomous Vehicles – Tesla, Google
- Image recognition - CT scans
- Face recognition

MACHINE LEARNING IMPLEMENTATION IN PHARMACEUTICAL INDUSTRIES

Machine Learning is believed to have a huge impact in pharmaceutical industry since 1/3 of data comes from healthcare industry. Its implementation is still in fancy, but more and more pharmaceutical companies go into this area. The followings are some use cases or initial investments and partnerships of pharmaceutical companies on Machine Learning implementation.

- GSK signed 43 million-contract with Exscientia to speed drug discovery. GSK aims to reduce ¼ time (i.e., 5.5 to 1 years) and cost to identify a target for disease intervention to a molecule.
- Surgical Robotics in J&J partners with Google. It will leverage AI/ML to help surgeons by interpreting what they see or predict during surgery.
- Roche works with GNS healthcare to use ML to find novel targets for cancer therapy using cancer patient data.
- Pfizer works with IBM and utilize Watson for drug discovery. Watson has accumulated data from 25 million articles compared to 200 articles a human researcher can read in a year.
- Novartis partners with Watson to develop a cognitive solution using real-time data to gain better insights on the expected outcomes. With Cota Healthcare, Novartis also aim to accelerate clinical development of new therapies for breast cancer.

Many pharmaceutical companies try to utilize AI/ML to automate and speed up the following areas.

- Drug discovery
- Drug candidate selection
- System optimization

- Medical image recognition
- Medical diagnosis
- Optimum site selection or recruitment
- Data anomality detection
- Personized medicine
- Medical coding
- Maybe SDTM, ADaM and TLF developments?

MACHINE LEARNING WORKING GROUP IN PHUSE

PhUSE just started Machine Learning Working Group in 2018. Our goal is to introduce Machine Learning to statistical programmers and biostatisticians in pharmaceutical industry. Our working group believe that Machine Learning initiative and innovation should start within biometric departments because our expertise on programming, statistics and data. We live with data all the times. There are no other functions who know more about data and its analysis than biometric team do in pharmaceutical companies. Machine Learning WG wants to help programmers and statisticians to lead Machine Learning innovation in pharmaceutical industry.

Machine Learning WG needs your expertise and volunteer. We can do this together. For volunteer or participating in Machine Learning WG, please contact Kevin Lee (kevin.kyosun.lee@gmail.com, klee@clindatainsight.com), Nicholas Dupuis at ndupuis@protonmail.com or Wendy Dobson at wendy@phuse.com.

CONCLUSION

Machine Learning will greatly impact pharmaceutical industries as well as our daily lives. Its market is estimated to increase from 300 million in 2016 to 10 billion in 2024, expecting more than 40% annual growth rate. And we are short in talents. This will be great opportunities for biometric department – statistical programmers and statisticians. With Machine learning knowledge, the biometric department can lead the next innovation of pharmaceutical industries, data-driven medicine company that new Novartis CEO claimed Novartis to be.

REFERENCES

- SAS Visual Data Mining and Machine Learning, https://www.sas.com/en_us/software/visual-data-mining-machine-learning.html
- TensorFlow, https://www.tensorflow.org/
- Sci-kit Learn, http://scikit-learn.org/stable/

CONTACT INFORMATION

Your comments and questions are valued and welcomed. Please contact the author at

Kevin Lee
Director of Data Science
Clindata Insight
klee@clindatainsight.com

TRADEMARKS