The Computational Principles of Learning Ability

Hao Wu

MSc by Research
University of York
Computer Science
September 2015

Abstract

It has been quite a long time since artificial intelligence (AI) researchers in the field of computer science have stopped talking about simulating human intelligence or trying to explain how the brain works. Recently, represented by deep learning techniques, the field of machine learning has been experiencing unprecedented prosperity and some applications with near human-level performance bring researchers confidence to imply that their approaches are the promising candidates for understanding the mechanism of the human brain[1][2]. However apart from several ancient philological criteria and some imaginary black box tests (Turing test, Chinese room) there is no computational explanation, definition or criteria about intelligence or any of its components. Based on the common sense that learning ability is one critical component of intelligence and from the viewpoint of mapping relations, this paper presents two laws which explain what “learning ability" is, as we familiar with it and under what conditions a model can be acknowledged as a “learning model". Furthermore, corresponding corollaries prove the existence of a common learning model (L), and by comparing with traditional learning theory with the theoretical framework proposed in this dissertation, the author explains why traditional classification models are not able to learn spontaneously.

Contents

List of Figures

Acknowledgements

I wish to express my sincere gratitude to Dr Simon O'Keefe, although the program was stopped on half way, he showed great patience , kindness and support till the very last moment, which enable me to develop a brief framework that can be used to solve part of my doubts as described in my PhD research proposal. Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

Declaration

I hereby declare that this dissertation is an original work and I am the sole author of this dissertation entitled “ The Computational Principles of Learning Ability". This thesis has not been presented for any other award at this or any other institution.

CHAPTER 1

Introduction

The field of “AI” as being widely mentioned today is generally held to have started at the conference in July 1956 when the term “Artificial Intelligence” was first being used, two main approaches were developed since then. Based on the belief that intelligence is fundamentally related to neuronal and synaptic activity [3][4], the “Bottom Up” approach looked at the neuron level and worked up to try to create higher level functions, in other words, they were studying the fundamental mechanism of intelligence. The “Top Down” approaches focus on higher level functions of intelligence and trying to implement those. Before 1980’s, many industry and academic fields researchers believed that it would not take too long to create artificial system which could simulate human intelligence. Therefore large amount of money was invested into this field and several programs and methodologies were developed, such as “Virtual Mall 1952”[5], “Geometry Theorem Prover 1958”[6], “General Problem Solver 1959”[7], “Eliza 1966”[8] an early example of primitive natural language processing program, and “Deep Blue 1997” the most famous expert system, however, this once great promise of bringing real artificial intelligence to public leads to nowhere but inevitable futility and researchers in this filed had avoided using this tarnished term “artificial intelligence (AI)" during the AI winter.

Actually, started from early 1970’s, some scientists[9] were beginning to realise that creating real intelligence might be much more complicated than their first thought. After 1980’s, bottom-up approaches were abandoned by most AI researchers in the field of computer science, therefore instead of emphasising on the ability to simulate human intelligence, software and algorithms developed by computer scientists now work as a kind of support of different applications, this methodology is known as machine learning techniques, more specifically, machine learning techniques are being developed as means for satisfying the demand of different disciplines other than an ultimate purpose: simulating intelligence on computer system.

The classic definition of machine learning can be summarised as:

Being capable to use experience to improve the behaviour of the computer system.[10]

This operational definition is different from previous definition of AI areas, so the question "Can machines think" be replaced with the question "Can machines do what we can do?". Because machine learning techniques are now being used to identify common problems of many different subjects that needs to be addressed urgently, so it has become a highly interdisciplinary area which combines studies of artificial intelligence, probability and statistics, neurobiology, cognitive science, information theory, cybernetics, computational complexity theory, philosophy and other disciplines. And it has demonstrated significant practical value in many fields such as data mining, speech recognition, image recognition, robot,

automatic vehicle driving, bio-informatics, information security, remote sensing information processing, computational finance and industrial process control.
Recently, due to the great success of Deep Learning Techniques, Artificial Intelligence (AI) becomes a hot topic again. Deep Learning researchers imply that the question about “How Brain Works" can be partially explained by their approaches. However, except for many philosophical discussions and the famous black box test “Turing test”, there is no clear definition of intelligence and it is well accepted that the ability of “thinking” is difficult to define[11]. On the other hand, for machine learning researchers and artificial intelligence experts, once a problem is solved, the solution as a computational model, seems to have nothing to do with intelligence, it seems like only a problem to be solved is related to the understanding of intelligences[12]. Therefore, researchers are trapped in a paradox as shown in figure 1.1.

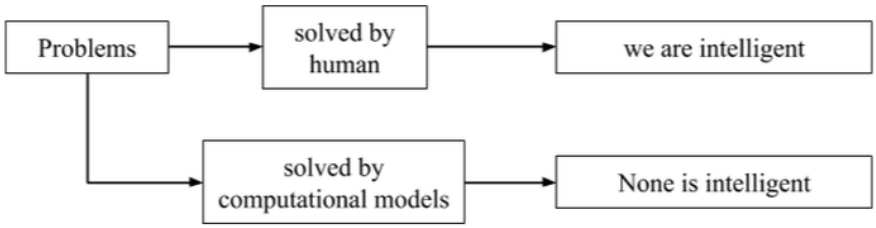


Figure 1.1: Paradox of AI effect. Problems solved by human, there is no doubt that we are intelligent. Problems solved by machine, there is no doubt that machine is not intelligent because we know exactly how it works.

Since all possible automated solutions implemented by computer systems are basically different computational models[13], it seems like there will be no computational model which could be acknowledged as possessing true intelligence forever. One possible solution of breaking this paradox is to find one or a set of criteria which can be used for white box testing of all computational models, as shown in figure 1.2.

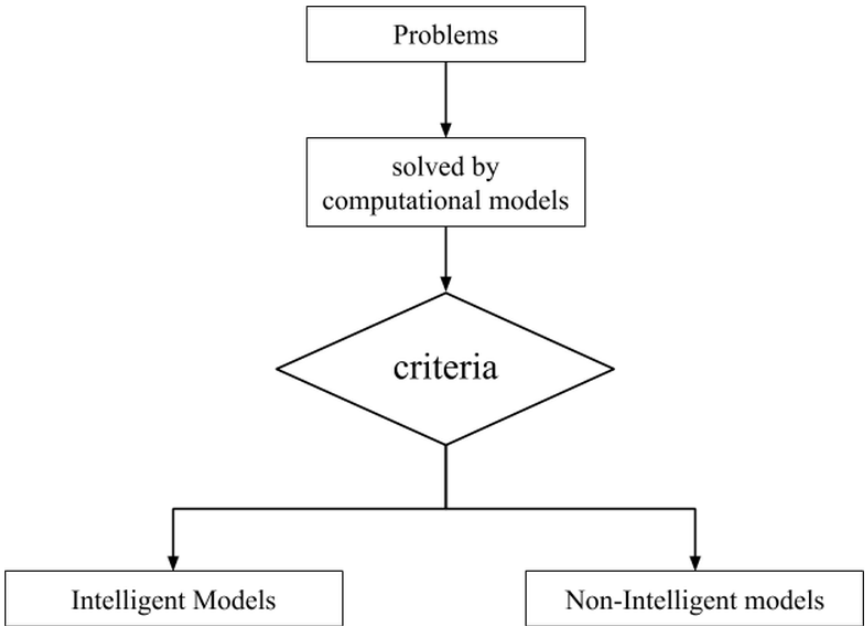


Figure 1.2: White box criteria for Intelligent Models

Instead of giving criteria for intelligences, based on the understanding that the learning ability is a critical component of intelligence, this dissertation proposes two laws for a computational model to be a learning model. With the help of these two laws, computational models can be classified as “Learning Model" and “Non-Learning Model" (figure 1.3), these two laws also provide a computational explanation about what the “Learning Ability" is.

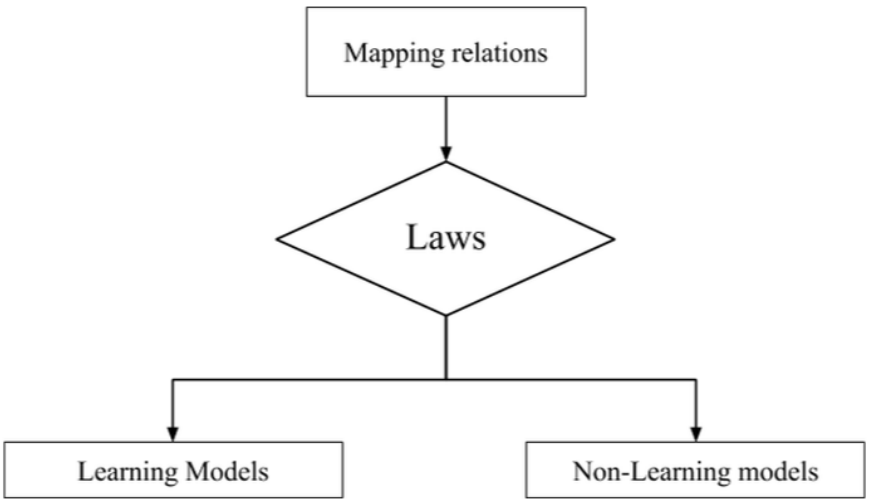


Figure 1.3: White box criterion for Learning model

1.1 Structure of this dissertation

This dissertation based on the belief that "There is no observation independent reality", and author’s explanation of this notion is " the existence of different concepts are the result of learning, it is not eligible to assume the existence of any concept before learning", or in other words " learning is the ability to define different concepts ". The same conclusion could be draw from an intuitive deduction of the "NO-FREE-LUNCH" theorem.

A general-purpose universal optimization strategy is theoretically impossible, and the only way one strategy can outperform another is if it is specialized to the objective function"[14][15].

In our daily life, there are many intuitive examples which reflect the notion of NFL theorem. Normally, we use coffee grinder to make coffee not juice extractor, we would like to wear soccer shoes while playing football not slippers, and paratroopers are equipped with parachute not umbrella. The fact that specific task requires specific solution is the most common knowledge which we take it for granted, and all these solutions are basically specialized optimization strategy for specific objective function. The most important fact is that in every using case of all specific strategies, there is an inevitable component: us. We identify the objective requirements, we decide or design what strategy will be used, and we cooperate with all kinds of strategies designed by us. We are drivers, we are pilots, we are captains and we are astronauts, the best way to guarantee the well functioning of a system is to integrate us with the task specific solution, because we are the inevitable core part and ultimate information source which support every specific strategy under exceptional circumstances, we are the general purpose universal optimization strategy. Therefore, does the existence of us disprove the NFL-theorem?

Assume that there is an universal optimization strategy **Z**, then the only reasonable explanation is that there is not any specific "objective function" exists before **Z** is being implemented, furthermore the existences of all "specialized objective function" are products of implementing **Z**, once learned the existences of different specialized object functions, it is possible for **Z** to draw the conclusion that the only way one strategy can outperform another is if it is specialized to the objective function. More specifically, we are this optimization strategy **Z**, we define the existence of all objective functions and based on that we could get the conclusion of NFL-theorem. This interpretation supports the "Single algorithm" assumption of neuroscience, actually readers will notice that three hypothesises regarding the primate neocortex[16] will be uniformly explained as coherent parts of the theoretical framework proposed in this dissertation. In the next chapter, an introduction of traditional learning theory will be given and by carefully analysing from the view point which has been missed by previous researchers, two issues of traditionally learning theory will arise, and these two issues directly lead to two corresponding laws which can be used to verify whether a model possess the learning ability as we intuitively familiar with. In chapter three, five definitions will be given, these definition not only necessary for the following discussion but also convey the author's interpretation view point which is very important for understanding this dissertation. In chapter four, two laws will be introduced in detailed, and one central conclusion is that different concepts are defined by learning model, and learning ability of a model is to be able to defined the differences of concepts.

CHAPTER 2

Literature Review

For any machine learning problem, one of the most important assumptions is that patterns exist in our observation. Expressing our observation as a pair $< X, Y >$, then a pattern is $g : X \to Y$, and X is the input space, usually a d dimensional subset of R^d, where Y the output space is a subset of real number. Therefore the domain $X \times Y$ contains all possible observation results and the pattern is a relation between X and Y. In a classification problem, Y is also known as the indicator set.

Machine learning is essential when people are interested in knowing the relation between X and Y, but cannot observe every possible element (x, y) of the domain $X \times Y$ nor know the precise mathematical expression of g. Therefore all machine learning approaches can only depend on two kinds of information that we do know.

- Our limited observation of the domain $X \times Y$: (x_1, y_1),(x_2, y_2),(x_3, y_3)...(x_m, y_m) denoted as $< X, Y >_{seen}$.

- A hypothesis set about the relation between X_{seen} and Y_{seen}, denoted as $H = f(x, y, \Theta)$, which is usually called a model.

Each candidate of the hypothesis set is distinguished by a unique k dimensional parameter $\theta \in \Theta$, and the ultimate goal of all machine learning algorithms is to pick one hypothesis θ_t and hopefully this $f(x, y; \theta_t)$ could approximate the target unknown pattern g better than any other candidate in the hypothesis set[1], or in other words we hope $f(x, y, \theta_t)$ could generalise the unseen part $< X, Y >_{unseen}$ [2] with highest accuracy. However, since the target pattern g is unknown and there will always be a future observation which has not been seen yet in every machine learning problem, how could people possibly know whether a hypothesis $f(x, y; \theta_1)$ possesses higher generalisation accuracy than any another hypothesis $f(x, y; \theta_2)$? For example, in a binary classification problem B, there are always at least two functions from either the same or different hypothesis sets which could behave exactly the same on the observation set $< X, Y >_{seen}$ yet behave differently on the unseen set $< X, Y >_{unseen}$. Actually, the no-free-lunch(NFL) theorem indicates that the average accuracy of implying machine learning techniques against all possible hypothesis sets on B is no better than a random guess [17].

[1] when the unknown pattern is a binary function, it is usually called a concept, and our corresponding hypothesis set is mentioned as concept class or class of concepts

[2] We assume that the set of unseen part is always non-empty.

2.1 PAC Learning

PAC learning indicates that the NFL theorem does not necessary mean the immediate doom of machine learning.

Assume our observation of an unknown pattern g is $< X, Y >_{seen} = (x_1, y_1), (x_2, y_2), ..., (x_m, y_m)$, and we have a finite hypothesis set $H = h_1, h_2, ..., h_n$. Then we could define $Error(h)$ as follows:

$$Error(h) = Prob\{g(x) \neq h(x)\} \tag{2.1}$$

Based on our observation we obtain $Error(h)_{seen}$, which is usually referred to as a training error, while the error of future observation $Error(h)_{unseen}$ which is usually referred to as Off-Training-Set Error (OTS error) remains unknown; PAC learning shows that the differences (ϵ) of these two errors is bounded as follows:

$$Prob[|Error(h)_{seen} - Error(h)_{unseen}| > \epsilon] \leq 2|H|e^{-2\epsilon^2 m} \tag{2.2}$$

Where $|H|$ is the size of the hypothesis set, h is the hypothesis which best fit our observation, m is the number of observations,ϵ is the upper bound of the tolerance, and once the difference between $Error(h)_{seen}$ and $Error(h)_{unseen}$ is bigger than this tolerance we could say that $Error(h)_{seen}$ has lost track of $Error(h)_{unseen}$.

Instead of trying to approximate g with some $h \in H$ directly, Inequation 2.2 shows that it is possible to approximate the error we cannot observe ($Error(h)_{unseen}$) using the error that can be observed ($Error(h)_{seen}$). The only thing left to do is to minimise the training error, so that we can guarantee that h will generalise the unknown target g with great accuracy. However there is a trade-off between minimising the training error and increasing the chance that our observation ($Error_{seen}$) would keep tracking what cannot be observed ($Error_{unseen}$).

Intuitively, as shown in Inequation (2.2) the bigger the hypothesis set we have, the larger chance that we could find an h with an even smaller training error. However a more sophisticated model would require more observations; otherwise, the error we can see $Error(h)_{seen}$ could easily lose track of the error that cannot be seen $Error(h)_{unseen}$. This situation is usually referred to as over-fitting.

2.2 Shattering Effect and VC Dimension

In previous section we assume that the size of our hypothesis is finite, and $|H| = \infty$ means there will be no guarantee that what can be observed ($Error_{seen}$) will keep track of what cannot be observed ($Error_{unseen}$) and learning in this case is infeasible. However, in the real world, the size of most frequently used real number parameter hypotheses is infinite. It would seem hopeless to implement machine learning techniques based on these infinite size hypothesises.

Instead of focusing on the exact size of a given hypothesis set, the idea of shattering is to try to analyse the effective behaviour of a given hypothesis set quantitatively. Assume $C(x, \Theta)$ is a concept class mapping from the input space X into {-1,+1}, so that C is said to be able to shatter $X\prime \subseteq X$ if every possible mapping relation $g\hat{}\prime : X\hat{}\prime \rightarrow \{-1, +1\}$ can be computed by a concept $x \rightarrow c(x; \theta \in \Theta)$ in the concept class C, and the VC dimension of the concept class C is the size of the largest possible $X\hat{}\prime$. More precisely, if the VC dimension of the concept class C is m, then the largest size of $g\hat{}\prime$ will be 2^m.

The idea behind the shattering effect is straightforward, even though a concept class C could contain infinite concepts defined by an infinite set Θ, what really matters is their behaviour. A million different concepts could behave as effectively the same as one concept, as shown in Figure 2.1. In Picture A, there are infinitely many hypotheses which are all able to separate dots from crosses, and in Picture B there is only one hypothesis which separates dots from crosses. Therefore, it is reasonable to assert that this infinite number of hypotheses behave as if there is only one hypothesis.

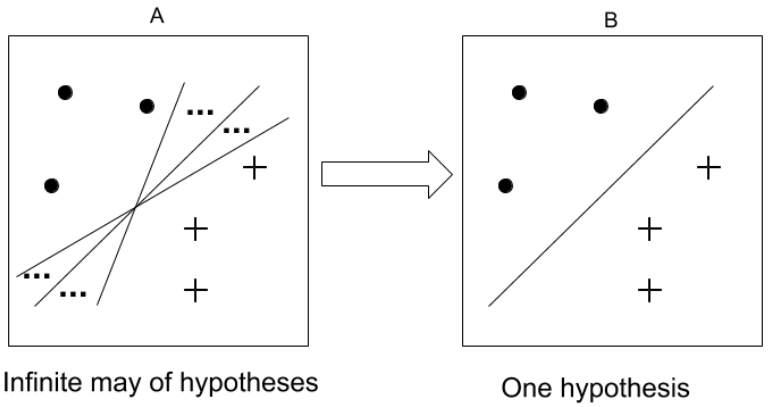


Figure 2.1: Infinitly many hypothesis versus one hypothesis

Therefore, although the size of a hypothesis set could be infinite, it could still behave like a finite size hypothesis set. By adopting the shattering effect, the effective behaviour of a hypothesis set could be analysed quantitatively and the mathematical definition of shattering is:

$$\forall g : X\prime \rightarrow \{0,1\} \exists \theta \forall x \in X\prime (H(\theta, x) = g(x)) \tag{2.3}$$

In equation 2.3 the number of possible is the VC Dimension of hypothesis set H. In the real world, this abstract quantity of many infinite hypothesis sets happens to be finite, and it is clear that the concept of VC dimension is independent from many other concepts of machine learning, such as 'learning algorithm' , 'unknown pattern', and 'sampling distribution'. It reveals the intrinsic property of a hypothesis set and therefore can be used to compare the capacity or representation power of different hypothesis uniformly, by replacing the size of a hypothesis set with VC dimension, we could rewrite equation (2.2) as follow:

$$Error(h)_{unsee} \leq Error(h)_{seen} + \sqrt{[H_{VC} \log (2m/H_{VC}) + H_{VC} - log(\delta/4)]/m} \tag{2.4}$$

As introduced by [18, p.76], in this new equation, H_{VC} is the VC dimension of hypothesis set H, m is the number of instance being observed, and equation (2.4) holds with the probability of $1 - \delta$, or in other words, δ is the chance that our hypothesis will behave badly ($Error_{seen}$ lose track of $Error_{unseen}$). According to Equation (2.4) we can see that:

1. Low VC dimension compared to the size of dataset would suggest that $Error_{seen}$ approximates $Error_{unseen}$ with higher accuracy, and larger VC dimension is good for minimising the the error that can be observed ($Error_{seen}$) but bad for generalisation.

2. Infinite VC dimension still means that the learning is infeasible.

3. For keep tracking what cannot be observed, the number of necessary instances of observation are proportionally increase with the value of the VC dimension. Only when exposed to this large number of instances, a hypothesis set is said to be able to be generalised the unknown target g with the probability of $(1 - \delta)$.

2.3 Traditional Classification Task and Two Issues

In this section, one simple binary pattern classification problem is investigated. The discussion of this example will not only show how VC dimension can be used to analyse the generalisation ability in practice but also reveal two important issues about traditional learning theory of classification tasks, and these two issues will be used to illustrated what is the learning ability. Furthermore, two laws of learning ability which will be introduced in Chapter 4 correspond to these two issues.
Suppose there is a binary multilayer feed-forward neural network N_1 which will compute an unknown target function:

$$g_1 : R^n \rightarrow \{0, 1\} \tag{2.5}$$

As shown in Figure 2.3, the neural at layer t is the label of g_1 which has a binary value $\{0, 1\}$.

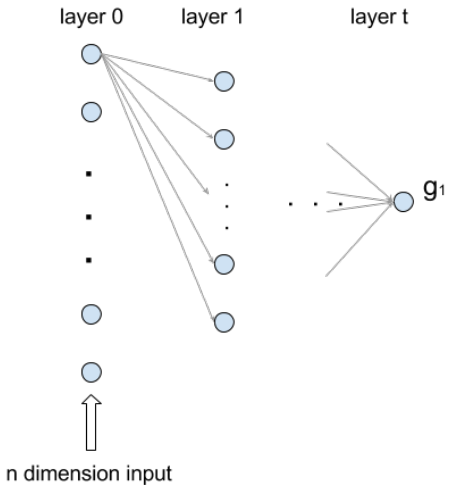


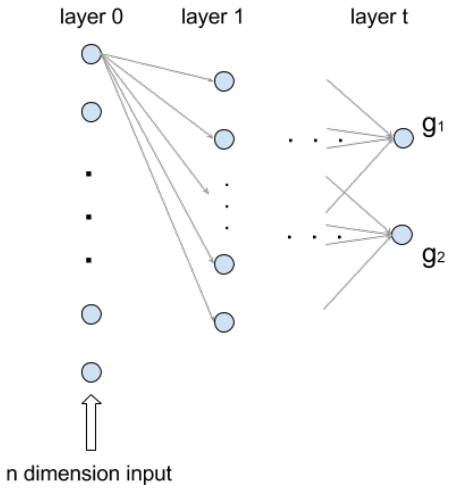
Figure 2.2: Multilayer feed-forward neural network N_1

The activation function of each neuron is a linear threshold activation function:

$$F = \{f_W(x) = \rho[w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n] \mid W \in R^{n+1}\}, \tag{2.6}$$

The VC dimension of our neural network is $VC_{dim}(N_1) = O(nl \log n + nl^2)$ [19]and because a linear threshold function is a piecewise polynomial function, the VC dimension of N_1 is also proportional to $w\dot{\log} w$ [20], so the VC dimension of N_1 is also:$\Omega(w \log w)$ [21], where w is the number of weights of a neural network.

A binary classification neural network could be easily extended to handle multiclass classification tasks as shown in Figure 2.3. Recently by adopting layer-wise pre-training techniques, it is possible for a deep neural network to generate hierarchical representations of high-dimensional data, and meanwhile preserve its neighbourhood structure in nonlinear mapping [22]. Therefore, theoretically, in neural network N_d the neuron g_1 could represent the concept of the front image of human[22], the neuron g_2 could represent the concept of a cat[1], and with enough observations of human faces or images of cats, our neural network N_d could actually learn an unknown concept $g_1 : R^n(images) \rightarrow \{0(not - human), 1(human)\}$, $g_2 : R^n(images) \rightarrow \{0(not - cat), 1(cat)\}$.



4

Figure 2.3: Neural network N_d for multiclass classification task

2.3.1 Two issues

In this section two new terms will briefly be introduced in this dissertation. In the example above, the neural network N_d must be exposed to enough instance, so that it could generalise the unseen data with high accuracy. Therefore, we can say that all those instances collectively represent the concept of humanity, and the output +1 is another representation of humans. The term "concept" in this case means the existence of humanity which is a distinguishable notion from other species.

At this stage, we could call the set of all instances the local representation of humanity, and +1 in conjunction with the map the global representation of humanity. Presumably, the unknown pattern that exists in our observations could always map the local representation (an infinite set) to a single global representation.

There is another important assumption of learning theory: the elements x of $< X, Y >_{seen}$ and $< X, Y >_{unseen}$ are drawn from the same distribution independently. It is apparent, after training, when being exposed to pictures of pig, dog,cup,etc, that it is impossible to determine the input picture by observing the output of N_d. In other words, with all these local representations which we human beings know are from different categories, our model N_d can only offer an indistinguishable global representation. The differences being contained in the input are lost [3] in the global representation and all these input are noise with respect to the input that we expect our model to recognise. Currently, classifiers in traditional learning theory are able to determine what is not expected to be recognised; however all differences among unexpected inputs are lost This is the counter event issue of traditional learning theory.

If we want the model to be able to recognise dogs or cups or pigs. we would include pictures of these different categories in our training set, and replace our old hypothesis set N_d with a new one$_d^+$ and go over all the training processes again. After the training, we acknowledged that a new model has "learned" how to recognise these different objects, but is the learning process the same as our ability of learning?

The author assumes the following facts are well accepted by most of people:

- Human beings and some other animals (like dogs) are able to recognise new objects which are different from all objects we have ever seen.

- We do not need to see a large amount of instances of a new object that we not seen before.

When dealing with classification problems, whenever we need to include a new class in our model, it is inevitable to construct a new hypothesis set (model) and go over the entire training process all over again. Different assumptions about the number of categories t expected to be recognised will lead to different hypothesis sets H_t (t is the number of categories expected to be recognised). In addition, one necessary condition of reaching the best generalisation performance of a given hypothesis set H_{t1} is to implement H_{t1} to a classification problem which is expected to contain $t1$ different categories.

If a model is assumed able to identify different categories in a domain with great accuracy, it must be built based on the knowledge about how many categories there are in the training set at the first place; however if knowledge about differences in the domain can only be gained by implementing machine learning techniques, then how it would be possible to know how many different categories there are beforehand? For example, in Figure 2.3.1, the training set contains images of different objects, outputs are labels that represent different categories, and the necessary information for selecting a model H_t is the number of labels. In reality, this information is usually being provided by us, the model creator. Therefore, in summary, machine learning techniques cannot solve randomly selected classification problems spontaneously.

[3] The central philosophy of this dissertation is that the differences exist or can be identified only because the global representations are different; otherwise there is no way for us to know that a picture of pigs is different from a picture of dogs.

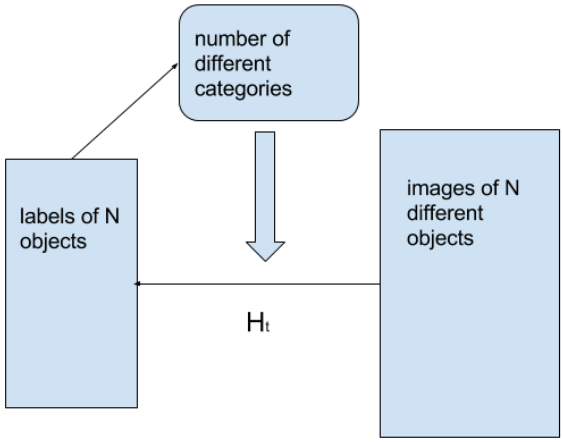


Figure 2.4: Spontaneity issue: the information loop

In the NFL theorem, this conclusion was being summarised as " A general-purpose universal optimisation strategy is theoretically impossible, and the only way one strategy can outperform another is if it is specialised to the objective function"[14][15]. Therefore, with respect to human-level ability of learning which is usually acknowledged as a general purpose learning algorithm, the only possible conclusion is that there is no such 'objective function'. The existence of a pattern is the result of learning. Because of the ability of learning, learning models (human beings) should be able to sense the existence of a pattern. It is the learning ability that brings us the differences of a pattern, not the other way around.
These counter event issue and spontaneity issues are fundamental differences between traditional learning theory and our intuitive understanding of learning ability. These two issues will lead to Law one and Law two directly which will be introduced in fully details in chapter 4, before that some formal definitions need to be given.

CHAPTER 3

Definitions

3.1 Intuitive Examples

When we look outside the window, we could see birds, butterflies, clouds, and trees. When we take a deep breath, we could taste the sweet smell of the freshly cut grass. Although we have no direct access to the world other than through our sensors [23], we can always rely on different kinds of apparatus to discover the world. In fact, all apparatus can be regarded as extension of our biological senses. However what if something cannot be detected by all means? Is it necessary to insist on its existence? The question has been answered perfectly by Carl Sagan's famous story *"The Dragon In My Garage"*[24]. Discussions in this dissertation are built based on the belief that:

There is no observation independent reality.

This notion will naturally lead to a hierarchical relation between two representations of a single "existence" (The meaning of this quotation mark will be revealed soon).

3.2 Global and Local

Imagine there is a world that contains three dice and a sub-world in it. There is only one very narrow channel between World One and its sub-world, and the inner state of this sub-world can be affected by the surfaces of these dice through this narrow channel. All these dice have surfaces with six different colours: R(red), G(green), B(black), P(purple), Y(yellow), and O(orange), the creator of the sub-world and the dice guarantee that these dice exert their power through the channel by the following rules:

- These three dice will exert their power through the channel one by one.
- The rotation order of die one (D1) and die two (D2) are the same: red->green->black->purple->yellow->orange->red->...->orange and repeat.
- Each surface of these dice will be face the channel for a certain amount of time; surfaces with the same colour as dice one (D1) and dice three (D3) will exert their power with the same proportion of time.

- People of the sub-world have no ability to affect anything outside the sub-world, in other words their behaviour cannot affect states outside the sub-world directly or indirectly[1].

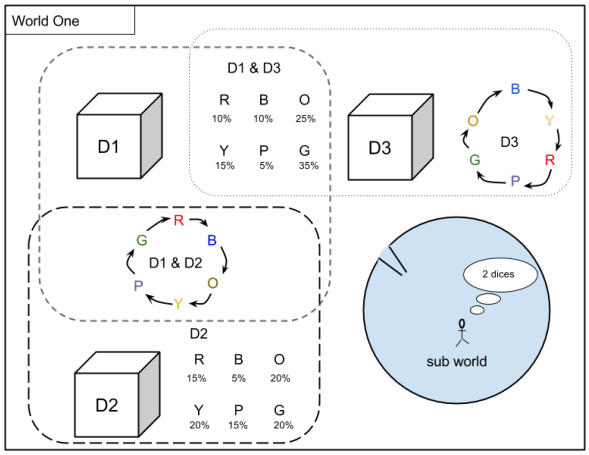


Figure 3.1: The relativity between global and local.

- Scenario A
  If the sub-world will be affected only by how long each colour is being shown to it, then people living in this sub-world could have the notion that there are two dice in World One.

- Scenario B
  If the sub-world will be affected only by the order of the colour being shown to it, then people living in this sub-world could also have the notion that there are two dice in World One.

In Scenario A, the differences between D1 and D2 vanished; in Scenario B, the differences between D1 and D3 vanished, and we know that. However, people who live in the sub-world can only be affected by local effects of the dice on the sub-world, and based on those local effects , the people of the sub-world learn of the existence of the two dice. Therefore, we can say that local effects in this sub-world collectively represent the existence of two dice, and so does any distinguishable labels being used by people from the sub-world to record the existence of these two dice[2]. The existence of these two dices is indisputable truth for people of the sub-world, as long as what can be detected by people in the sub-world still follows either Scenario A or B as introduced above. However, this “truth" can only be verified by people who live in World One or any world that contains World One, which has the relative global vision with respect to people in the sub-world[3], even though this verification is meaningless for people in the sub-world. Therefore, with respect to observers in the sub-world, all those local effects collectively represent the existence of these dice, and the concepts of the existences of two dice (no matter how it is recorded) are global[4] information which could be “wrong" but will affect only global events.

[1]Actually, our skull is such a sub-world which keep our brain inside, the only difference is we could affect states of the world which contains our skull.

[2]Could also be a certain combination of activated neurons in their brains.

[3]Physicians of the sub-world might want to find a unified field theory which would eventually describe all local effects in term of one die.

[4]The term “global" is used in contrast to “local". This usage also reflects a nested relation between World One and the sub-world. What the sub-world learned from local information can only be verified by the world it is nested in. The result of this verification could has no effect on it. Most of all, the author’s viewpoint of this relation can be revealed by using these two terms.

Actually, by inspecting our daily life with a little bit more patience, we can find that global information and local information appear in pairs very often and almost everywhere. People just take it for granted most of the time. Here are two examples below:

- Global information: File extensions. Local information: a set of certain schema.
- Global information: DNA. Local information: corresponding environment.

When we write some simple applications, if an interface was designed to process XML files, this means the implementation of this interface will expect an input file with XML schema. If some file other than expected was passed to this interface, usually we obtain an some error message and simply modifying the suffix would not offer any help. There are infinitely many files of different types that could be passed into this interface, and all their different local information will be lost in a single error message. If a mammal from the Serengeti national park was somehow brought to Antarctica, the error message would be "DEATH", all those local information are lost in this single error message as death in here [5].
Especially in software engineering, the more global information being used in coding (known as hard coding or tight coupling), the less generalisation ability the application will have. As introduced in the file extension example, one typical solution of making our application behave more smartly is to include a vast number of exception handlers; however, ironically, this method ultimately needs humans to decide what exceptions this application can expect to meet[6]. Now, readers might have realised that when a student starts learning a new programming language, there is plenty of global information such as the key words “int", “double", “String", and a student needs to learn their corresponding local information and the different contexts in which these variables should be used. When a variable is passed to an unexpected place, we will get the “TypeError" message which is usually the most common error message; all information contained in that variable cannot be obtained from that “TypeError" message.

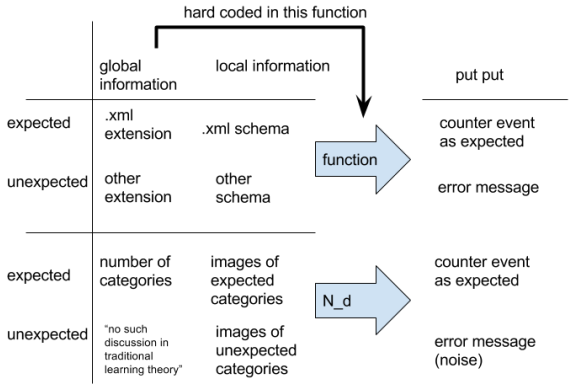


Figure 3.2: Error message of hard-coded function and traditional learning theory.

[5]One may argue that the category of an animal can be identified by inspecting its remains, and therefore all local information can be recovered completely, such as appearances, habits, and colour of the skin. Try to imagine if this were the remains of an animal from other planet, could all this local information still be recovered?

[6]Even though, we could come up with a sophisticated model which could somehow detect whatever schema the input file could use, how can this model be integrated into an application? Currently, there is only one such sophisticated model: us, and we are also responsible for creating and adjusting all other applications.

It is also true that when we extend this concept to natural language, a combination of some marks known as words are global information with respect to all different kinds of local representations, which are usually referred as “MEANING". Dictionaries are full of this pairwise information and our daily lives also relies on this pairwise relation. The word “HOME" would generate feelings of easy and comfortable, while the word “GRANNY" might mean delicious cookies or strong hugs. It is also true when you heard these words, global information is just a different combination of sounds, however, they could carry the same meaning (local information) as corresponding words. When people are learning a foreign language, one easy task is to remember foreign words of substantial objects because it is just a set of local information and extra global information.

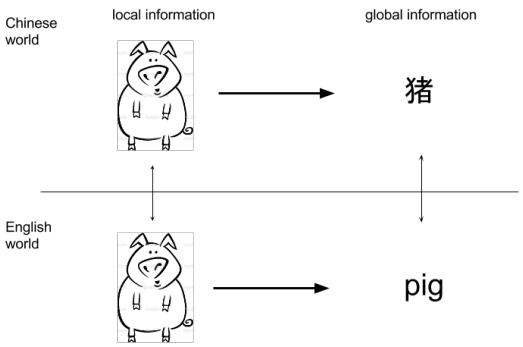


Figure 3.3: As a substantial existence, it is easy to directly map its English global information to Chinese global information.

What could be difficult is trying to remember the word of an abstract concept because the corresponding set of local information may have never been collectively used to represent some concept and therefore has no single global representation, such as the German word “WALDEINSAMKEIT". There is no single corresponding word in either Chinese or English that has the same meaning. Worse, the corresponding set of local information might not even exist in a student’s life, what shown in Figure 3.4 is almost impossible to translate into English. Learning a foreign language would literatelly brings people a new vision of our world.

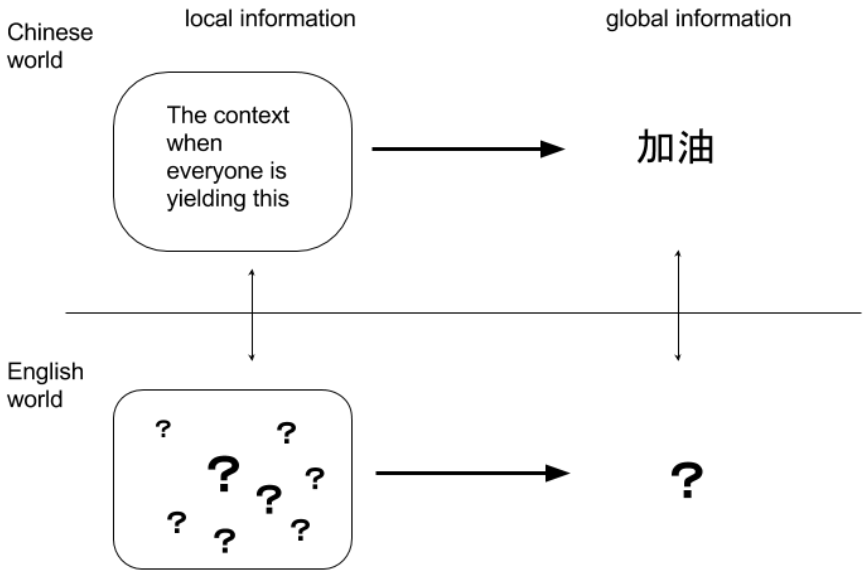


Figure 3.4: As an abstract existence, it is almost impossible to directly map this Chinese global information to English global information, even though there are similar activities in the English world; however all the local information just does not match perfectly.

These intuitive examples not only briefly illustrate the relative global and local relation but also indicate some intrinsic casual relations between this local and global relation. In the following section, based on the World One and Sub-world example, five definitions will be given. These definitions not only aim to express this global and local relation mathematically, but also aim to provide utilities to analyse whether a model is able to learn, or in other words possesses the ability to always give counter events spontaneously.

3.3 Definition 1

For a given mapping relation $M : D \to O$ ***(*** D ***and*** O ***are subsets of high-dimensional Euclidean space), an element in the range (*** $o \in O$***) is a piece of global information and each element in its corresponding domain (*** $\{e \mid e \in D, M(e) = o\}$***) is local information.***

At this stage, this definition means nothing more than giving names to elements of the range and domain in the image of a mapping relation. Any apparatus used to detect the world can be regarded as a mapping relation from appearances (subset of apparatus' domain) to the concept (elements in the range). Biological systems can also be regarded as one kind of apparatus ,which includes us. A concept could be a substantial object, such as cat, or an abstract concept, such as gravity or electromagnetic waves.

3.4 Definition 2

For a given mapping relation $M : D \to O$ [7]***, one subset of the domain and each element of the***

[7] In the following sections, all domain and range are subsets of high dimensional Euclidean space

corresponding subset of the range are local and global representations, which define the same concept ,respectively.

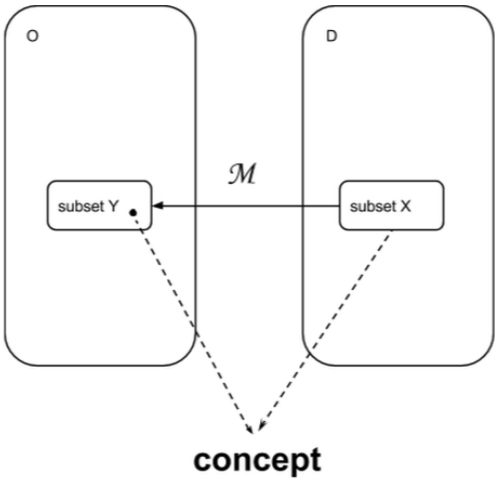


Figure 3.5: Local and global representation. A set of local information is the local representation of a concept; each global information is one global representation of a concept.

A set of appearances are detected by an apparatus M (the domain of a mapping relation M). These appearances indicate the existence of a concept, so the local representation of this concept is defined as the subset X ,and the global representation of this concept is defined as each element of the subset Y. Therefore, all local information about a concept is its only local representation, and the global information of the concept is equivalent to its global representation. Furthermore, it is necessary to define local and global representation in this way because of the hierarchical structure of this global and local expression. More specifically, it means the existence of a concept depends on its appearances, and this concept itself could also be one of many appearances that define a higher level concept. This can be explained by a simple mental experiment.

Imagine sleeping on the backseat of a minivan and the shaking caused by the speed bump wakes you up. You do not know how long you have slept, and you watch the scenery passing outside the window. Suddenly, you realise that you are approaching “Some Place". What is included in the passing landscape depends on appearance. It could be a cottage, church or supermarket. All these rapidly passing views are appearances which enable you to recognise that you are near this “some place”, and this “some place" is a higher level concept. Therefore, Definition *2* and Definition *1* should be able to precisely describe all parts of this hierarchical relationship of different information. Readers might argue that it is redundant to define global and local relations as Definition *2* and Definition *1*. As shown in Figure 3.6, Set A collectively represents a concept **A**, so according to Definition *2*, Set A is the local representation of Concept **A** and a is one global representation of Concept **A** (Other possible global representations of concept **A** are not shown here or there could be only one global representation). And concepts a, b, and c belong to another Set E which could collectively represent a Concept **E**, therefore from the viewpoint of the mapping relation f_4, a is a piece of local information and $e = f_4(a)$ is a piece of global information (Definition *1*) which is also a global representation of Concept **E**. More intuitively, A could be the appearances of a cottage, B could be the shake caused by the speed bump, a and b are their global representations, respectively, and they collectively as E they represent some place e.

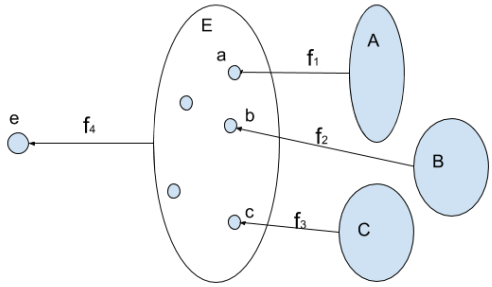


Figure 3.6: The hierarchical structure of local and global. Capital letters mean sets which are different local representations and its elements are local information.

In summary, Definition *1* and Definition *2* indicate that:

1. These four expressions are equivalent:
   A concept≡A set of appearances≡A set of local information (Definition *1*)≡A local representation (Definition *2*)

2. These three expressions are equivalent: A concept≡A piece of global information≡A global representation

3.5 Definition 3

For a given mapping relation $M : D \to O$***, it defines the type of global information as*** M**.**
As the notion being introduced in Definition *1* is:

There is no observation independent reality

The existence of a certain concepts depends on whether it is observable. Furthermore, the nature of the concept depends on the method of observation. As shown in Figure 3.6 set E collectively represents a concept, and the elements of E could be the output of different types of mapping relations f_1, f_2, and f_3. Intuitively, the phrase "Pepperoni Pizza" (a global representation) means slightly spicy, red colour, Ron Cooke hub[8], and all this local information that collectively represents "pepperoni pizza" contains different types of information: vision, taste, and even memory.
Definition *3* guarantees that it is the appearances and the way these appearances are processed that decides not only the existence but also the nature of the concept because the appearances (local information) form a subset of the domain, and the way this local information is processed (mapping relation) gives the global representation (global information) of a concept.
Even when facing the same domain, different mapping relations will give different types of information. One typical example is the camera where the domain is provided by the CCD array. Different functions of

[8]Not necessarily true for others.

the camera will give different types of information, such as the focusing information used to adjust the lens and the information recorded as photos.

3.6 Definition 4

For two given mapping relations: $M_l : S_1 \to S_2$ ***and*** $M_h : S_3 \to S_4$***, if*** $S_1 \cup S_2 \subset S_3$ ***then*** $\forall s \in S_2$ ***is homologous global information with respect to*** M_h***.***

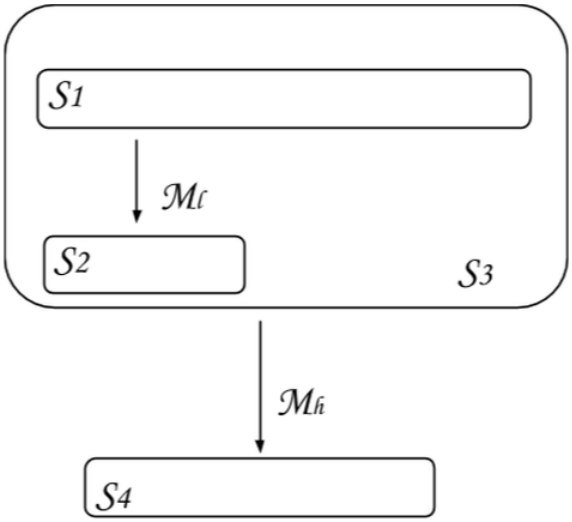


Figure 3.7: Homologous global Information. Elements of $S2$ are homologous global information with respect to the mapping relation M_h.

3.7 Definition 5

i ***For two given mapping relations:*** $M_l : S_1 \to S_2$ ***and*** $M_h : S_3 \to S_4$***, if*** $S_2 \subset S_3$ ***then*** $\forall s \in S_4$ ***is first order global information with respect to*** M_l***.***

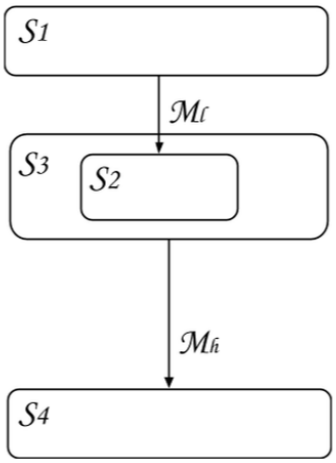


Figure 3.8: First order global Information. Elements of $S4$ are first order global information with respect to M_l.

Apparently, all elements in S_4 are beyond the field of vision of mapping relation M_l. If M_h has a one-to-one mapping relation between $S_2 = S_3$ and S_4, it is also true that all elements in S_2 are beyond the field of vision of mapping relation M_l. Therefore, for a given mapping relation $M : D \to O$, all elements in its range are first order global information with respect to M itself. In addition, together with Definition *2* we know that these three concepts are equivalent: global information, a global representation, and first order global information.

3.8 Motivaton of Definition *4* and *5*

The motivation of these two definitions above can be explained from the viewpoint of trying to write a function with two arguments (a, b) and a return value c. Suppose this function is defined as follows:

int c exampleFunction(int a, StringList b)

This example Function takes an integer and a list of strings as inputs and returns an integer value. It is perfectly normal that the value of a will affect the behaviour of example Function, and a could also be a return value from another function which take the same list of strings as input. This is an intuitive example of Definition *4*.

According to Definition *5* the return value of c is first order global information with respect to example Function, there is no way that this value or any other function that takes this value as input could affect example function's behaviour that produced c, because it violates causality [9]. However, in traditional learning theory as being shown in Figure 2.3.1, the need for this information loop causes the spontaneous issue. If we would like to construct a complex mapping relation by integrating many simple components, then definitions *4* and *5* illustrate all possible relations among these components, they provide information which would be either homology or ordered. Furthermore, as shown in this example, the return value c means nothing to example Function, which means c must be an input of another function; otherwise, all the hard work of example Function will be in vain. The function that takes c as an input value will be referred to as the harvest function of example Function. Up to this stage, readers should have an intuitive understanding of local and global relations, and these two issues of traditional learning theory also appear in several intuitive examples in this chapter. Before we start the formal analysis of these two issues, the author would like to summarise the basic assumptions of this dissertation briefly:

- There is no observation independent reality
- There is no "unknown pattern", the existence of different concepts is the result of learning, not the other way around.
- As we are intuitively familiar, a model with learning ability could always define different concepts according to different appearances, spontaneously.

[9] In recursive functions, a return value cannot affect the behaviour of the function that produced this return value; it is a different scenario.

CHAPTER 4

The Laws of Learning

At the end of Chapter 2, two issues are raised from analysing the traditional learning theory In this section, to simplify the expression, the term 'model' will be used to replace the term 'hypothesis set'.
A model $M_L(D;\Theta) : D \rightarrow O$ is said to be a learning model that is able to learn from its domain D (D and O are high dimensional Euclidean space); it must follow these two laws.

4.1 Law 1

$$\forall X_S \subset D, \exists \theta, \exists X_N \subset D \setminus X_S : Y_S = M_L(X_S;\theta), Y_N = M_L(X_N;\theta), Y_N \cap Y_S = \emptyset$$

For any subset X_S of the domain D, there exists θ so that we can always find at least one such pair $< X_N, Y_N = M_L(X_N;\theta) >$, X_N is disjoint with X_S and Y_N is also disjoint with $Y_S = M_L(X_S;\theta)$.

This law would be best understood by assuming that there is a model M_{non} that does not obey law *1*. Then there will be four[1] possible scenarios as follows:

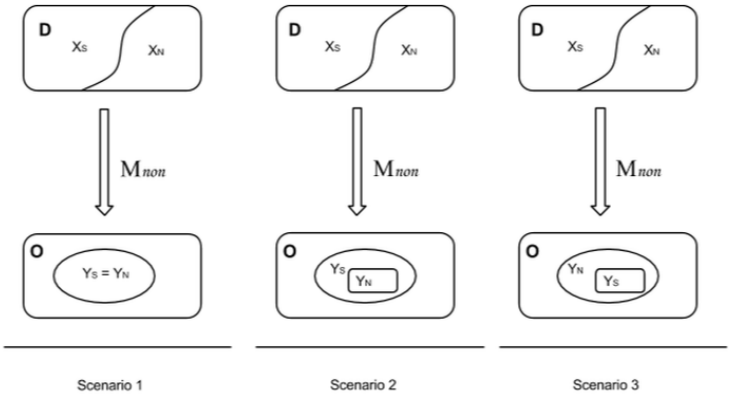


Figure 4.1: Interpretation of Law One: Scenarios 1-3

In the three scenarios in Figure 4.1, X_S and X_N contain different local representations; however the model M_{non} will map them to the same set of global representation. Therefore, as a detector, M_{non} fails to detect different concepts in the domain. In other words, models that do not obey Law *1* are basically

[1] Scenarios two and three are equivalent.

information black holes, which could allow possible information of the existence of many concepts to devolve into the same state [25, p.43]. This expression might seem vague at first glance; however recall the intuitive examples at the beginning of Chapter 3, reading the files with unexpected extensions that all lead to the same error message. There is no way to tell all their differences from the error message; it is the same for the counter event issue of the traditional learning theory, as shown in the example at the end of Chapter 2. The Hegelian dialectic events happened only on expected classes; therefore we could say that a typical classifier is remembering what being expect to remember in the training set in a fancy way.

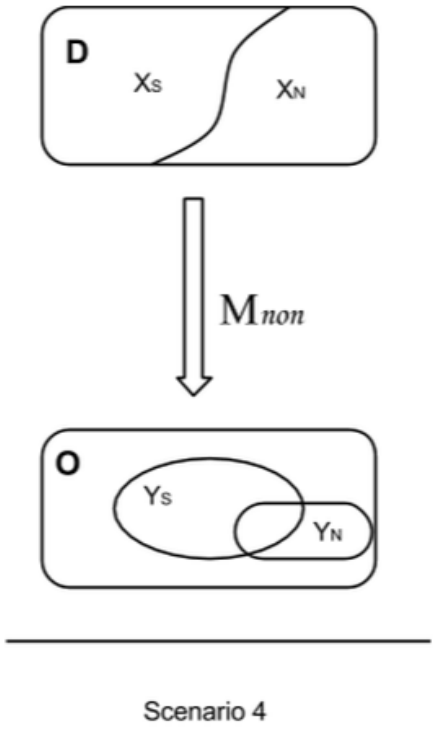


Figure 4.2: Interpretation of Law One: Scenarios 4

What if Y_S and Y_N are partially overlap as shown in Figure 4.2? We could assume that Y_S is subject to constraint C_S and Y_N is subject to constraint C_N, so that the differences contained in local representation X_S and X_N can be preserved. However, adopting unique constraint indicates that we need to apply a certain constraint on a unique event before the uniqueness of this event is actually being defined. In other words, the model need information I_U about the uniqueness of input set to make the decision about when it should apply what constraint to the output set, which means in this scenario the model cannot define the uniqueness of an input set spontaneously.
In summary, Law *1* guarantees that the possible existence of a new concept will not be missed.

4.2 Law 2

The training process*[2] *of $M_L(\alpha \subset \Lambda) : D \rightarrow O$ ***should not depend on any of*** M_L***'s first order global information*** M_H ***which follows Law 1 and Law 2 as well.*** The intuitive explanation of law *2* is that a learning model should be able to define the existence of a new concept all by itself. Here is an example:

[2]The notation Λ represents the parameter space of a model M_L.

The second issue is about learning spontaneously. The number of categories being contained in a domain is first order global information, and the exampleFunction example at the end of Chapter 3 also indicates that include any first order global information will violate causality.
Definition *2* ends all similar arguments, such as 'Should a NLP system linked with a dictionary be labelled as 'Intelligence'?'. Furthermore, together with Law *1*, models can be classified into two categories, learning model and non-learning model (memory system), so that further analysis could be possible. Readers have noticed that in the example at the beginning of Chapter 3, hard coded and tight coupling always limit the generalisation ability of an application, in other words, to make an application behaves smartly, we need to reduce the use of hard coded information and decouple, so the same for reducing the use of global information for good generalization ability.

4.3 Corollary 1

Given a model $M_L : D \to O(O \subset R^n, D \subset R^m)$***, if*** M_L ***follows Law 1*****, then there exists a family of functions** $H : R^m \to R$ **which can be used to harvest information contained in the range of** M_L**.**
As shown in Definitions *4* and *5*, elements in the range as first order global information are the products of a model which is also beyond the vision of a model itself. Therefore, all that global information can only be identified as local information which will be processed by other higher level mapping relations, it is just like the return value of a function, if not assigned to some variable, this return value will be lost. Therefore, the author chose to use the term 'harvest' to express this operation. Rice would not appear in you mouth automatically whenever you are hungry, it must be harvested from the field first.

4.3.1 Lemma 1

For a given $D \subset R^n$***,*** $\forall X_j \subset D$ ***and*** $\forall r \in R$***,*** $\exists \Phi(x; x \in X_j) = r$**.**
For any subset X_j of a n-dimensional real number in domain D and any given real number r, there exists an equality constraint $\Phi(x)$, so that $\Phi(x; x \in X_j) = r$.

Proof:

For a given subset $X_j \subset D$, there exists an equation[3]:

$$G(t) = (t - x_1) \bullet (t - x_2) \bullet ... \bullet (t - x_i) \tag{4.1}$$

so that for any $x \in X_j$, we have $G(x) = 0$.
And for a given real number r, there exists:

$$\Phi(t) = G(t) + r \tag{4.2}$$

so that for any $x \in X_j$, we have $\Phi(x) = r$.

Lemma *1* shows that for a given subset X_j of a n-dimensional real number space[4], there exists a family of equality constraints: $\Phi_{Xj}(t, r; X_j)$, so that $\forall x \in X_j$ and $\forall r \in R$ we have $\Phi^r_{Xj}(x; r, X_j) = r$.

[3] Solutions of this equation could be not an element of X_j; this equation '$G(t)$' is just an example which proves the existence of an equality constraint Φ.

[4] Currently, we assume that this is a discrete finite subspace of the continuous real number space.

The constraint family $\Phi_{Xj}(t, r; X_j)$ can also be expressed as a column that includes infinitely many constraints, as shown in Figure 4.3.

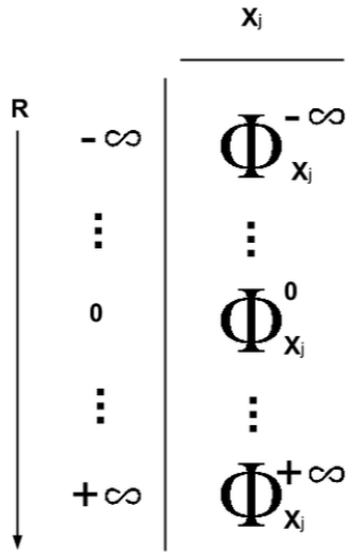


Figure 4.3: A family of constraints defined by X_j

Furthermore, constraint family $\Phi(t, r, X_k)$ is defined by the power set of the domain D ($\forall X_k \in \mathcal{P}(D)$), and each member of this family is $\Phi^r_{X_k}(x) = r$ and the constraint family $\Phi(t, r, X_k)$ can also be expressed wit a matrix as shown in Figure 4.4.

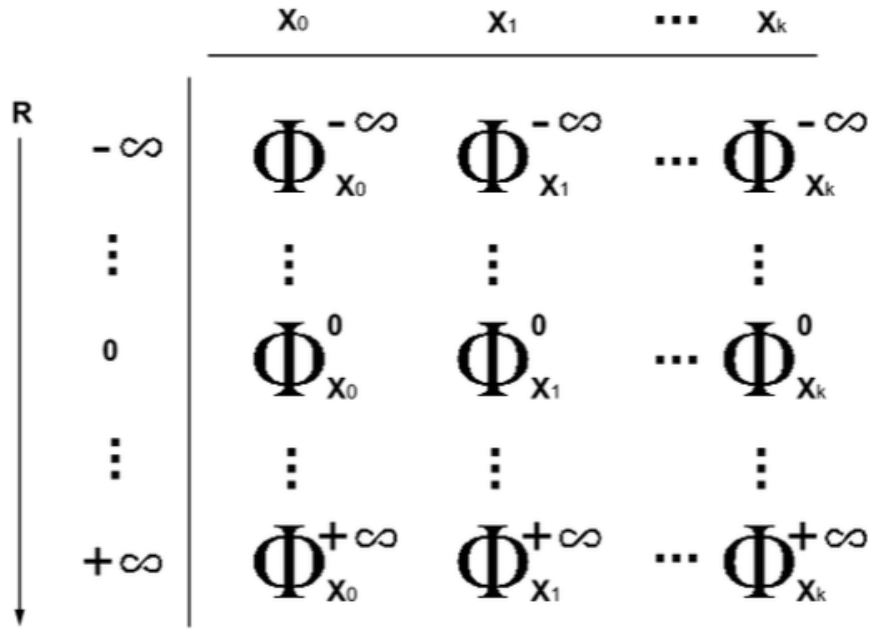


Figure 4.4: Constraint families defined by powerset of D

Proof of Corollary 1

It is known that model $M_L : D \rightarrow O(O \subset R^n, D \subset R^m)$ satisfies Law *1*, so that for possible learning result Y_1 there exists an equality constraint $\Phi_{Y_1}^{r1}$, and for a new learning result Y_2, there exists infinetly many equality constraints $\Phi_{Y_2}^{r2 \neq r1}$, so for a new learning result Y_i, there always exists an equality constraint $\Phi_{Y_i}^{ri}$ which is defined by Y_i and a real number r_i ($r_i \neq r_1, r_i \neq r_2$). Then the constraint in which all possible learning results can be expressed is shown in Figure 4.5.

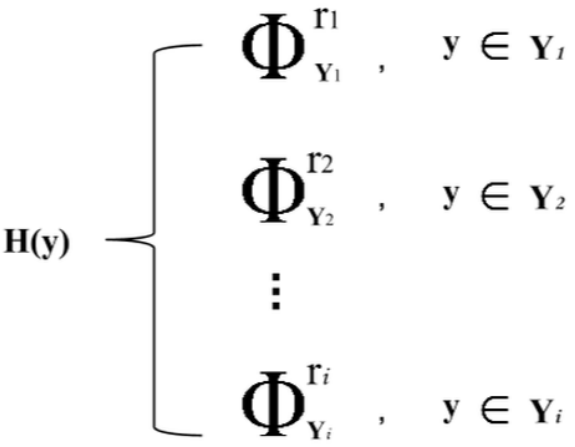


Figure 4.5: Information harvesting function H

Because Y_1,Y_2, to Y_i are disjoint sets(Law *1*) and the corresponding formula $\Phi_{Y_I}^{ri}(y)$ equals a unique real number, $H(y)$ is by definition a function. There could be infinite possible $H(y)$. Thus, H is the harvesting function that can harvest information from a mapping relation, which follows lLaw *1*.

4.4 Corollary 2

The composition of mapping relations M_L that follows Law *1* and its corresponding harvesting function H still follows Law *1*.

Proof:

Denote $F_l = H \circ M_L$.
Because M_L follows Law *1*then:
$\forall X_S \subset D, \exists X_N \subset D \setminus X_S$
so that $Y_S = M_L(X_S), Y_N = M_L(X_N)$ and $Y_N \cap Y_S = \phi$
Then according to the definition of (H):
$H(y; y \in Y_S) \neq H(y; y \in Y_N)$.
Therefore, we know $\forall X_S \subset D, \exists X_N \subset D \setminus X_S$:
$F_L(X_S) \neq F_L(X_N)$

Corollary *2* indicates that it is possible to represent different concepts of information type M_L using different real numbers.

4.5 Corollary 3

The composition of mapping relations M_L and any of its harvesting function H is equivalent to a common constraint V_C and a model 'L' which also satisfies Law 1.

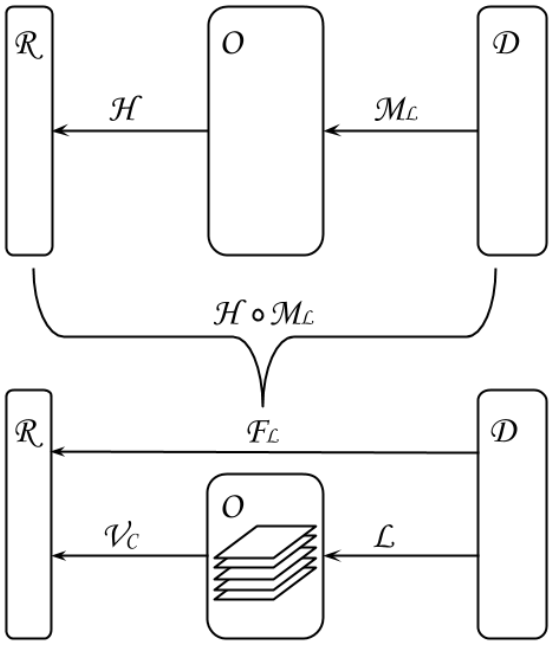


Figure 4.6: Common constraint V_C and common learning model L. The constraint V_C is independent from any prior knowledge about R, in other words it does not depend on any sort of first order global information of F_L.

Proof:

Real numbers can be represented by parallel hyperplanes defined by a vector set V_C. Because of corollary *2* we know:
$\forall X_S \subset D, \exists X_N \subset D \setminus X_S$: $F_L(X_S) \neq F_L(X_N)$ (Corollary *2*)
Therefore, $L(X_S) \cap L(X_N) = \phi$ (satisfies Law *1*)

This corollary carries double meaning:

1. Linear separability is the common constraint for all models that could convert to that satisfy Law *1*.
2. Without the harvesting function, the learning result of model M_L cannot be recognised, so the model F_L (as shown in Figure 4.6) is supposed to be the mapping relation that could eventually provide useful information. However, in spite of the existence of infinitely many possible H, it is almost impossible to locate a suitable harvesting function without violating Law *2*. Corollary *3* shows that V_C is independent of any first order global information of F_L (prior knowledge about the undetected concepts)[5]. Therefore it has so far been the only known family of implementable

[5]It is a common sense that intelligent creatures are able to learn concepts from completely unfamiliar environment.

harvesting functions[6] and it cooperates only with model 'L'. This corollary also explains researchers' intuitive preference for linear separability.

4.6 The point of corollary two and three

As shown in figure 4.6 two subsets in space R^n could tightly entangled with each other, corollary two and three proves that any two subset of R^n could have corresponding set of global representations which are parallel to each other. Thus, the behaviour of the common learning model L is variance reduction.

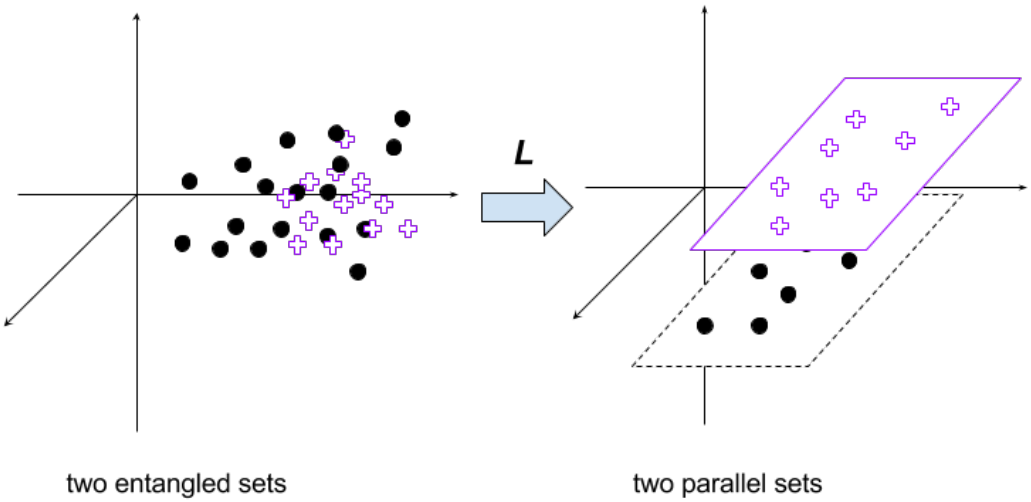


Figure 4.7: Variance reduction of common learning model L

By expressing each point in a high-dimensional space, as shown in Figure 4.7, with a periodic function:

$$f_X(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t... \tag{4.3}$$

Function $f_X(t)$ is plotted on the range $-\pi < t < \pi$, and X is a high dimensional point[26]. Therefore, L could be seen as a transformation of functions, and the behaviour of L can be expressed as being shown in Figure 4.8.

[6] V_C is a set of vectors.

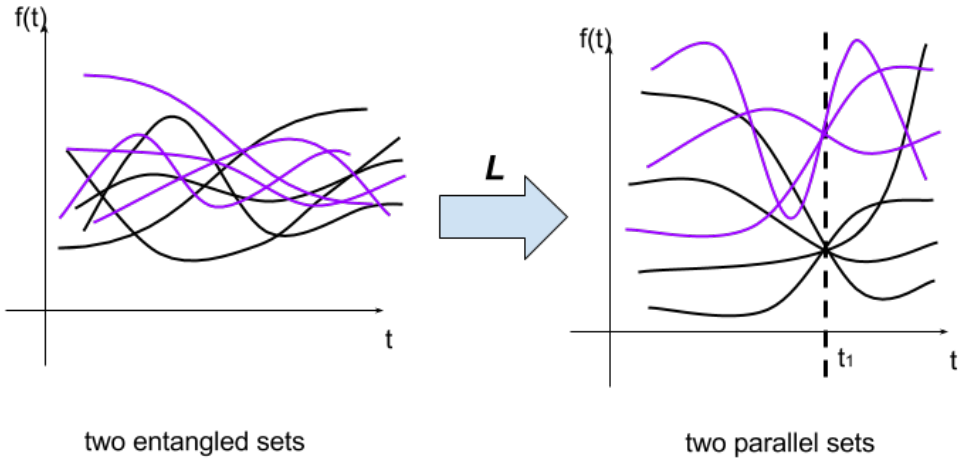
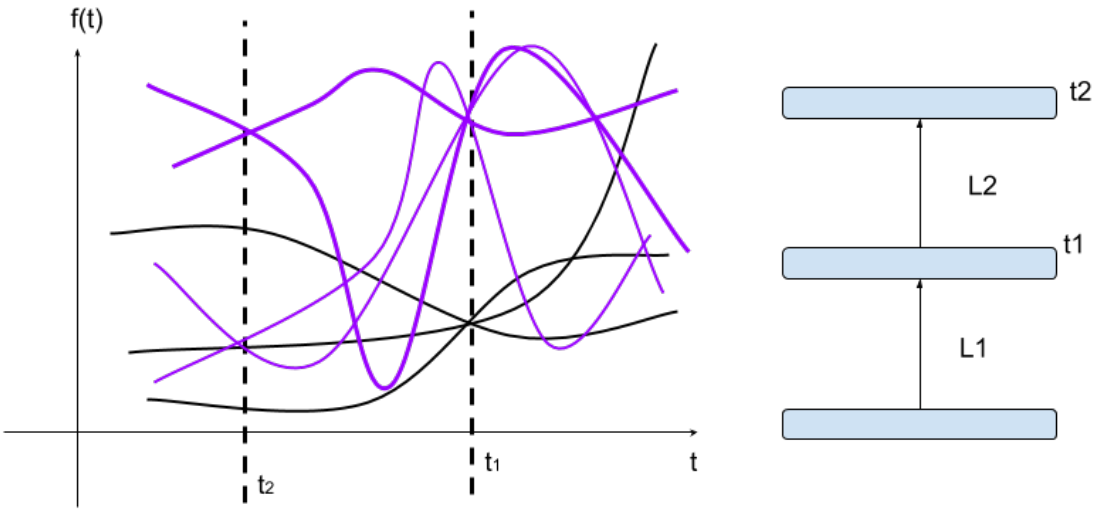


Figure 4.8: Variance reduction of common learning model L expressed as family of functions.

Two set of points as shown in Figure 4.8, will be expressed as two sets of functions (Black and Purple), after being transformed, each set of functions will cluster to each other at some value t_1. One important work is to solve L by adopting the notion of harmonic analysis[7]. Because of the hierarchical structure of global and local relation, repeat the transformation L, differences within each family of functions can be defined at higher level, as shown in figure 4.9.

[7]This work has been stopped since last year this time.



parallel between different categories and within the same category

Figure 4.9: L_1 and L_2 are the same model, because this hierarchical structure. They could defined different concept with different level of grain.

This mechanism solve author's question (this question is proposed in the original PhD research proposal) about the relation between learning to identify different races and learning to identify different individuals of the same race. For example, global representations of two races are parallel to each other at the hyperplane defined by t_1, global representations of two individuals of the purple category will parallel to each other at the hyperplane defined by t_2.
The behaviour of a common learning model L is variance reduction (single algorithm hypothesis), because of the hierarchical structure and the different property of different input space, different information will be learned (Definition *3, modular mind hypothesis.* and by keep extending this hierarchical structure different grain of concept could be defined (scalable cortex hypothesis).

4.7 Explanatory Comment

This section explains the context of the two laws and three corollaries above in order to be understood. To understand the essence of these laws and corollaries, it is necessary to understand what information can and cannot be gained and what cannot from the viewpoint of a model rather than as a creator of a model.

1. Since the existences of concepts are defined based on different appearances, the notion of 'right' or 'wrong' is redundant in this situation. More precisely, for a learning model $M_L : D \to O$ talking about whether the model $x \to y(x \in X \subset D; y \in Y = M(x))$ is correct or not is meaningless; in other words, any criteria that can be used to test this model only provides first order global

information with respect to M_L. For example, when explaining the object recognition problem using the theoretical framework proposed in this dissertation, two seemingly counter-intuitive deductions are dataset separation and dataset merge problems.

- Dataset separation problem: When a dataset generated by one object is separated, there could be two different sets of invariant representations, such as A and B in Figure 4.10.
- Dataset merge problem: When two datasets of different objects are merged together, there could be a new set of invariant representations, such as C and D in Figure 4.10.

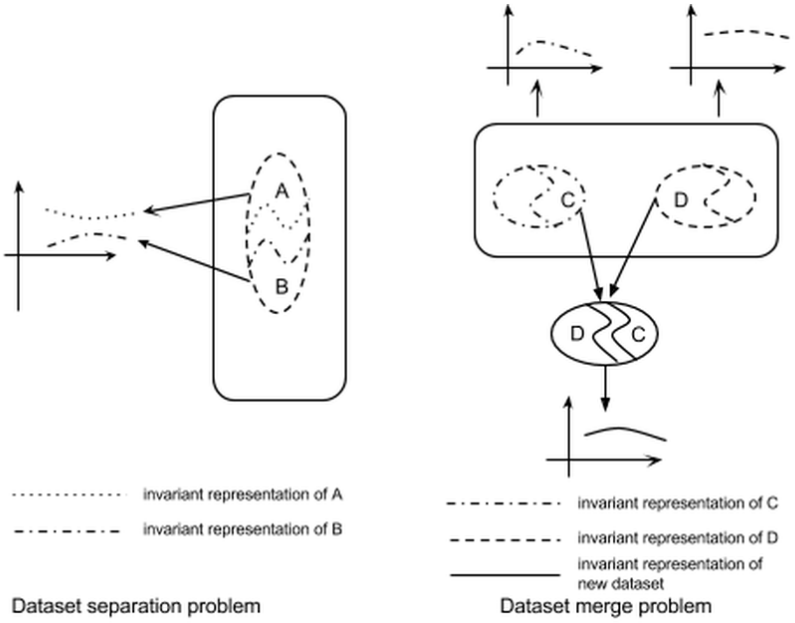


Figure 4.10: Dataset separate and merge problems

The above description is the common intuitive understanding of these two problems, but there are two mistakes concerning this understanding:

- The two concepts are defined by the model, so these two problems are equivalent.
- Since there is no observation independent concept, solely discussing the existence of a certain concept is meaningless and the example shown in Figure 4.10 is often known as an over-fitting problem in the field of machine learning. However, from the viewpoint of a learning model there is no notion of 'right' or 'wrong'; therefore it is not a problem. It is a phenomenon that can be used to verify the validity of future implementations based on this theoretical framework. This phenomenon has also been proved by neuroscientists[27].

2. Law *1*, together with Corollary *3* indicate that harvest function V_C could utilise global representations for all possible concepts without the existence of L. This implication seems irrational at first glance; however by assuming the well accepted learning model (human) is a learning model as described in Corollary *3* ($V_C \circ L$), the correctness of this implication is undeniable. This implication is equivalent to:

   People can see everything they could possible see in the future.

Most people do not realise this fact in their daily lives, but this is the foundation of creative activities such as painting or sculpture [8]. It is not hard to image that a painter would still be able to create a masterpiece once he/she lost his/her sight, and for normal people who do not master this painting skill the experience of dreaming of strange-looking creatures is not unusual. In this case, when V_C and L were being constructed, all dimensions of the domain contain same type of information and the appearances of a concept are fairly straightforwardly provided. What if the domain consists of different types of information? We know that people who merely remember the answers of a certain examination or all past examinations will not be acknowledged as having learned the concept of the corresponding subject. On the other hand, people who learned concepts of a certain subject could not only give answers to every possible related question but also see facts that cannot be seen by people who merely remember the answers. This is an example of a non-learning model(memory system) and learning model that are related to a high-level concept. In this case a high-level concept will be defined based on different types of information and apparently how different types of information being clustered will largely affect whether higher level concept can be effectively learned or not[9]. This belongs to the discussion of intelligence which is beyond the scope of this dissertation.

3. For model $M_L : D \to O, (D \subset R^m, O \subset R^n)$, m, n and the range of each dimension are assumed to be a very large number, to infinity ideally . The reason for this assumption is straightforward. It is apparent that a domain $D \subset R^3$ is less likely to contain less information than only two of its recorded dimensions, and it is also less likely to contain less information than only the integer part of all recorded dimensions. The direct consequence of having a domain which contains only a small amount of information is that there will be not be enough information to define different concepts. An intuitive example of this scenario would be when a person with a high degree of myopia accidentally loses his/her glasses. Since the domain is supposed to be big enough to carry large amounts of information, it is reasonable to assume that the dimensionality of the range is high enough to contain enough parallel hyperplanes, and the range of each dimensionality seems unimportant at this stage, but it is directly related to the further analysis of the model L. It is worth mentioning that shrinking the range of each dimensionality is clearly a strategy that enable a non-learning model which does not satisfy Law *1* behaves like it is able to learn the domain.

Up to this stage a summary that explains relations in each chapter of this dissertation will be given, so that readers would have a better understanding of this dissertation.

First, by giving definitions of local and global information (representations), the intuitive understanding of the relation between global and local can be explained in the framework of mapping relation. Based on this formalised expression of information, Law *1* indicates that for a mapping relation to be a learning model, it should at least not be an information black-hole. Law *2* confines possible information used in constructing the so that from a s point of view it will not be able to utilise information that goes beyond its field of vision (first order global information). Corollary *1* proves the existence of a set of equivalent constraints which can harvest global information from a learning model, Additionally, all possible sets of equivalent constraints are functions. Corollary *2* proves a learning model M_L, together with any of

[8] The dynamic property of a model is out of scope here.

[9] Examples about high level concepts are provided at the beginning of chapter 3, such as learning the meaning of a word.

its possible harvest function H still satisfies Law *1* and different concepts of any type of information M_L can be represented by different real numbers although there are infinitly many possible choices and it could be impossible to find any of them. Corollary *3* proves it is possible to have a harvest function V_C that is independent from any possible first order global information and all learning model M_L has a corresponding family of mapping relation L that are also learning models.

CHAPTER 5

Conclusion and Future Works

Recently, the success of the Go-playing program Alpha-Go has surprised the public. People started comparing its learning ability with top level Go players. However, necessary information used to construct Alphago represents fundamental differences between a tradition learning model and a real learning model (us). Assuming we double the size of a Go board, would any one still claim that Alphago has actually 'learned' how to play Go? There size of a Go board is first order global information with respect to every position of the board, and unlike Alphago, what human learned about GO would not be affected very much by simply changing the size of the board.
Without a fundamental interpretation regarding what a given ability is, there could be dramatic differences between constructing a near-this-ability-performance system and constructing a having-this-ability-performance system. This dissertation gives two laws, which are necessary conditions for a model to be acknowledged to have learning ability (to be a learning model). This dissertation also illustrates these following facts:

- "Learning" is the ability of identifying the existence of a new concept.
- A model M_L will be acknowledged as a learning model of information M_L only when it is able to possess this "learning" ability with no help from other learning models. In other words, being able to learn spontaneously.
- If the mathematical expression (model) of our hypothesis of observation is not constructed carefully, information provided by us (the creator of the model) can exceed the vision of the model very easily and cause the model to be a non-learning model. Therefore, further development based on this model for achieving a human-level learning ability can only lead to inevitable failure. For example, using labels to represent our observations in typical classification problems will directly lead the hypothesises of observation to fail to satisfy Law *2* and *1*.
- There exists a common learning model that utilises the power of hyperplanes.
- The behaviour of a common learning model is variance reduction.

By inspecting from the viewpoint of mapping relations and treating them equally (human, animal, and apparatus), these key ideas of this dissertation can be appreciated more clearly.
This dissertation, as the primary version of the first part of my original PhD thesis, brings more questions than it has addressed. The following two parts include a discussion about the model $L(x, \Theta)$ and how the

parameters change dynamically. Further illustration and explanation on these two issues will address the following questions:

- Under what circumstances is the hierarchy structure[1] of a learning model necessary?
- In a typical machine learning problem, the learning result is a $\theta \in \Theta$, however, for a learning model, compared with obtaining desirable θ, a more valuable question is "how a learning model encodes the unknown constraint of local representations of all concepts in the domain?".
- How could the possible existence of concepts depend on the information being contained in the domain quantitatively?
- The dependency relation between learning model and non-learning model (memory system).
- How could harmonic analysis be used to solve (to get L) the transformation being discussed in Chapter 4 ?

These days, swing into the saddle does not mean people are going on a long journey; jogging on the pavement is usually just for exercise. After the industrial revolution, people have continued inventing all different kinds of machines that enable us to exceed the physical limitations of our biological blueprint. Therefore, the ability of invention, or more broadly, intelligence is the proudest property of human and it has not been simulated by any man-made machine successfully, yet. This dissertation is one of the many steps to the inevitable future when humans might not be an absolutely necessary information resource for automatic systems and we could harvest knowledge that cannot be provided by our learning ability.

[1] Currently, there is no explanation of the necessity of having layer structure for both biological and artificial neural networks.

Bibliography

[1] Quoc V Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.

[2] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

[3] Edward L Thorndike. The fundamentals of learning. 1932.

[4] Donald Olding Hebb. *The organisation of behaviour: a neuropsychological theory*. Wiley, 1952.

[5] AG Oettinger. Simple learning by a digital computer. *Proceedings of the Association for Computing Machinery, September7*, 1952.

[6] HL Gelernter and N Rochester. Intelligent behavior in problem-solving machines. *IBM Journal of Research and Development*, 2(4):336–345, 1958.

[7] Allen Newell, John C Shaw, and Herbert A Simon. Report on a general problem-solving program. In *IFIP Congress*, pages 256–264, 1959.

[8] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[9] John McCarthy. Professor sir james lighthill, frs. artificial intelligence: A general survey. *Artif. Intell.*, 5(3):317–322, 1974.

[10] Tom M. Mitchell. *Machine Learning*. WCB McGraw-Hill, 1997.

[11] Alan M Turing. Computing machinery and intelligence. *Mind*, pages 433–460, 1950.

[12] Pamela McCorduck. Machines who think. 2004.

[13] David Berlinski. The advent of the algorithm: The idea that rules the world. *AMC*, 10:12, 2000.

[14] Yu-Chi Ho and David L Pepyne. Simple explanation of the no-free-lunch theorem and its implications. *Journal of optimization theory and applications*, 115(3):549–570, 2002.

[15] Thomas M English. Optimization is easy and learning is hard in the typical function. In *Evolutionary Computation, 2000. Proceedings of the 2000 Congress on*, volume 2, pages 924–931. IEEE, 2000.

[16] Thomas L Dean, Greg Corrado, and Jonathon Shlens. Three controversial hypotheses concerning computation in the primate cortex. In *AAAI*, 2012.

[17] David H Wolpert and William G Macready. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82, 1997.

[18] Vladimir Vapnik. *The nature of statistical learning theory*. springer, 2000.

[19] Peter L Bartlett and Wolfgang Maass. Vapnik-chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*, pages 1188–1192, 2003.

[20] Wolfgang Maass. Bounds for the computational power and learning complexity of analog neural nets. *SIAM Journal on Computing*, 26(3):708–732, 1997.

[21] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *Electronic Computers, IEEE Transactions on*, (3):326–334, 1965.

[22] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[23] Sebastian Luft. *Subjectivity and lifeworld in transcendental phenomenology*. Northwestern University Press, 2011.

[24] Carl Sagan. *Demon-haunted world: science as a candle in the dark*. Ballantine Books, 2011.

[25] R.W. Anderson. *The Cosmic Compendium: Black Holes*. LULU Press, 2015.

[26] David F Andrews. Plots of high-dimensional data. *Biometrics*, pages 125–136, 1972.

[27] Nuo Li and James J DiCarlo. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron*, 67(6):1062–1075, 2010.