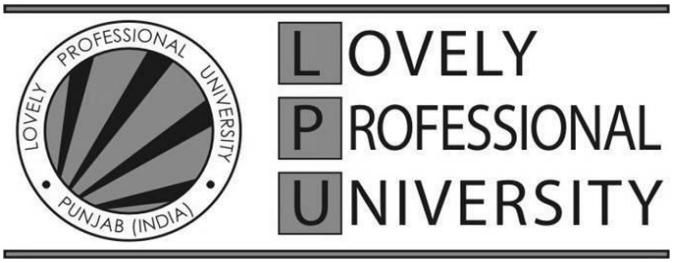



**SIX WEEKS SUMMER TRAINING REPORT**

on

# Machine Learning

Submitted by

Mayank Pratap Singh

**Registration No. : 11713175**

**Programme Name:** Bachelor Of Technology

Computer Science & Engineering

Under the Guidance of

**Sarthak Nigam**

**School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

(June-July, 2019)

## DECLARATION

I hereby declare that I have completed my six weeks summer training at Board Infinity from June 5th,2019 to July 20th,2019 under the guidance of Sarthak Nigam. I have declared that I have worked with full dedication during these six weeks of training and my learning outcomes fulfill the requirements of training for the award of degree of B. Tech CSE, Lovely Professional University, Phagwara.

(Signature of student)

Name of Student : Mayank Pratap Singh

Registration No. :11713175

Date: August 8th, 2019

**ACKNOWLEDGEMENT**

In preparation of my assignment, I had to take the help and guidance of some respected persons, who deserve my deepest gratitude. As the completion of this assignment gave me much pleasure, I would like to show my gratitude to Sarthak Nigam, Course Instructor, on Board Infinity for giving me good guidelines for assignment throughout numerous consultations. I would also like to expand my gratitude to all those who have directly and indirectly guided me in writing this assignment.

In addition, a thank you to Saurabh Singh, Yash Sinha, Aditya Mehta who introduced me to the Methodology of work, and whose passion for the “underlying structures” had lasting effect? I also thank the University for consent to take the course and pursue our passion.

Many people, especially my classmates(in the course) have made valuable comment suggestions which gave me an inspiration to improve the quality of the assignment.

# CERTIFICATE OF COMPLETION

THIS CERTIFIES THAT

## Mayank Pratap Singh

successfully completed the summer training program in the domain of Machine Learning

| July 2019 | BOARD INFINITY |
| --- | --- |
| ISSUED ON | ISSUED BY |

BOARD

**TABLE OF CONTENTS**

# INTRODUCTION

The term Machine Learning was coined by Arthur Samuel in 1959, an American pioneer in the field of computer gaming and artificial intelligence and stated that “it gives computers the ability to learn without being explicitly programmed”.
And in 1997, Tom Mitchell gave a “well-posed” mathematical and relational definition that “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Machine Learning is a latest buzzword floating around. It deserves to, as it is one of the most interesting sub-field of Computer Science. So, what does Machine Learning really mean?

Let’s try to understand Machine Learning in layman terms. Consider you are trying to toss a paper to a dustbin.

After first attempt, you realize that you have put too much force in it. After second attempt, you realize you are closer to target but you need to increase your throw angle. What is happening here is basically after every throw we are learning something and improving the end result. We are programmed to learn from our experience.

This implies that the tasks in which machine learning is concerned offers a fundamentally operational definition rather than defining the field in cognitive terms. This follows Alan Turing’s proposal in his paper “Computing Machinery and Intelligence”, in which the question “Can machines think?” is replaced with the question “Can machines do what we (as thinking entities) can do?”
Within the field of data analytics, machine learning is used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to “produce reliable, repeatable decisions and results” and uncover “hidden insights” through learning from historical relationships and trends in the data set(input).

Suppose that you decide to check out that offer for a vacation. You browse through the travel agency website and search for a hotel. When you look at a specific hotel, just below the hotel description there is a section titled “You might also like these hotels”. This is a common use case of Machine Learning called “Recommendation Engine”. Again, many data points were used to train a model in order to predict what will be the best hotels to show you under that section, based on a lot of information they already know about you.

So if you want your program to predict, for example, traffic patterns at a busy intersection (task T), you can run it through a machine learning algorithm with data about past traffic patterns (experience E) and, if it has successfully “learned”, it will then do better at predicting future traffic patterns (performance measure P). The highly complex nature of many real-world problems, though, often means that inventing specialized algorithms that will solve them perfectly every time is impractical, if not impossible. Examples of machine learning problems include, “Is this cancer?”, “Which of these people are good friends with each other?”, “Will this person like this movie?” such problems are excellent targets for Machine Learning, and in fact machine learning has been applied such problems with great success.

**Classification of Machine Learning**

Machine learning implementations are classified into three major categories, depending on the nature of the learning "signal" or "response" available to a learning system which are as follows:-

1. **Supervised Learning :** When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of Supervised learning. This approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples.
2. **Unsupervised Learning :** Whereas when an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of un-correlated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms.
   As a kind of learning, it resembles the methods humans use to figure out that certain objects or events are from the same class, such as by observing the degree of similarity between objects. Some recommendation systems that you find on the web in the form of marketing automation are based on this type of learning.

## TECHNOLOGY LEARNT

### 1. Programming in R

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac. This programming language was named **R**, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language **S**.

# Objects: character, numeric, integer, complex, logical

# Attributes of Objects

# Vector, Matrices, List, Data Frame. Factors : Create, Coerce, Edit, Convert

# Column Bind, Row Bind

# Data Table : Why, Create, Edit, Convert

## Creating variables

# R Object "atomic" Classes : character, numeric, integer, complex, logical

## R Objects : Vector

## Attributes : Name, Dimension, Class, Length

## 2. Learning Model Building in Scikit-learn : A Python Machine Learning Library

Important features of scikit-learn:

1. Simple and efficient tools for data mining and data analysis. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, etc.
2. Accessible to everybody and reusable in various contexts.
3. Built on the top of NumPy, SciPy, and matplotlib.
4. Open source, commercially usable – BSD license.

**Step 1**: Load a dataset
A dataset is nothing but a collection of data. A dataset generally has two main components:

Features: (also known as predictors, inputs, or attributes) they are simply the variables of our data. They can be more than one and hence represented by a feature matrix ('X' is a common notation to represent feature matrix). A list of all the feature names is termed as feature names.

Response: (also known as the target, label, or output) This is the output variable depending on the feature variables. We generally have a single response column and it is represented by a response vector ('y' is a common notation to represent response vector). All the possible values taken by a response vector is termed as target names.

Scikit-learn comes loaded with a few example datasets like the iris and digits datasets for classification and the Boston house prices dataset for regression.
Now, consider the case when we want to load an external dataset. For this purpose, we can use **pandas library** for easily loading and manipulating dataset.

In pandas, important data types are:

**Series**: Series is a one-dimensional labeled array capable of holding any data type.
**Data Frame**: It is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict. of series objects. It is generally the most commonly used pandas object.

**Step 2:** Splitting the dataset

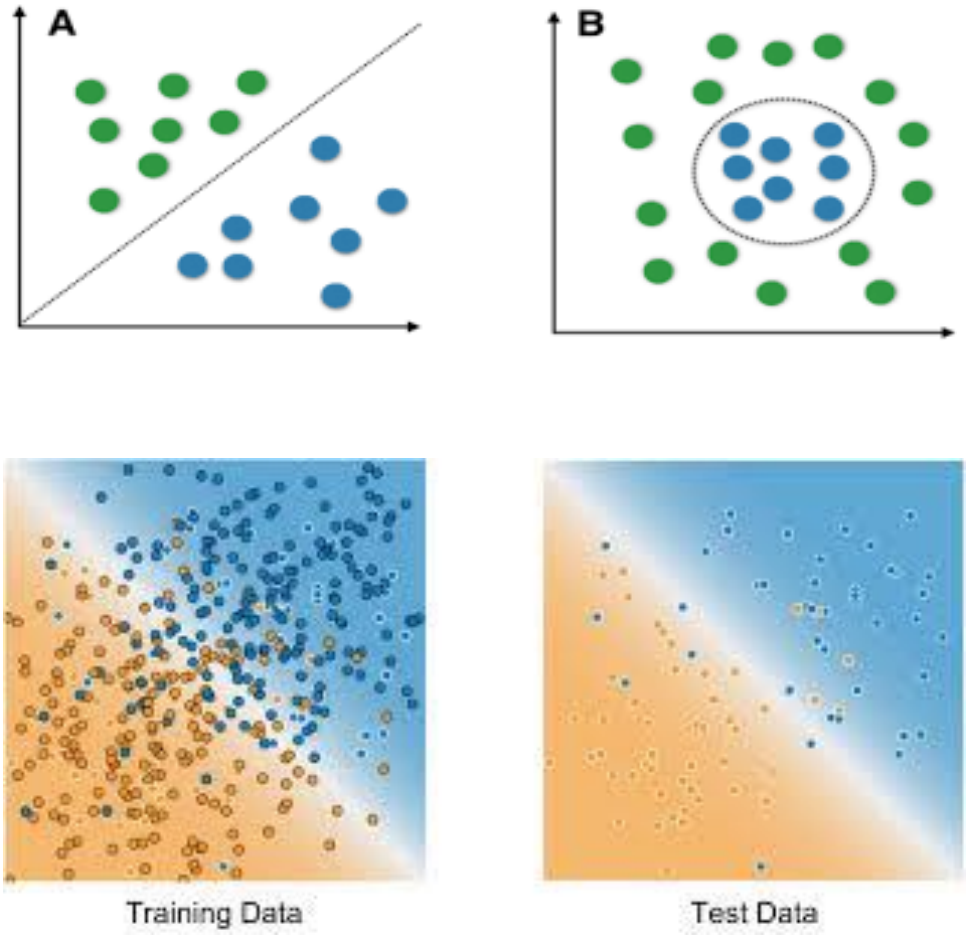
One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model.

But this method has several flaws in it, like:

1. Goal is to estimate likely performance of a model on an out-of-sample data.
2. Maximizing training accuracy rewards overly complex models that won't necessarily generalize our model.
3. Unnecessarily complex models may over-fit the training data.

**Step 3:** Training the model

Now, it's time to train some prediction-model using our dataset. Scikit-learn provides a wide range of machine learning algorithms which have a unified/consistent interface for fitting, predicting accuracy, etc.



### 3. Convolutional Neural Networks(CNN)

Convolutional neural networks are neural networks used primarily to classify images (i.e. name what they see), cluster images by similarity (photo search), and perform object recognition within scenes. For example, convolutional neural networks (ConvNets or CNNs) are used to identify faces, individuals, street signs, tumors, platypuses (platypi?) and many other aspects of visual data.

The efficacy of convolutional nets in image recognition is one of the main reasons why the world has woken up to the efficacy of deep learning. In a sense, CNNs are the reason why deep learning is famous. The success of a deep convolutional architecture called AlexNet in the 2012 ImageNet competition was the shot heard round the world. CNNs are powering major advances in computer vision (CV), which has obvious applications for self-driving cars, robotics, drones, security, medical diagnoses, and treatments for the visually impaired.

Convolutional networks can also perform more banal (and more profitable), business-oriented tasks such as optical character recognition (OCR) to digitize text and make natural-language processing possible on analog and hand-written documents, where the images are symbols to be transcribed.

CNNs are not limited to image recognition, however. They have been applied directly to text analytics. And they be applied to sound when it is represented visually as a spectrogram, and graph data with graph convolutional networks.

### 4. Using Google Colab for Model(s) Implementation

Colaboratory is a research tool for machine learning education and research. It's a Jupyter notebook environment that requires no setup to use. Colaboratory works with most major browsers, and is most thoroughly tested with latest versions of Chrome, Firefox and Safari. All Colaboratory notebooks are stored in Google Drive. Colaboratory notebooks can be shared just as you would with Google Docs or Sheets. Simply click the Share button at the top right of any Colaboratory notebook, or follow these Google Drive file sharing instructions.

**5. Natural Language Processing (NLP)**

Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software.

The study of natural language processing has been around for more than 50 years and grew out of the field of linguistics with the rise of computers.

Natural language refers to the way we, humans, communicate with each other.

Namely, speech and text.

We are surrounded by text.

Think about how much text you see each day:

- Signs
- Menus
- Email
- SMS
- Web Pages

The list is endless.

Now think about speech. We may speak to each other, as a species, more than we write. It may even be easier to learn to speak than to write. Voice and text are how we communicate with each other.

Given the importance of this type of data, we must have methods to understand and reason about natural language, just like we do for other types of data.

## REASON FOR CHOOSING THIS TECHNOLOGY

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new – but one that has gained fresh momentum.

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations big data – over and over, faster and faster – is a recent development. Here are a few widely publicized examples of machine learning applications you may be familiar with:

- The heavily hyped, self-driving Google car? The essence of machine learning.
- Online recommendation offers such as those from Amazon and Netflix? Machine learning applications for everyday life.
- Knowing what customers are saying about you on Twitter? Machine learning combined with linguistic rule creation.
- Fraud detection? One of the more obvious, important uses in our world today.

## LEARNING OUTCOMES

When you take a minute to stop and look around, the technological advancements of today could be perceived as something out of a futuristic novel. Cars are learning to drive, hands-free devices can turn on your lights or toast your bread, and flying drones are circling the skies. This is 2019. While the manifestation of Artificial Intelligence (AI) and Machine Learning (ML) haven't been realized, impressive progress has certainly been made.

This led us to ask, what can we do today that we couldn't do 10 years ago? And what does the future look like from here? While none of us have a crystal ball, that doesn't stop us from making educated predictions.

Let's say in **10 years**, AI/ Machine Learning will…

1) Increase security

**Drones are going to change the way we live**

2) Generate new services (and potentially social issues)

**Entirely new AI-based products and services will have created new consumer and industrial markets.**

3) Empower businesses

**The industry excitement around this technology and the rich experiences built on top of it I hope and believe represent the tides of change.**

4) Improve healthcare

**Artificial intelligence can access a much larger set of patient data of how they were treated and what the outcomes were.**

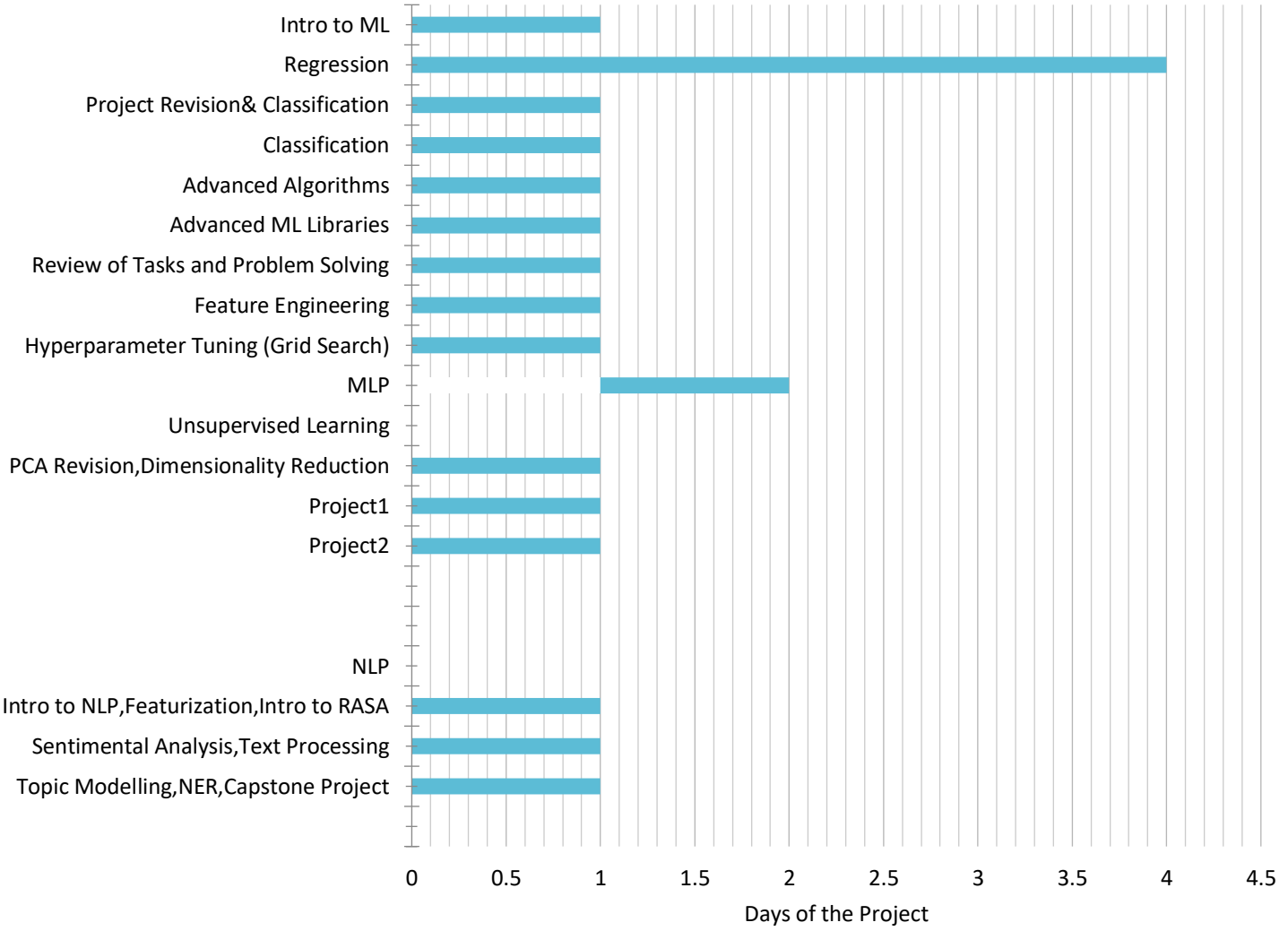
5) Make us smarter

**Ultimately, training an AI platform — it is very much like molding a child. If you treat it the right way and teach it the right things, train it to know what's right and wrong, it will inherently grow up to become a productive member of society that cares about people and the future. Just like any one of us.**

**6)** Project Link: **https://www.kaggle.com/mayank1212/kernelc1ac177b28/notebook**

# GANTT CHART

| TASK NAME | START DATE | END DATE | START ON DAY* | DURATION* (WORK DAYS) |
|---|---|---|---|---|
| | | | | |
| Intro to ML | 15/6 | 16/6 | Sunday | 1 |
| Regression | 17/6 | 17/6 | Monday | 4 |
| Project Revision& Classification | 20/6 | 21/6 | Thursday | 1 |
| Classification | 21/6 | 22/6 | Friday | 1 |
| Advanced Algorithms | 22/6 | 23/6 | Saturday | 1 |
| Advanced ML Libraries | 23/6 | 24/6 | Sunday | 1 |
| Review of Tasks and Problem Solving | 25/6 | 26/6 | Tuesday | 1 |
| Feature Engineering | 28/6 | 29/6 | Friday | 1 |
| Hyperparameter Tuning (Grid Search) | 29/6 | 30/6 | Sunday | 1 |
| MLP | 30/6 | 1/7 | 1 | 1 |
| Unsupervised Learning | | | | |
| PCA Revision,Dimensionality Reduction | 18/7 | 19/7 | Thursday | 1 |
| Project1 | 19/7 | 20/7 | Friday | 1 |
| Project2 | 20/7 | 21/7 | Saturday | 1 |
| | | | | |
| NLP | | | | |
| Intro to NLP,Featurization,Intro to RASA | 3/8 | 4/8 | Saturday | 1 |
| Sentimental Analysis,Text Processing | 4/8 | 5/8 | Sunday | 1 |
| Topic Modelling,NER,Capstone Project | 5/8 | 6/8 | Monday | 1 |


Intro to ML
Regression
Project Revision& Classification
Classification
Advanced Algorithms
Advanced ML Libraries
Review of Tasks and Problem Solving
Feature Engineering
Hyperparameter Tuning (Grid Search)
MLP
Unsupervised Learning
PCA Revision,Dimensionality Reduction
Project1
Project2
NLP
Intro to NLP,Featurization,Intro to RASA
Sentimental Analysis,Text Processing
Topic Modelling,NER,Capstone Project
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Days of the Project

# BIBLIOGRAPHY

| Source | | Material | Date of Implementation |
|---|---|---|---|
| Aditya Mehta (Supervised Learning) | Regression | https://colab.research.google.com/github/adityakshay/Machine-Learning/blob/master/Regression_Exercise.ipynb#scrollTo=aHoQhLQgqcPt | 16-05-2019 |
| | Classification | NA | 20-05-2019 – 21-05-2019 |
| | Based on Tensor-flow | https://colab.research.google.com/github/adityakshay/Machine-Learning/blob/master/tensorflow.ipynb | 22-05-2019 |
| | Advanced Algorithms | https://colab.research.google.com/github/adityakshay/Machine-Learning/blob/master/Advanced_Algortihms.ipynb | 23-05-2019 |
| | Feature Engineering | https://colab.research.google.com/github/adityakshay/Machine-Learning/blob/master/homeworkFE_Varun_Reddy.ipynb | 28-05-2019 |
| Yash Sinha (Unsupervised Learning) | PCA and Dimensionality Reduction | NA | 18-06-2019 |
| Natural Language Processing | Intro to NLP, RASA, Featurization | NA | 03-08-2019 |
| | Sentimental Analysis | NA | 04-08-2019 |
| | Topic Modelling | NA | 05-08-2019 |