A Brief History of Machine Learning

Norbert Cristian BERECZKI

April 2018

Abstract

One of the most actual and interesting topics of the new millennium is how to make machines learn and help our society. In this paper we will present a brief history of machine learning and how it came to be the hot topic that it is today. We will also put much emphasis on Deep Learning and its potential.

1 Motivation

In today's world, Machine Learning is an indispensable part of our life. Many of the technologies that are build in this time are based on Machine Learning. Moreover, it creates new jobs and automates many simple-to-do tasks. So, it looks more and more that people will stop doing mundane tasks and start concentrating on complex ones. Consequently, it is natural to investigate how it came to the capabilities that ML-tools provide and the direction the domain is heading towards.

This paper consists mostly of 2 chapters. The first one presents a more general view of how ML tools developed and the most important milestones that occurred. The second chapter puts emphasis upon Deep Learning(DL), and why, even though neural networks were invented a long time ago, it became so popular only in the past two to three decades.

2 A broader view of ML

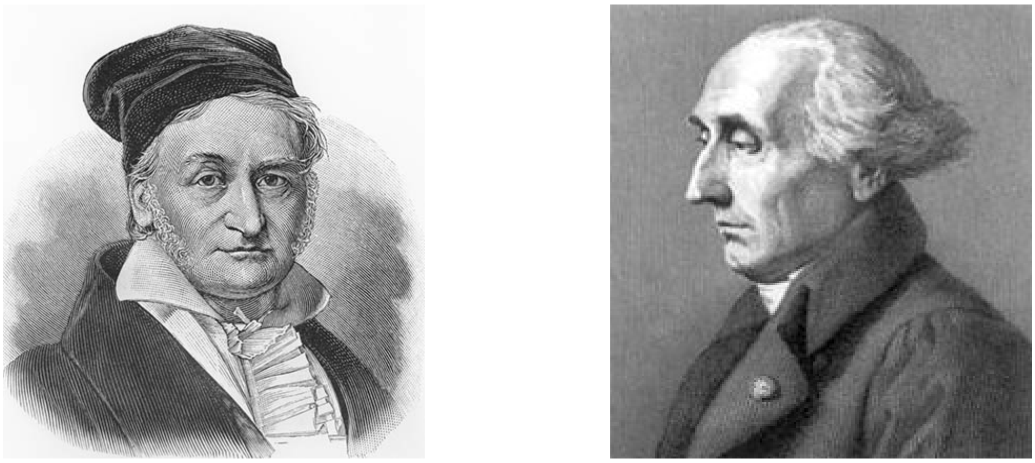
This section comprises of more subsections that describe parts of the Machine Learning.

2.1 Statistical Learning

We begin with this topic to emphasize its importance to the field in discussion. It is fair to say that without statistics there would be no Machine Learning nor there would be any advanced AI algorithm that we have today. However, we will only concentrate upon the part which includes *learning*.

Interesting enough, the first tabulating machine made by Herman Hollerith was created for the government to offer support for some statistical computations related to the population (1890). Statistical Learning is preoccupied with the employing of statistics into modeling data. Most of the state-of-the-art techniques use at some degree statistics, but in this part we will deal only with the ones that make much use of them.

The revolutionary work made by Lagrange and Gauss (in the 19th century) lead to the discovery of the *`linear least square method`* which was a sort of **`regression`** method. By the time of approximately 1950, more and more linear classifications and regression methods were developed. However only after 1970 most people started using these methods in practical applications such as predictive analysis, anomaly detection and classification. This is mostly due to the rise of the more powerful computers which could support complex computations and had much more memory and to the rise of non-linear classification and regression techniques. Not to forget the importance of the creation of the R language which boosted the field of statistics and provided very good (and optimized) tools for the development of statistical models and inference models. [10]



(a) Gauss

(b) Lagrange

Figure 1: Pictures of Gauss and Lagrange

2.2 Unsupervised Learning

Methods provided by this topic are widely used in subfields such as Data Science, Data Analysis and Business Intelligence. Unsupervised learning implies finding structure and ordering data where there is unlabeled data. So you will not be able to know for certain how well your model works.

The most trivial example for this is **`clustering`**. The apparition of clustering techniques (including here the most known clustering algorithm: **`k-means`**) is a little bit debatable many people referring to it as clustering or using k-means but not publishing it. Although, the first one to publish work related to the k-means algorithm is Stuart Lloyd in 1957. In the early 2000s the rise of data mining led to the invention of yet another clustering algorithm, namely the hierarchical clustering algorithm. Although it is not as used as other, more advanced algorithms, the hierarchical clustering algo had much popularity until approximately 2010s.[1]

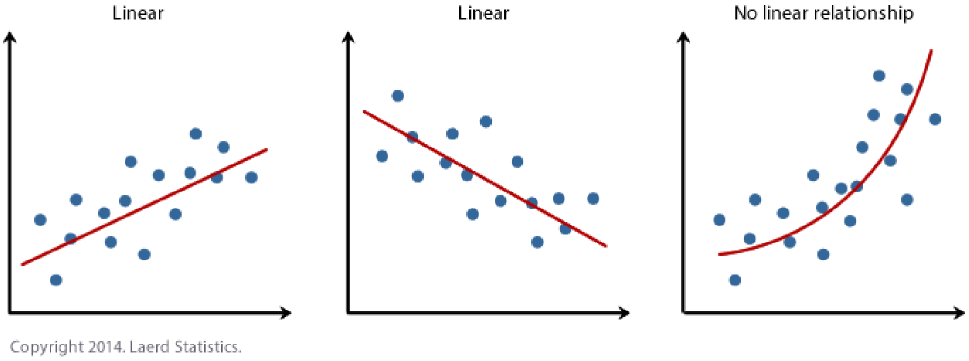
Other areas where unsupervised learning shows its prowess are the fields of Anomaly Detection, Structural Prediction and feature reduction (mainly PCA). The PCA method was created in 1901 by Karl Pearson (as [2] states) and since it has been used in many fields. However, the years 2000-2013 have known a large increase in papers upon PCA, mainly the year of 2013 where most of the papers on PCA were published, as stated in [1].

2.3 Supervised Learning

When taling about supervised Learning, it is usually implied that for a problem of regression or classification we must train a model, using some already known examples, to give us the desired output.

One of the most common set of techniques is the one regarding `regression analysis`. Here [3] it states that: 'Regression analysis is a set of statistical processes for estimating the relationships among variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.'

Figure 2: Image of fitting a regression relationship for some variables



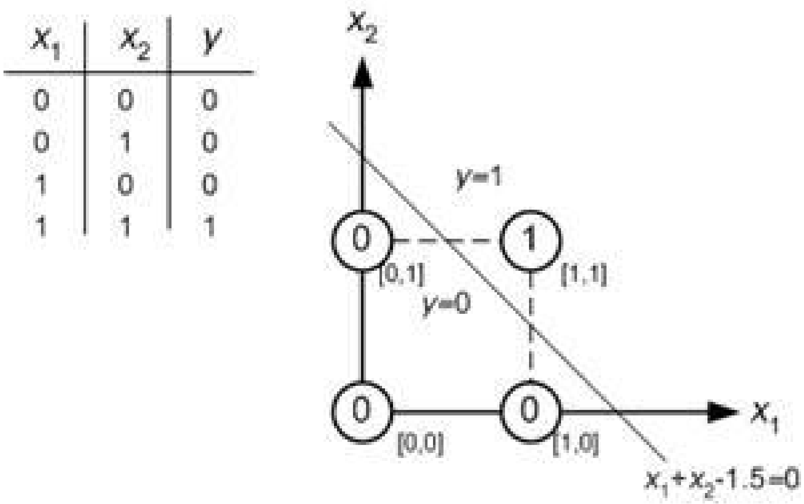
In this article [11] there is much being said about the early years of regression. We find out that Karl Pearson, who was at the time the first professor of Statistics in Britain chose Galton over Auguste Bravaris as the first one to offer a reasonable definition to regression. He accounts that the idea came to Galton when he was sheltering from a rainstorm. At that time Galton worked on eugenics and tried to see how geniuses can be bred. He was interested in the behaviour of outliers and if they will regress back to the mean. Methods for regression evolved over time by the works of Pearson and later Fisher. By the 1950s, regression was used in many branches of our society including economics.

3 Deep Learning

3.1 Artificial Neural Networks

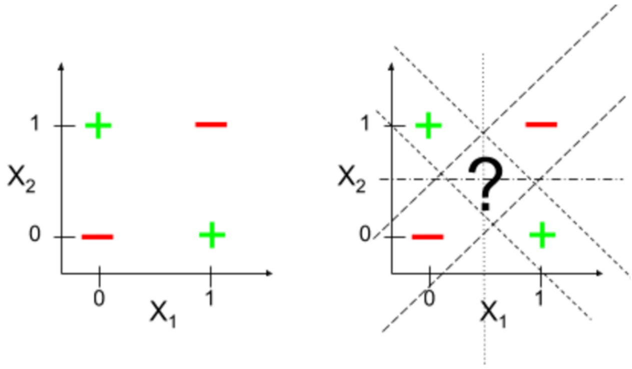
The roots of neural networks lies in the Hebbian Learning Theory which was proposed in the late 1940s. After that period, researchers tried to bring the hebbian learning model into the computational world. The first major breakthrough was made by Rosenblatt, namely the discovery of the `Perceptron` which could classify linearly separable classes. For instance, take the logical AND operation. You can draw a line to separate the output(0's and 1's) made by the AND function when inserting a 2-input value (Figure 1). In spite of this big breakthrough, much research did not follow until later on. This was due to mostly two things: firstly, the perceptron could not classify more complex functions such as the logical XOR, secondly, the computing power was weak compared to these days.

Figure 3: Example of the logical and operation



x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Figure 4: Example of the logical xor operation

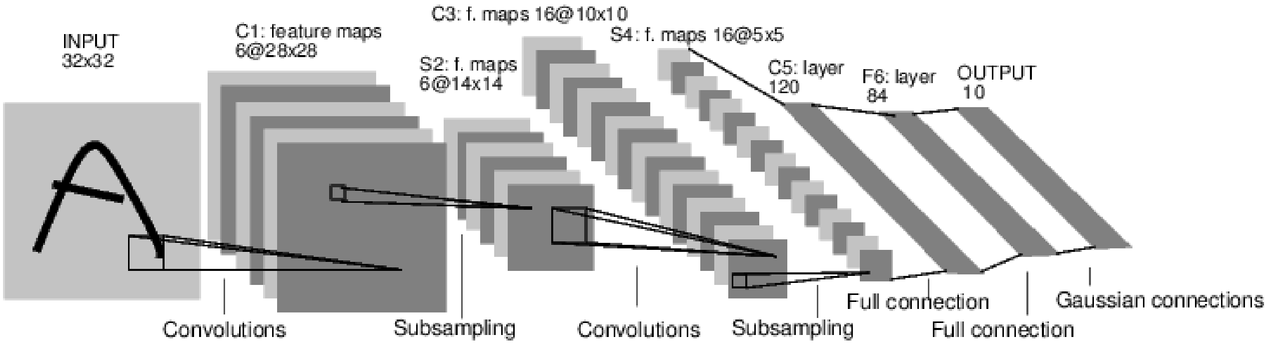


A big step futher was taken in 1975 by Werbos when he proposed a variant of the backpropagation algorithm, which was simply gradient descent applied recursively on each node, now having a network of perceptrons. Because of this, the network was able to learn the XOR function. G. Hinton had also worked to demonstrate the generalization of backpropagation on multi-layer networks. That was the spark that triggered the Deep Learning Revolution. To give some clarification, by Deep Learning we understand the methods to train ”deep models”, such as neural networks with many layers (more than 10).

3.2 Deeper Networks

This idea to stack layers in order to get deeper network stroke a revolution in Machine Learning. However, an important discovery was the Neocognitron (created by Kunihiko Fukushima in 1980) which was mostly a neural network that was classifying handwritten digits. This had a great impact on the work of Yann LeCun, called the ”Father of Convolutional Nets”. One of the first important works that LeCun[4] did, was during his PhD. and was on the backpropagation algorithm. He continued his work later on at Bell Labs (where, from his saying, he got his own cluster to experiment on). There, he created the first Convolutional Neurla Network called LeNet 1, which did handwritten digit classification. He improved his model until iteration 5 (LeNet 5). After that he went on working on the famous DjVu image compression system. His groundbreaking work enabled researchers to invest and develop a stil-growing set of tools related to Convolutional Neural Networks.

Figure 5: Architecture for LeNet 5 (which was used for digit recognition). The convolutional layers are shown in the image and at the end it can be seen the fully connected layers.



Other architectures began to rise, including: the Cresceptron(which had the job of detecting objects in a 3D space) and the SOFT (a weightless self-organizing feature extraction neural network). Although there were many issues related to deep learning, research stil continued. The issues were connected to the time of the training phase(for most models days or weeks were needed to train on even a small amount of data) and to the lack in labeled data. So in this time other methods such as SVMs(Support Vector Machines) and Gabor-Filter-based models were used for computer vision tasks and Hidden Markov Models for sequence modeling tasks.

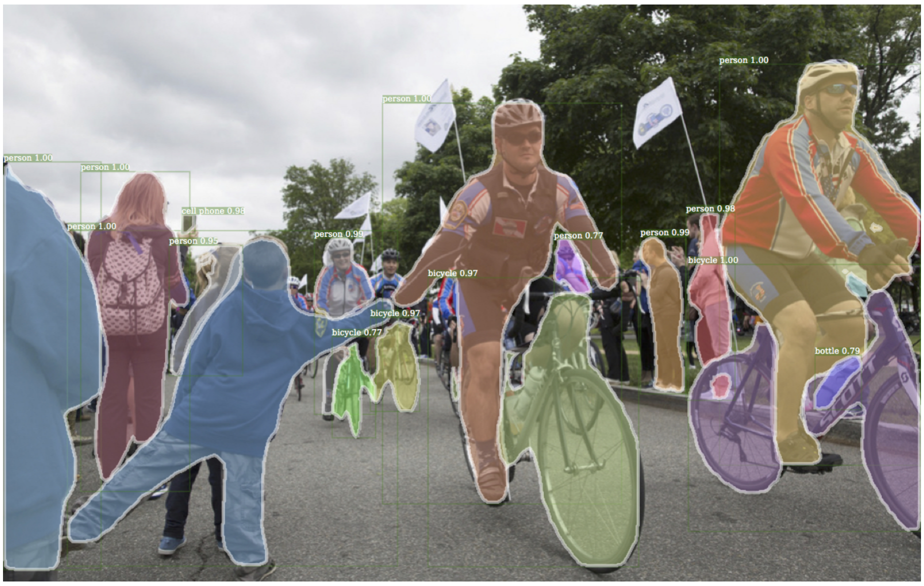
NSA and DARPA began to invest very much in SRI International in order to develop technologies that could achieve speech and speaker recognition. Although, they could not achieve very good speech recognition, they did it with speaker recognition using a deep neural network. Consequently, many researchers started investing time into speech recognition. The invention of LSTM (Long-short term memory), a type of Recurent Neural Networks(this are networks that have back connections between layers deeper in the net with layers at the beginnning of the net), had a lot of success in the task of speech recognition by the year of 2003. Google included this technique at industry level (embedding Recurent Neural Networks into Google Voice Search) after combining LSTMs with the novel CTC(connectionist temporal classification), which led 49% performance jumps. As of 2016 big companies such as Apple, Microsoft, Amazon, Google, started using more and more the LSTM approaches into products such as Siri, Amazon Alexa, etc.

Starting with 2009, people used NVIDIA gpus for their computing power and began reducing model-training time from weeks to days. This enabled people develop models faster and the ”bloom” of deep learning began.

One big achievement was done by Alex Krizhevsky, G. Hinton et. al, when they came up with AlexNet which won the ILSVRC(ImageNet Large-Scale Visual Recognition Challenge). This great achievement showed people that they can train successfully deep learning models on GPUs. After the AlexNet, there were done experiments on trying to stack more layers. However, these experiments showed that models with 17 layers could achieve greater classification accuracy than models with 30 layers. This is mostly because of exploding/diminishing gradients effect(which meant that the gradients that helped optimize your model would zero-out and would have no effect in minimizing the loss). This problem was solved by the development of Residual Neural Networks, where the network only needed to learn a residual function(meaning that it only needed to learn what to add or subtract from your input).

In the time the Residual Networks were developed, another strong mechanism rose up, namely the R-CNN (Region based Convolutional Neural Networks) which was the work of Ross Girshick et al. The R-CNN represented a new way to compute the bounding boxes of objects in images. While at Microsoft, Girshick worked on this R-CNN and developed Fast R-CNN and Faster R-CNN which achieved better performance in training and inference time and better accuracy. Building on these works, Kaiming He (the main researcher of Deep Residual Networks) et al. published a meta-algorithm, named Mask R-CNN, which would achieve great results in the image segmentation challenge. Image segmentation, usually means that the shapes of objets are extracted from an image.

Figure 6: Result of Mask R-CNN on an image



4 Conclusion and Fun Facts

4.1 Fun Facts

One interesting fact is that according to [5] 90% of the data has been produced in the last 2 years. This highlights the importance of data in supervised models. It is known that most of the deep learning models that are complex (have many layers) need large amounts of data to learn from.
In 2016[6], a farmer automated the process of classifying cucumber using deep learning. According to [7] in 1997 IBM's DeepBlue beat Kasparov at chess and in 2016 AlphaGo (produced by DeepMind) beat the world champion at Go. Both achievements were making use of deep reinforcement learning.

In 2010, a platform where people can compete using machine learning models was founded with the name of Kaggle. Until then more than 200 competition have taken place. These competitions helped in research areas including HIV research. Kaggle was aquired in 2017 by

Google, but nothing major changed.

In 2016 Amazon Go is launched[8]. It was the first store into which you would enter, take from the shelves the things you want to buy, then get out of the store and it would automatically subtract from your card the amount equal to the things you got from the store. This was implemented using advanced techniques of computer vision and until 22 january 2018 it was still under beta testing.

In 2014, according to [9] the chatbot ”Eugene Goostman” passed the Turing Test in which a person A sits in front of a curtain and asks the one after the curtain question, so that A would determine if the one after the curtain is a computer or a human. If person A decides that after the curtain there is a person and in fact there is a computer, the computer is regarded to have passed the Turing Test.

Moreover, Google had a big impact in the development of Machine Learning. Many of its employees say that it has become a company that puts ”Machine Learning First”. In 2017 they started building on distributed machine learning models by launching Google Cloud Machine Learning Engine and in 2018 they launched google.ai which is designed to attract more and more interest in machine learning and in the use of it.

4.2 Final words

In conclusion, it is obvious to underline that we have reached a point where more and more people are reasearching into machine learning every day. New products are build for it. Many tools are created to support large scale machine learning systems. And thus, in this time, machine learning is helping people fly, drive, detect diseases they have, support surgeries, help them interconnect and in many other ways.

Having an idea of how machine learning evolved over centuries we can understand better the direction it is heading now and understand why so many companies are pouring so much research into this field. However, we must hope that this helps us build a better world for tomorrow.

References

[1] URL: https://f1000research.com/articles/6-2012.

[2] URL: https://en.wikipedia.org/wiki/Principal_component_analysis.

[3] URL: https://en.wikipedia.org/wiki/Regression_analysis.

[4] URL: https://www.youtube.com/watch?v=JS12eb1cTLE&t=1063s.

[5] URL: http://www.mellanox.com/interconnected-planet/related-docs/machine-learning-infographic.pdf.

[6] URL: https://cloud.google.com/blog/big-data/2016/08/how-a-japanese-cucumber-farmer-is-using-deep-learning-and-tensorflow#closeImage.

[7] URL: https://en.wikipedia.org/wiki/Timeline_of_machine_learning.

[8] URL: https://en.wikipedia.org/wiki/Amazon_Go.

[9] URL: https://cloud.withgoogle.com/build/data-analytics/explore-history-machine-learning/.

[10] Gareth James et al. An Introduction to Statistical Learning. Springer, 2013. ISBN: 978-1-4614-7137-0.

[11] T.J. Barnes. “A history of regression: actors, networks, machines, and numbers”. In: Environment and Planning 30 (1998), pp. 203–223. DOI: http://journals.sagepub.com/doi/pdf/10.1068/a300203.