Author/s:
Juliandri, M;Ristanoski, G;Aickelin, U

Title:
Untangling the Concept of Artificial Intelligence, Machine Learning, and Deep Learning

Date:
2022-01-01

Citation:
Juliandri, M., Ristanoski, G. & Aickelin, U. (2022). Untangling the Concept of Artificial Intelligence, Machine Learning, and Deep Learning. Artificial Intelligence in Medicine Applications Limitations and Future Directions, pp.3-21. Springer Nature Singapore.

Persistent Link:
https://hdl.handle.net/11343/311467

Untangling the concept of Artificial Intelligence, Machine Learning, and Deep Learning

Muhammad Juliandri

School of Computing and Information Systems
The University of Melbourne
Melbourne, AU
juliandri.muhammad@gmail.com

Goce Ristanoski

School of Computing and Information Systems
The University of Melbourne
Melbourne, AU
gri@unimelb.edu.au

Uwe Aickelin

School of Computing and Information Systems
The University of Melbourne
Melbourne, AU
uwe.aickelin@unimelb.edu.au

<< the fame of AI, ML, and DL>>

The words *Artificial Intelligence (AI)*, *Machine Learning (ML)*, and *Deep Learning (DL)* have been used extensively in recent years. They suddenly become the top computer science terms that appear in other research domains such as *AI for medical diagnosis*, *ML in finance*, or *DL for linguistics*. Even the mainstream media like news or articles now promote the benefit of exploiting them to face Industry 4.0[1]. As a result, there is an increasing number of organizations that start adopting these concepts. They believe that AI, ML and DL can help them to improve business processes in their organizations.

<<why this chapter is important>>

Despite their growing popularity, not many people are able to clearly understand the main ideas behind AI, ML and DL. Most of them find it a challenge to tell the difference between them and use those terms interchangeably. This common misconception needs to be rectified since they are different in several aspects. In this first chapter we walk through the concept of AI, ML, and DL by showing the definition and methodology involved for each of them. We also discuss *Data Mining* (DM) as a separate concept as well.

[1] https://www.forbes.com/sites/bernardmarr/2018/09/02/what-is-industry-4-0-heres-a-super-easy-explanation-for-anyone/#448f66849788

A. Artificial Intelligence

1. Definitions and Objectives

<<Understanding AI by examples>>

The use of computing machines to assist humans in their tasks has been significantly evolving over the past few decades. Activities such as calculating numbers, storing documents, or reading articles can now be performed with a computer. However, none of these tasks can be categorized as a valid intelligent system because they still need a large portion of human instruction in reaching their end goals. Their computation is simply an execution of a quantified human intention, rather than a mimic of some form of human perception.

Contrary to such tasks, chatbot, machine translation, or autonomous vehicle are more fitted to be classified as applications of Artificial Intelligence. They require more than a simple function to be implemented on a machine in order to perform intellectual tasks – they require an amount of decision making conditioned by other inputs than human instructions. We can label them as systems that hold some intelligence (1).

<<The basic idea of AI>>

Provided with above examples, we define Artificial Intelligence as a collection of processes in a system with the capability to learn and solve a problem like a human (2). If a human needs a thinking process to solve a certain problem and a system could simulate the process into the expected solution, that means the system is leveraging Artificial Intelligence. However, this is not the only definition of Artificial Intelligence. There are in fact many different views on the definitions since the concept of Artificial Intelligence has shifted in accordance with the desired objectives.

<<AI definitions in literatures>>

Given the basic notion of AI, we observe other perspectives in literature. Russell and Norvig (3) highlighted 4 technical terms that briefly summarized multiple definitions on AI: *Acting Humanly*, *Thinking Humanly*, *Acting Rationally* and *Thinking Rationally*. They adopted the *Acting Rationally* as the essential characteristic of AI. There are two reasons for this: first is because it is more general compared to others, as in order to achieve intelligence, the machine does not only rely on a rational thinking but also on actions taken accordingly to any conditions; second, it is more adaptable since the improvement of Artificial Intelligence would not only revolve around human behaviour. The idea of making a machine smarter without explicitly giving it instructions is also supported by these two reasons. Hence, we consider this as the fundamental definition of AI.

<<Acting rationally in humans and machines >>

Despite that definition, we still have another question left that is how to implement AI so the machine can act rationally. Humans can act rationally because they learn from their environment. The way they learn is by experiencing it or it is something given as knowledge. We know fire is hot because we may have touched it, or we have been taught about this as other people had touched it before. The result of learning process is pairs of actions and consequences that can be used as responses. By comparing consequences, we can choose the rational response in a particular condition to obtain an expected result. A machine needs a similar process to act rationally. It needs to digest information in various form, extract knowledges, perform measuring consequences, and perform responses.

<<The growing domain inside AI>>

Interestingly, to build the machine that can respond to incoming information it would require knowledge from several domains. For instance, to make the machine have the capability of doing two-ways communication with humans, it needs to incorporate Natural Language Processing engine; to make the machine learn the pattern from the data, it needs Machine Learning techniques; and to make the machine distinguish objects, it needs Computer Vision proficiency. All of these have brought AI to a broader concept that covers all those domains. Thus, now we can see Natural Language Processing (NLP), Machine Learning (ML) or Computer Vision (CV) as some branches inside AI.

2. Branches of Artificial Intelligence

Natural Language Processing

<<Why does NLP require human intelligence?>>

Natural Language Processing (NLP) is Artificial Intelligence's branch that handles textual and linguistic data. This type of data is quite challenging to address because it may come in an unstructured form. Aa example of this type of data is medical patient records – the examination performed by doctor involves the documentation process of the patient's condition. By employing NLP this documentation can be conducted automatically by recording the conversation between the patients and doctor during examination (15). It can save the time taken for each examination and the doctor can focus more on the patient. However, this conversation data may have different accent, incomplete sentences, or any variation of voices. Yet, humans may easily understand this conversation and transform it into medical text while it is a difficult task for a machine to achieve the same goal.

Computer Vision

<<Why does CV require human intelligence?>>

While NLP deals with textual data, Computer Vision (CV) is another branch of AI that treats digital images as their primary source of knowledge. This kind of input is also challenging to process since it does not have any recognizable features that a machine can learn from. To better illustrate this complexity, let us assume that we need to identify pneumonia from frontal chest radiograph images (15). When radiologist see those images, they can directly interpret the possibility of pneumonia without much conscious cognitive effort, but for a machine, this picture is just a set of pixels represented as numbers. Each of the object's characteristics like colours and shapes need to be extracted from this numeric representation before the machine can detect if a patient has pneumonia or not.

Recommendation System

<<Why does RS require human intelligence?>>

We are often suggested with some suggestion videos on streaming service platforms. It looks like they certainly know what kind of videos that are relevant to us. Different people will have different video recommendations, but more interestingly these suggestions happen without having us to search for videos manually. The reason for this is because they are adopting Recommendation System (RS) in their platforms. Recommendation System can be seen as an engine that provides a user with relevant items, products, or even services. In the context of streaming platforms, the items can be videos or songs. However, the discovery process of these

relevant items is challenging as the engine needs to understand how relevant an item is for a user. In addition, this relevancy is measured on large collections of items where the manual approach will be infeasible.

B. Data Mining

1. Definitions and Objectives

<<Challenges in extracting knowledge from data>>

Having defined the idea and branches of AI, we can clearly see that data plays a central role in achieving intelligence since it contains a lot of valuable knowledge. If the machine is able to obtain that knowledge, then it may be able to act like humans. However, finding the knowledge from data is a difficult task due to many factors. Occasionally data produced may grow significantly over a period of time. This affects an exhaustive process of knowledge identification. In addition, data may also consist of many attributes that need to be considered. An advanced technique is required to perform the knowledge extraction. In such a case, we can incorporate Data Mining as a technique to handle that process.

<<Data Mining in literature>>

Data Mining (DM) actually is not a new concept even though its popularity just accelerated recently. The term itself was coined in the 1990s by database community (4). However, this term is inaccurate as what we want to discover is not data, but useful information trapped in it. That is also the reason why many researchers formerly called Data Mining as *Knowledge Discovery from Data* (KDD) (5). Additionally, in KDD, the definition is presented in a wider context including data pre-processing, data transformation, data mining, and data evaluation.

<<Data Mining vs Machine Learning>>

Provided with Data Mining definition, it will be also necessary to know the differences between Data Mining and Machine Learning. This is because Data Mining is commonly compared to Machine Learning (ML). Even though the clear line that separates them is now hardly to be seen, we are still able to find the distinctions. While Data Mining explores knowledge from a pattern, Machine Learning is more concerned about the algorithm to extract that pattern (6). As a result, Data Mining used a lot of Machine Learning techniques. Another significant difference between them is the objectives. In ML, the primary objectives involve making improvements to the machine based on the experience so it can predict the future well. Data Mining on the other hand, is leveraged massively on finding patterns, looking for trends, or discovering outliers on the current data. Typical examples in medicine would be multi-objective semi-supervised clustering to identify health service patterns for injured patients (16).

2. Data

<<Understanding the data properties>>

Before diving into some techniques used for Data Mining, it is necessary to understand a common terminology used to describe an essential part of data. Data is used as the input of a learning process in a machine and it can be divided into three different parts: features, labels, and instances.

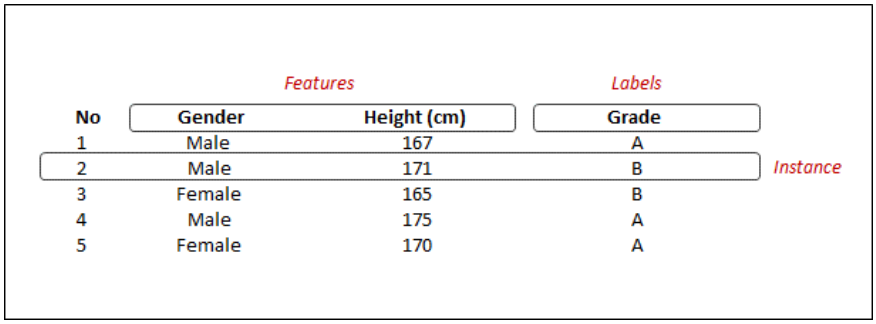

Figure 1. Instances, Features, and Labels in Data

<<Explanation of Features, Labels, and Instances>>

Features are part of data that represent characteristics or attributes in our observation while labels are part of the data that we aim to learn or predict. One important thing to note is a label is actually a feature, but its value can be determined by other features. Hence, the feature and label pairs are also known as explanatory and response variables. Moreover, instances can be considered as individual parts of data known as examples or samples. To illustrate how they relate to each other, assume that we are given data consisting of genders and heights to predict grades of each person (Figure 1). In that case, the features are genders and heights; the labels are grades; and an instance is each person's data.

We can separate the data based on their value into two major types: Categorical Data or Numerical Data.

Categorical Data

<<What is categorical data?>>

When the information is presented in a way that describes a certain type, most often in a distinct and discreet format, and an order of the value may or may not be present, then we are referring to this type of data as categorical data. The values in categorical data are denoted as classes. Examples of categorical data are blood type, age group, education level, location(by name), food group etc.

Numerical Data

<<What is numerical data?>>

In contrast to categorical data, numerical data deals with a wide range of continuous numbers. It often involves a real value number such as weights, blood pressures, and ages. The important thing with numerical data is we can perform any mathematical operations with it unlike the categorical data.

3. Methodology

So far, we have presented an overview of Data Mining and how it employs Machine Learning techniques to learn the pattern. For that reason, in this section we shall show two broad classes of Machine Learning techniques that are extensively used for discovering patterns in Data Mining: Classification and Clustering.

Classification

<<What is classification?>>

Classification is a strategy in supervised learning to find a relation between features and labels in the data. In supervised learning, data comes with labels. The most important thing with this strategy is it is only applicable for the labels of categorical data. Once the relation is extracted, then it can be used to predict labels of the unseen data. A number of applications used classification in detecting patterns such as spam detection, weather forecasting, and medical diagnosis.

Clustering

<<What is clustering?>>

Clustering is another strategy in data mining to find patterns from the data. It is part of Unsupervised Learning as the data does not come with labels. We can only compare one instance to another instance. If these two instances shared similar features, then we can group them into one cluster. Otherwise, that instance forms a separate cluster with other instances matching its characteristics. Some applications of clustering are topic modelling, anomaly detection, and image segmentation.

C. Machine Learning

1. Definitions and Objectives

<<Machine Learning's definition>>

In the previous section, we see Machine Learning from the perspective of Data Mining without explaining its actual definition. In literature, we may find many definitions of machine learning introduced by *scientist*s. Despite some differences in definitions, most of them agree that machine learning in general can be thought of as a program that can learn from the past information to predict the future information. We put an emphasis on words *learning* (7) and *predicting* (8) since these two also represent essential processes in Machine Learning.

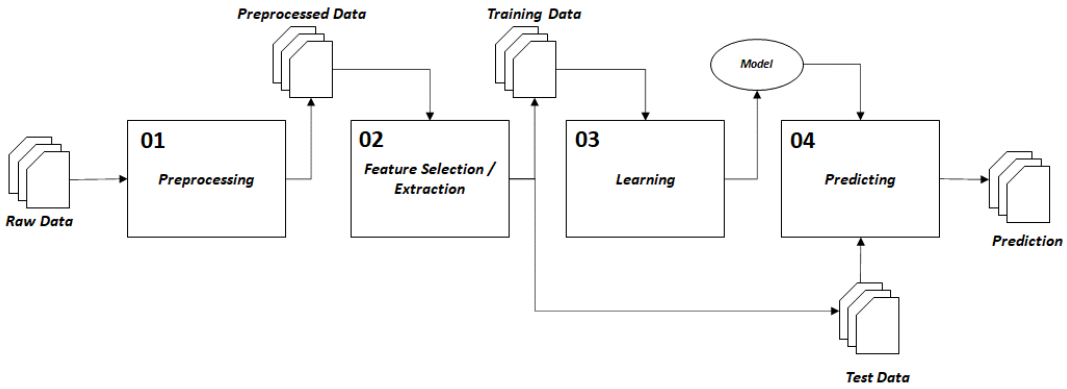

Figure 2. Machine Learning's Workflow

<<Common workflow in Machine Learning>>

However, learning and predicting are not the only processes. It may be preceded with two additional processes, *Pre-processing* and *Feature Engineering* (Figure 2). Pre-processing is part of transforming the raw data into a cleaned version and this step is usually adopted if we have

textual data. An example of this is to clean the document from all html tags leaving only the relevant contents of interest. Following the pre-processing we may also perform feature engineering to extract attributes from data. We need this since the data may not come with identifiable attributes so we need to extract them before it can be processed further. For instance, in textual data, we need to convert them into a numeric representation because some techniques in machine learning only accept numbers as their input.

<<Main factors behind machine learning's performance>>

Provided with the workflow, we can see that machine learning performance is highly dependent on two factors. The first factor is the input data. The larger the data means the machine can learn more, but it is also important to note that this would influence the running time. Therefore, we need to consider carefully if we want to use more data. The second factor is the learning technique to extract the knowledge (*model*). We should pick the appropriate learner based on the data types and goals that we want to achieve. Failing to pick the correct learner would also affect the performance. Hence, it is necessary to understand two big categories of learning: *Supervised Learning* and *Unsupervised Learning*.

Supervised Learning

<<Overview of Supervised Learning>>

In supervised learning, the model is generated by giving the labelled data to the learner. What we mean by labelled data is the data comes in pairs, the input with its corresponding output (8). The learning process can be seen as finding the best model that maps that input to the output. We can divide the supervised learning based on the types of output. We call it classification if It handles the categorical data as the output. *Decision Tree*, *Support Vector Machine*, and *K-Nearest Neighbours* are examples of classification techniques in Machine Learning. Alternatively, If the output is a continuous number, then we call it Regression. Some of the regression techniques are *Linear Regression, Polynomial Regression,* and *Ridge Regression*.

Unsupervised Learning

<<Overview of Unsupervised Learning>>

While in supervised learning we have a labelled data in our disposal, in unsupervised learning we may not have or need it. It is because sometimes we cannot define all the possible output or class in the data. We are still able to extract some knowledge from the data. It is performed by measuring the similarity among inputs. For instance, in the topic modelling task, we use the number of overlapping words as the similarity metrics. As a result, documents with a high number of overlapping words will form a cluster. In such a task, we call this technique as clustering since the main goal is to group the similar objects in a cluster. However, this is not the only technique covered in unsupervised learning. *Outlier Detection*, *Recommender System*, and *Association* are other techniques that fall within this category.

2. Feature Engineering

<<Why feature extraction?>>

Despite the fact that feature engineering is not necessarily to be performed in machine learning, for some cases we would not be able to continue to the next process without having it done. For example, in a topic classification task, we have documents as the input. As a result, we need to convert this document to its numeric representation since the optimization algorithm is only able

to handle numerical data. One way to meet this requirement is to extract features from documents such as the size of document, the number of special characters and the number of stop words. They can be the distinctive characters to differentiate one document to another. However, we may also use more advanced techniques of feature extraction such as *bag of words* and *term-frequency inverse document-frequency* (TF-IDF).

<<Why feature selection?>>

In addition to features extraction, there is another feature engineering technique called feature selection. The aim of this process is to select the most important features in the data to obtain a better performance. It is because irrelevant features would make the solution not unique and the learning process slower. A further instance of this is having size of documents and number of characters as features. One of these features can be ignored since they are highly correlated to each other. We can obtain the size of documents based on the number of characters and vice versa. As a result, we can have a lower number of features involved in the learning process and the time to build the model can be reduced. Feature selection can be executed with a variety of methods such as *Principal Component Analysis*, *Akaike Information Criterion (AIC)*, and *Singular Value Decomposition (SVD)*.

3. Learners

In this section, we explain the linear model for regression and classification. These two learners are quite important in Machine Learning since they form a foundation for more complex models (9).

Linear Regression

<<Example of simple linear model>>

In high school, we learnt how to generate a linear equation that goes through two data points. This equation is usually expressed in the form of:

$$y = mx + c$$

where m is the slope, and c is the intercept. If we have two data points called A and B, with their coordinates: (1, 3) and (2, 5), then we can calculate the slope and intercept for a straight line that passes through A and B in following steps:

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{5 - 3}{2 - 1} = 2$$

$$c = y - mx = 5 - 2\,(2) = 1$$

Having the slope and intercept, we can plug them into the variables to produce the full equation.

$$y = 2x + 1$$

If a new point C comes with $x = 10$ then using that equation, we can predict the value of y = 21. That linear equation is known as the model of those data points. However, what if we have more than two data points and there is no straight line that runs through them? In such a case, we still need to find the closest line to be the model of that data. One approach to find that line is using Linear Regression technique.

<<the idea behind Linear Regression>>

Linear regression finds the fittest line that represents the relationship between features x and labels y. In measuring the fittest line, linear regression employs a performance metric and a commonly used metric is sum of squared error. This sum of squared error is computed by comparing the actual label y and the predicted value $\hat{y}$ for each data. Hence, the smaller this value is the better the model generated.

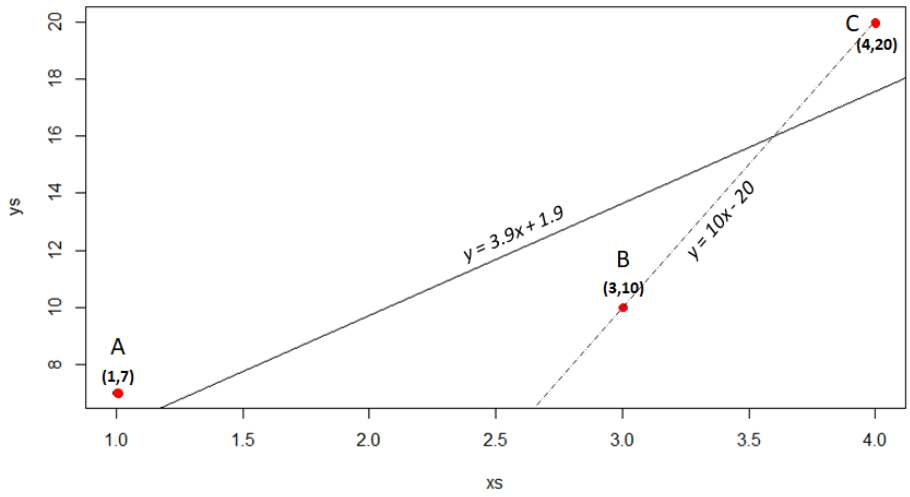

Figure 3. Two different model for same data

<<Evaluating fittest line using Sum of Squared error>>

In Figure 3, we present two different linear models for three data points: A, B and C. The first model is shown by the dashed line that runs through two data points (B &C) and the second model is depicted by the solid line which did not sweep any points. At a glance, the dashed line looks like a fitter model for the data. However, the sum of squared error generated by the dashed line is much larger compared to the solid line. The sum of squared error for the first model is

$$\sum(\hat{y} - y)^2 = (-10 - 7)^2 + (10 - 10)^2 + (20 - 20)^2$$

$$= 289$$

while the error for the second equation is

$$\sum(\hat{y} - y)^2 = (5.8 - 7)^2 + (13.6 - 10)^2 + (17.5 - 20)^2$$

$$= 20.65$$

Hence, we prefer to use $y = 3.9x + 1.9$ as the model to precisely describe the data. However, in a real case, the linear models for three data points above can be infinite in number. Therefore, we need a shortcut to find the most suitable line and it can be done by minimising the sum of squared error. This shortcut is known as *the method of least square*.

<<Linear regression for more complex model >>

The linear regression model is not only able to handle a single variable problem like above but far beyond that. They can also address multidimensional data. In fact, they can also be used to generate more complex models like polynomial functions to fit a non-linear data.

Logistic Regression

<<Issues with Linear Regression>>

Often, we are faced with a problem that has only two different possible outcomes. For instance, to model if a person had a farsightedness issue based on his age. We can define this as a linear regression model, but it would not fit well at least for two reasons. First, linear regression may produce output outside of the expected values. Note that it can even produce a negative value as the outcome. Second, the larger the value of a feature may greatly influence the error for the model. In fact, for a binary classification problem, this cannot become the issue otherwise the generated model would be susceptible to outliers.

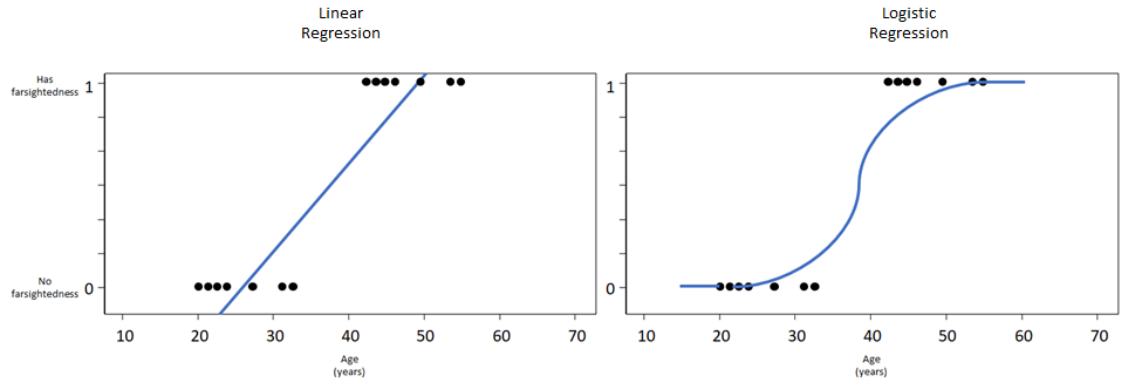

Figure 4. Linear Regression and Logistic Regression function

<<the idea behind Logistic Regression>>

Instead of using a linear regression as the learner, we need to use binary classifiers for such a case and logistic regression is one of them. Logistic regression used a logistic function to fit the non-linear model. It has the characteristic to map all real numbers into numbers between 0 and 1. This is the main factor why this function is suitable for classification. In terms of appearance, the logistic function has an s-shaped line (Figure 4). The formula of this logistic function is as follow:

$$P\,(y\ =\ 1\,|\,X\ =\ x)\ =\ \frac{e^{mx\,+\,c}}{1\ +\ e^{mx\,+\,c}}$$

To make it as linear model, it is usually transformed into the logit (*inverse of logistic function*) where it adopts the concept of probability and odds. Probability can be thought as how likely an event occurred, denoted as p, while the odds can be defined as the ratio between the probability of event occurred and probability of it does not occur:

$$odds\ =\ \frac{p}{1\ -\ p}$$

Having the odds defined, the previous equation then can be converted into a new expression:

$$log\ (odds)\ =\ mx\ +\ c$$

<<Evaluating fittest model using MLE>>

To generate the most fitted model for the data, unlike linear regression which uses a *least square method*, logistic regression employs a *Maximum Likelihood Estimator* (MLE). The idea behind MLE is to find parameters that will increase the likelihood of the outcome most.

<<Logistic regression for more complex model >>

It should be noted that all the equations shown above are only applicable for a single feature model. If we have more than one feature, then it should be modified accordingly. For example, if there are two features, then $mx + c$ should be replaced by $m_1x_1 + m_2x_2 + c$ since different features will have their own slope (*parameter*).

D. Deep Learning

1. Definitions and Objectives

<<The state-of-the-art classifiers in Machine Learning>>

Deep Neural Networks or Deep Learning has become one of the popular methods in machine learning to address difficult tasks ranging from computer vision to pattern recognition (10). In computer vision domain, Deep Learning can assist systems to recognize images of handwritten digits and classify them to correct labels automatically (11). This capability was not limited to digits, but also to other objects in images such as planes, cars, or animals (12). Having the knowledge that makes Deep Learning a good classifier, many works propose this method as their core algorithm to perform more complicated tasks in related fields such as robots (13) and self-driving cars (14).

<<Understanding the idea core behind deep learning>>

Provided with many powerful applications above, we have important questions to answer. *What is deep learning and why it is so powerful?* Before we jump into deep learning, we need to know what perceptron is. Perceptron is a binary classifier that uses the concept of neuron. A neuron in our brain is a unit that processes some signal and transmits it to other neurons. This process will influence how strength the output signal will be. For instance, humans can hear sounds and see images because of this processed signal. Perceptron adopts this concept by introducing a single neuron to process input and produce an output in classification tasks. This is also the main reason why perceptron is also called as an artificial neural network.

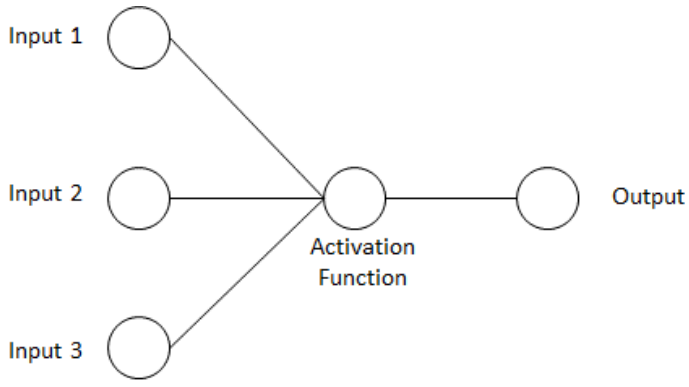

Figure 5. Perceptron Model

In a perceptron, the process of transforming some information into another information can be seen as a certain mathematical operation. This operation is called an activation function (Figure 5). Recall that in logistic regression such mathematical operation is a logistic function. In other words, perceptron is similar to logistic regression if we use logistic function as the activation function. The perceptron model, however, is not quite robust primarily if the input is not linearly separable. Put it simply, there is no straight line or plane that can divide the data into two

different classes. In addition, perceptron is only applicable for binary classification. Those are the main factors behind the invention of multilayer perceptron.

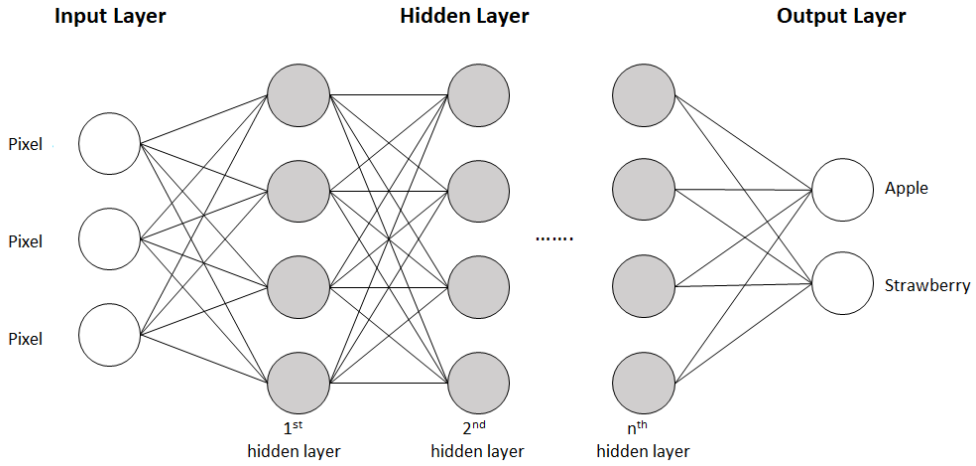

Figure 6. Deep Learning Model

<<What is Multilayer Perceptron?>>

In multilayer perceptron, we may have more than a single hidden layer and for each hidden layer we also can have more than a single neuron (Figure 6). For this reason, multilayer perceptron is also known as deep neural network (deep learning). By adopting more neurons and layers, it can solve the non-linearly separable issue and build more complex functions. It can be thought of as a composite function in mathematical theorem where we apply another function to the result of the previous function. As a result, this would transform our data into a higher dimensional representation. Imagine this as trying to separate data points into two classes either by using a straight line or curved line. Certainly, the accuracy would be better if we use the curved line instead of straight line for a complex problem. This answers our question why deep learning is powerful to approximate all classification tasks.

2. Variants

There are several variants of Deep Learning that are very popular nowadays: *Feedforward Neural Network*, *Convolutional Neural Network*, and *Recurrent Neural Network*. The variants differ in the way they process the input.

Feedforward

Feedforward neural network is the most common neural network used in many applications. It is a standard multilayer perceptron with a fully connected neuron (Figure 7). Feedforward neural networks learn by randomizing the initial value of parameters. After that, it will pass this value as the input for the neuron in the next layer. The way it passes the value to the next layer is the main reason why it is called feedforward. If the prediction is different with the actual output, then the parameter value may need to be reduced or increased. Otherwise, no penalty will be given to the current value of parameters.

Convolutional

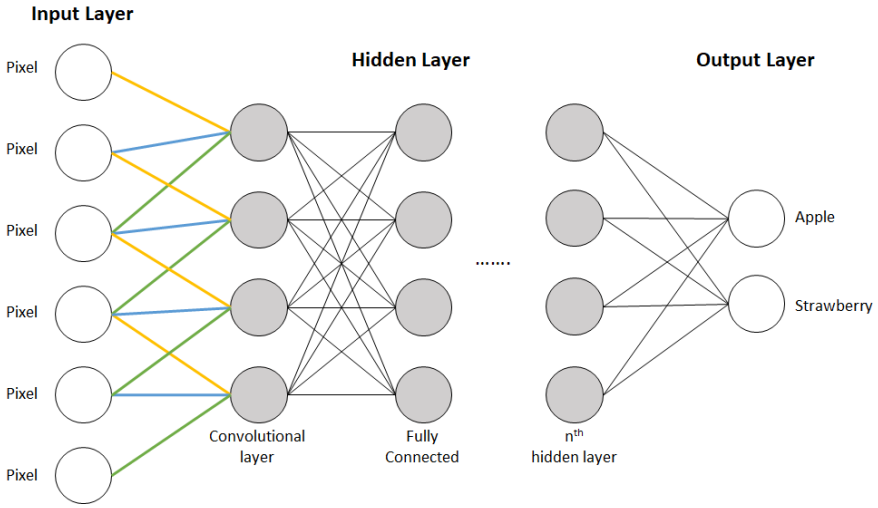

Figure 11. Example of Convolutional Layer

In contrast to fully connected neural networks, convolutional neural networks have some layer in the architecture, where its neuron is computed only by a specific number of neurons in the previous layers (Figure 11). This can be realized by using a sliding window and filter. The filter aims to extract some pattern in the images. This approach is believed to generate a more accurate prediction primarily in the visual recognition domain. It is because an object is constructed by multiple smaller objects which is only influenced by the pixels around the neighbourhoods. A further instance of this is in handwritten digits classification. Using a filter size of 3 x 3 means that it only considers the pixels in the region of 3 x 3 and as a result the filter will detect edges for each number as a pattern. For example, digit '4' can be seen as a construction of 3 pattern: left vertical edge, middle horizontal edge and right vertical edge (Figure 12).

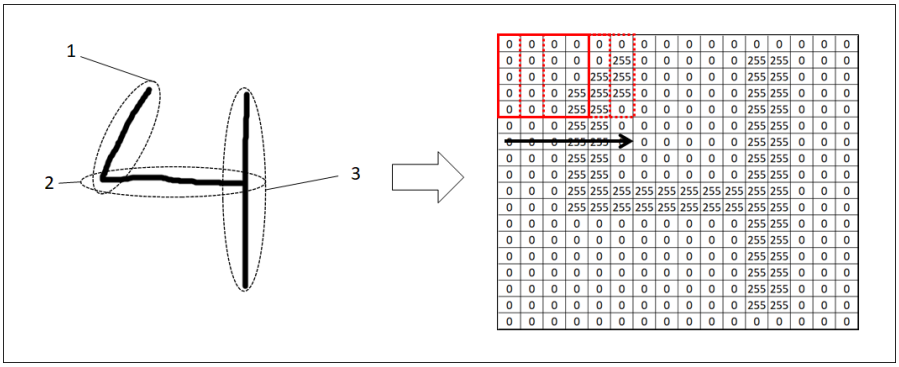

Figure 12. Illustration of handwritten digit construction.

Recurrent

Occasionally sequence of the data contains some knowledge that can be extracted to predict the output more precisely. A further instance of this is in the language generation model where we are provided with a sequence of words as the input and need to predict the next word. In such a case, the order of words in the input will influence the probability of the next word. If we know the input is "recurrent neural" then we can easily predict that the subsequent word is "networks".

Recurrent neural network exploits this idea in its architecture. Instead of processing the input in parallel, it operates sequentially. The reason for this is because they need some information from

the previous input to be processed along with the next input. This information is called a state vector with purpose to maintain the context. In terms of architecture, recurrent neural network is similar with feedforward neural network, but the neuron will have another edge that points to its own (Figure 13).

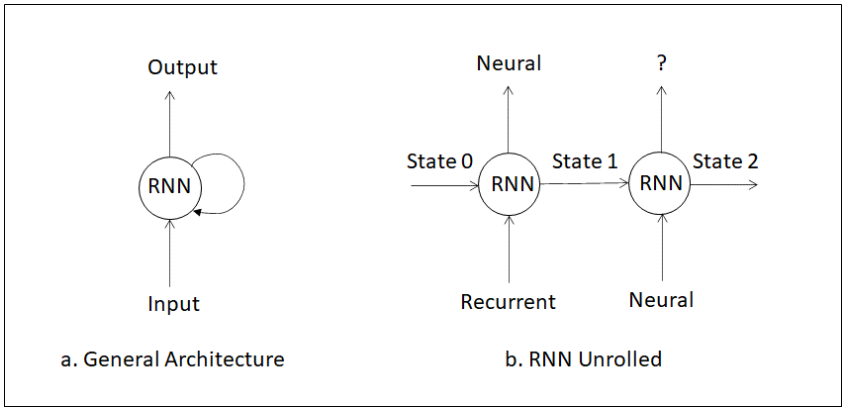

Figure 13. Recurrent Neural Network architecture

References

1. Nilsson NJ. Principles of Artificial Intelligence. Illustrated, reprint. the University of Virginia: Morgan Kaufmann Publishers; 1986.
2. Akerkar R. Artificial Intelligence for Business. First Edition. Springer International Publishing; 2019.
3. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. Third Edition. New Jersey: Prentice Hall; 2010.
4. History of Data Mining [Internet]. [cited 2020 Aug 11] Available from: https://www.kdnuggets.com/2016/06/rayli-history-data-mining.html
5. Han J, Pei J, Kamber M. Data Mining: Concepts and Techniques. Third Edition. Elsevier; 2011.
6. Tan PN, Steinbach M, Kartpane A, Kumar V. Introduction to Data Mining. Second Edition. New York: Pearson; 2019.
7. Mohri M, Rostamizadeh A, Talwalkar A. Foundation of Machine Learning. Second Edition. MIT Press; 2018.
8. Murphy KP. Machine Learning: A Probabilistic Perspective. London: MIT Press; 2012.
9. Bishop CM. Pattern Recognition and Machine Learning. Cambridge: Springer; 2006.
10. Deng L, Yu D. Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing. 2014; 197–387. doi: 10.1561/2000000039
11. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE. 1998; 309 – 318. Retrieved from http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf
12. LeCun Y, Bengio Y, Hinton G. Deep learning. 2015; 521(7553), 436. Retrieved from https://link-gale-com.ezp.lib.unimelb.edu.au/apps/doc/A415563174/
13. Lee J, Ryoo MS. Learning Robot Activities from First-Person Human Videos Using Convolutional Future Regression. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2017; 1-2. Retrieved from http://openaccess.thecvf.com/content_cvpr_2017_workshops/w5/html/Lee_Learning_Robot_Activities_CVPR_2017_paper.html

14. Kim J, Canny J. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. The IEEE International Conference on Computer Vision (ICCV). 2017; 2942-2950. Retrieved from http://openaccess.thecvf.com/content_iccv_2017/html/Kim_Interpretable_Learning_for_ICCV_2017_paper.html
15. Erik JT. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. Nat Med 25; 2019.
16. H. A. Khorshid, U Aickelin, G Haffari, B Hassani-Mahmooei, 'Multi-objective semi-supervised clustering to identify health service patterns for injured patients,' Health Information Science and Systems, vol.7, no.1, pp. 18, 2019