Logifold: A Geometric Fondation of Ensemble Machine Learning

Inkee Jung [1] Siu-Cheong Lau [2]

[1] PhD student of Mathematics
Boston University

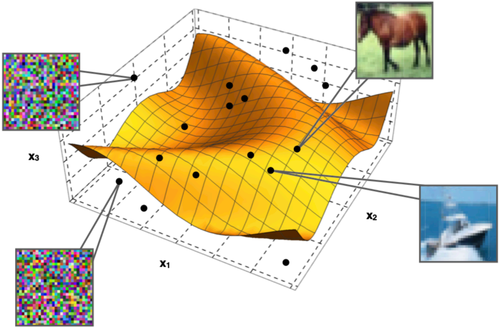
[2] Faculty of Mathematics
Boston University

November 4

International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME 2024)
4-5 November 2024, Male, Maldives

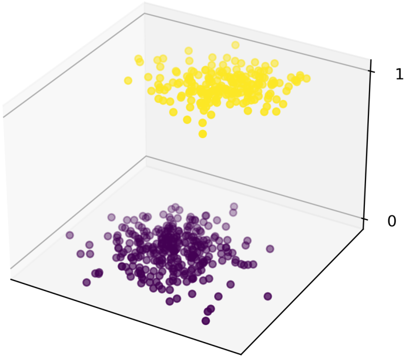
“Manifold” in Data Science

High-dimensional analogue of 2 dimensional surface in $\mathbb{R}^N$

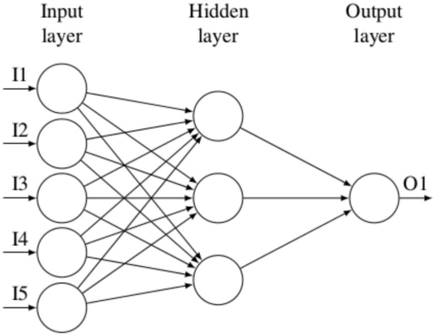


(Image from Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová)

Classification Dataset and Neural Network



Classification with two classes



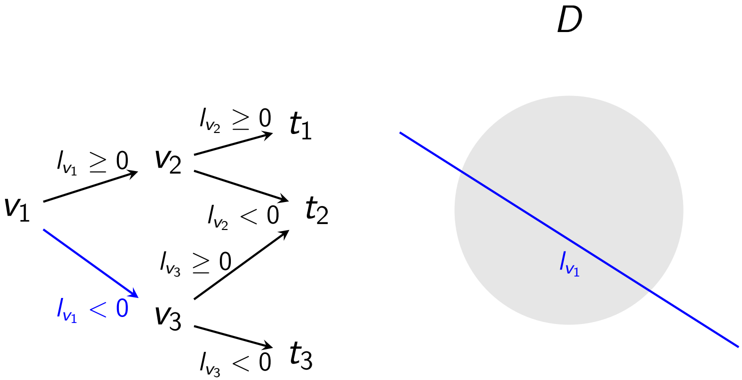
$$f = \sigma_2 \circ L_2 \circ \sigma_1 \circ L_1$$

- Network models gain tremendous success in describing datasets

Linear Logical Function

Motivated from Neural Network.

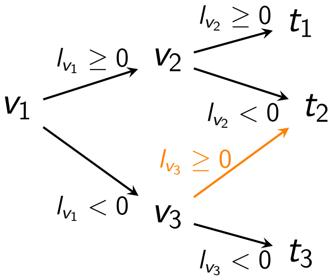
Example: Directed graph G & Set of affine maps $L = \{l_{v_1}, l_{v_2}, l_{v_3}\}$, $D \subset \mathbb{R}^2$

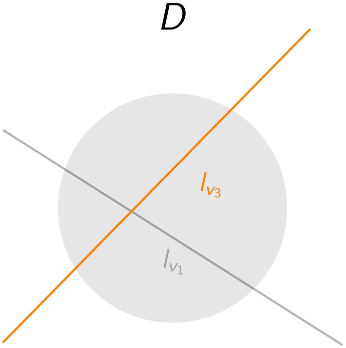


Linear Logical Function

Motivated from Neural Network.

Example: Directed graph G & Set of affine maps $L = \{l_{v_1}, l_{v_2}, l_{v_3}\}$, $D \subset \mathbb{R}^2$

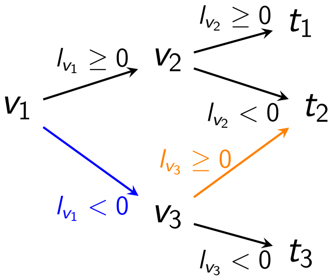


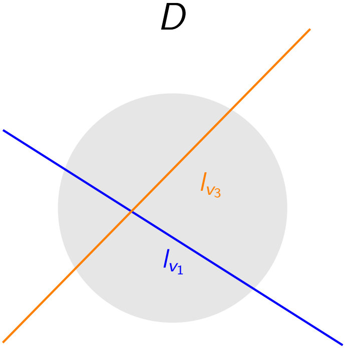


Linear Logical Function

Motivated from Neural Network.

Example: Directed graph G & Set of affine maps $L = \{l_{v_1}, l_{v_2}, l_{v_3}\}$, $D \subset \mathbb{R}^2$

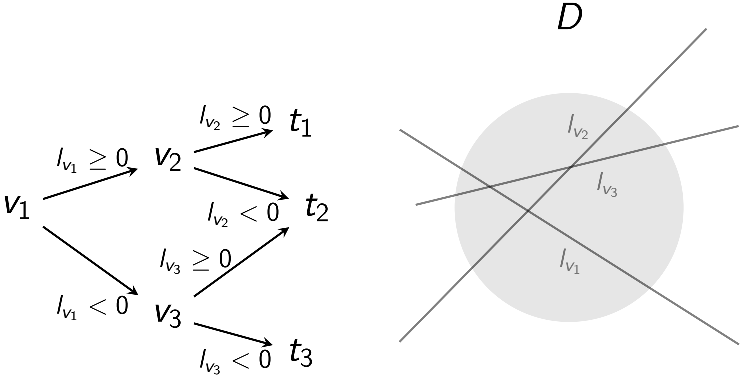




Linear Logical Function

Motivated from Neural Network.

Example: Directed graph G & Set of affine maps $L = \{l_{v_1}, l_{v_2}, l_{v_3}\}$, $D \subset \mathbb{R}^2$



$f : D \to \{t_1, t_2, t_3\}$ is a function defined by G and L.

Linear Logical Function

- Measurable set $D \subset \mathbb{R}^n$, Finite set T.
- Directed finite graph G without cycle
- Affine maps

$$L = \{l_v : v \text{ is a vertex with more than one outgoing arrows}\}$$

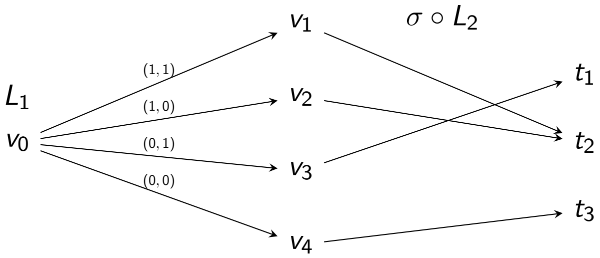
Definition

$f_{G,L} : D \to T$ is a linear logical function of (G, L) if $l_v \in L$ are affine linear functions whose chambers in D are one-to-one corresponding to the outgoing arrows of v.
(G, L) is called a linear logical graph.

Linear logical function : Example

$f = \sigma \circ L_2 \circ s \circ L_1$ where

- $L_1 : \mathbb{R}^n \to \mathbb{R}^2$ is affine map and s is a component-wise step function.
- $L_2 : \mathbb{R}^2 \to \mathbb{R}^3$ is affine map and σ is the index-max map.



f is a linear logical function with the above graph G and $L = \{L_1\}$.

Fuzzy linear logical function : Example

$f = \sigma \circ L_2 \circ s \circ L_1 : S^n \to S^3$ with SoftMax σ and ReLU s.

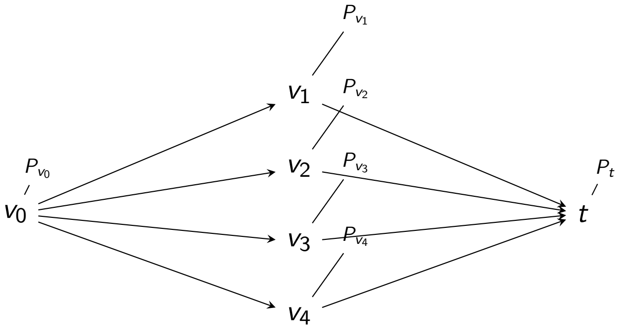
- G is a finite directed graph that has no oriented cycle with exactily one source vertex and target vertex t.

$$\begin{tikzcd}
 & v_1 \arrow[ddr] & \\
 & v_2 \arrow[dr] & \\
v_0 \arrow[uur] \arrow[ur] \arrow[dr] \arrow[ddr] & & t \\
 & v_3 \arrow[ur] & \\
 & v_4 \arrow[uur] &
\end{tikzcd}$$

Fuzzy linear logical function : Example

$f = \sigma \circ L_2 \circ s \circ L_1 : S^n \to S^3$ with SoftMax σ and ReLU s.

- Each vertex v of G is equipped with a product of standard simplices P_v, with domain $D = P_{v_0}$.

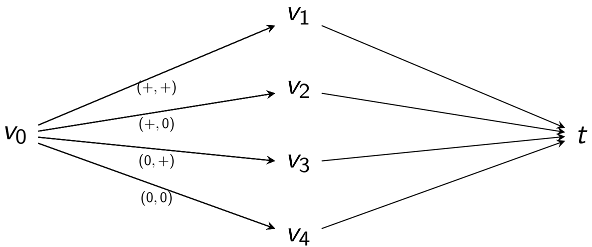


$$P_{v_0} = P_{v_1} = P_{v_2} = P_{v_3} = S^n, P_t = S^3$$

Fuzzy linear logical function : Example

$f = \sigma \circ L_2 \circ s \circ L_1 : S^n \to S^3$ with SoftMax σ and ReLU s.

- Arrow maps $p_a : P_{s(a)} \to P_{t(a)}$ for each arrow a, and affine map l_v whose chambers in P_v are one-to-one corresponding to the outgoing arrows of v.



$p =$ identity between input and hidden vertex $\qquad p = \sigma \circ l$

$L_{v_0} = L_1$ and l is the restricted affine linear map on chambers made by L_{v_0} and the ReLU activation s.

Fuzzy linear logical function

- G is a finite directed graph that has no oriented cycle with exactily one source vertex and target vertices $t_1, \ldots, t_K$.
- Each vertex v of G is equipped with a product of standard simplices P_v, where simplex is a set of the form $\{x \in \mathbb{R}^{d+1} : \sum x_i = 1\}$. Domain D is a subset of P_{v_0}.
- Each arrow a is equipped with a continuous function

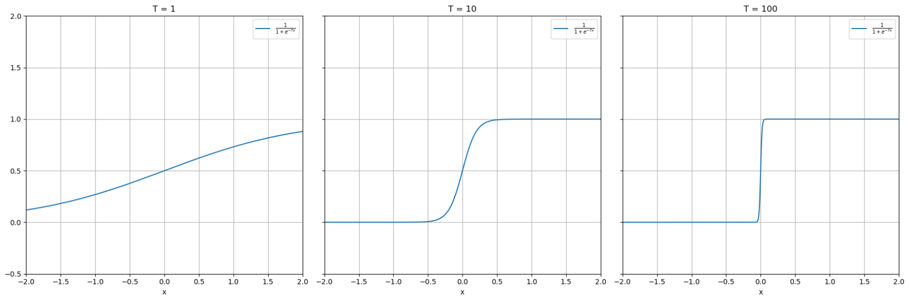
$$p_a : P_{s(a)} \to P_{t(a)}$$

  where $s(a), t(a)$ denote the source and target vertices respectively.
- Each vertex v that has more than one outgoing arrows is equipped with affine map l_v whose chambers in P_v are one-to-one corresponding to the outgoing arrows of v.

Given $x \in D$, L and p_a determine a path to a target, and $f_{(G,L,P,p)}(x)$ is defined by the composition of arrow maps along the path.

Tropical limits

Introduce formal parameter T to logistic functions.



$$\lim_{T\to\infty} \frac{1}{1+T^{-x}} = \begin{cases} 1 & (x>0) \\ 0 & (x<0) \end{cases}$$

$$\mathrm{SoftMax}(x) \xleftarrow{T\to e} \left(\frac{T^{-x_k}}{\sum_i T^{-x_i}}\right) \xrightarrow{T\to 0^+} \mathrm{Argmax(x)}$$

Universality of Linear logical function

- $D \subset \mathbb{R}^N$ with $\mu(D) < \infty$, where μ is the Lebesgue measure.
- T is finite

Theorem (I. Jung and S.C. Lau)

For any (Lebesgue) measurable function $f : D \to T$, we have a linear logical function that approximates to f.

Universality of Linear logical function

- $D \subset \mathbb{R}^N$ with $\mu(D) < \infty$, where μ is the Lebesgue measure.
- T is finite

Theorem (I. Jung and S.C. Lau)

For any (Lebesgue) measurable function $f : D \to T$, we have a linear logical function that approximates to f.

Corollary

There exists a family $\mathcal{L}$ of linear logical functions $L_i : D_i \to T$, where $D_i \subset D$ and $L_i \equiv f|_{D_i}$, such that $D \setminus \bigcup_i D_i$ is measure zero set.

Fuzzy linear logical function and fuzzy linear logifold

Definition

A fuzzy linear logifold is a tuple $(X, \mathcal{P}, \mathcal{U})$, where $(X, \mathcal{U})$ be a logifold and

- $\mathcal{U}$ is a collection of tuples (ρ_i, ϕ_i, f_i)
- $\rho_i : X \to [0,1]$ describe fuzzy subsets of X with $\sum_i \rho_i \leq 1_X$
- $U_i = \{x \in X : \rho_i(x) > 0\}$ be the support of ρ_i

In classification problems,

- $X = \mathbb{R}^n \times T$
- $\mathcal{P} : X \to [0,1]$ describes how likely an element of $\mathbb{R}^n \times T$ is classified as ‘yes’
- ρ_i can be ‘generalization performance’, or ‘constant’.

Example of logifold

$f : (0, 1] \to \{0, 1\}$ be a function defined as

$$f(x) = \sum_{n=0}^{\infty} \left(\frac{(-1)^n + 1}{2} \right) I_{E_n}(x)$$

where $E_n = (1 - 2^{-n}, 1 - 2^{-n-1}]$.
The graph of $f \subset [0, 1) \times \{0, 1\}$

with countably many 'jumps' or 'discontinuities' near at $x = 0$.

In classification problems, $X = \mathbb{R}^n \times T$ and each model $g_i : X \to T$ with $U_i = X$. Define $G_i : X \times T \to [0,1]$ by g such that $G_i(x,t) = (g_i(x))_t$. Let N be the total number of classifiers.

- If $\rho_i = \frac{1}{N}$ for any i, then $P : X \times T \to [0,1]$ is defined by

$$P(x,t) = \sum \rho_i(x) 1_{t_{i,0}(x)}(t)$$

, where $t_{i,0}(x) = \arg\max G_i(x,t)$ denoting 'the answer of g_i', and therefore the system employs majority voting.

In classification problems, $X = \mathbb{R}^n \times T$ and each model $g_i : X \to T$ with $U_i = X$. Define $G_i : X \times T \to [0, 1]$ by g such that $G_i(x, t) = (g_i(x))_t$. Let N be the total number of classifiers.

- If $\rho_i = \frac{1}{N}$ for any i, then $P : X \times T \to [0, 1]$ is defined by

$$P(x, t) = \sum \rho_i(x) 1_{t_{i,0}(x)}(t)$$

  , where $t_{i,0}(x) = \arg\max G_i(x, t)$ denoting 'the answer of g_i', and therefore the system employs majority voting.

- If $\rho_i = \frac{1}{N}$ for any i, then $P : X \times T \to [0, 1]$ is defined by

$$P(x, t) = \sum G_i(x, t)$$

  , which is simple average.

In classification problems, $X = \mathbb{R}^n \times T$ and each model $g_i : X \to T$ with $U_i = X$. Define $G_i : X \times T \to [0, 1]$ by g such that $G_i(x, t) = (g_i(x))_t$. Let N be the total number of classifiers.

- If $\rho_i = \frac{1}{N}$ for any i, then $P : X \times T \to [0, 1]$ is defined by

$$P(x, t) = \sum \rho_i(x) 1_{t_{i,0}(x)}(t)$$

  , where $t_{i,0}(x) = \arg\max G_i(x, t)$ denoting 'the answer of g_i', and therefore the system employs majority voting.

- If $\rho_i = \frac{1}{N}$ for any i, then $P : X \times T \to [0, 1]$ is defined by

$$P(x, t) = \sum G_i(x, t)$$

  , which is simple average.

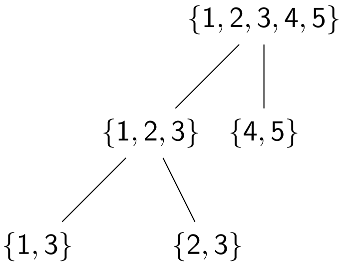
- If $\rho_i(x) = \frac{\max g_i(x)}{N}$ then $P(x, t) = \sum \rho_i(x) G_i(x, t)$ be the weighted average.

As our logifold formulation does not force to have X and T as domain/codomain of classifier, we allow classifier to have more flexibility in its target.

For instance, our classification problem is classifyting instances in X to $\{1, 2, 3, 4, 5\}$, and we have models $g_1, \ldots, g_7$ such that

Models	Targets
g_1	$\{1, 2, 3\}, \{4, 5\}$
g_2, g_3	$\{1, 2, 3, 4, 5\}$
g_4	$\{1, 2, 3\}$
g_5	$\{1, 3\}$
g_6	$\{2, 3\}$

As they can have various target, we make tree of targets.
For instance, with $\{1,2,3,4,5\}, \{1,2,3\}, \{1,3\}, \{2,3\}$, we have the following target tree.



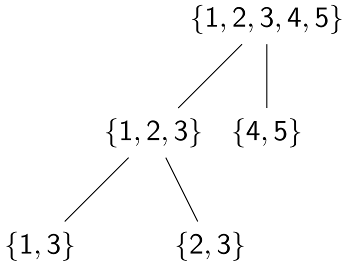
On validation dataset, define first certain domain of g under the certainty threshold α.

$$\text{Certainty} = \max g(x)$$

$$\text{Certain domain} = \{\text{certainty} > \alpha\}, \quad \alpha = \text{threshold}$$

Then compute accuracy (global, and in each target) of g.
For instance, g_2 has following fuzzy domain:

certainty threshold	Accuracy	Accuracy in each target
0	0.6	(0.7,0.8,0.45,0.5,0.45)
0.8	0.7	(0.7,0.9,0.5,0.7,0.6)
0.95	0.8	(0.8,0.9,0.75,0.8,0.75)



- For a given instance x, we can compute weighted voting for x at node $\{1, 2, 3, 4, 5\}$ according to the fuzzy domain of g_1, g_2, g_3 in each target $1, 2, 3, 4, 5$.
- If answer for $1, 2, 3$ is dominant, then we pass it to $\{1, 2, 3\}$ node. In this way, we have unique path in the target tree for each instance.
- On validation dataset, we can compute which (sub-)path and certainty threshold are optimal for best accuracy in each model.

Experimental Result 1

Dataset : CIFAR10
Six Simple CNN structure models trained on CIFAR10 (56.45% in average)
ResNet20 structure model trained on CIFAR10 (85.96%)
Simple average : 62.55%
Majority voting provides 58.72%.
Our logifold formulation : 84.86%

Experimental Result 2

dataset : CIFAR10, MNIST, Fashion MNIST (resized to 32*32*3 pixels)

- Filters are models classifying coarse targets. It only classify given data into three classes ; CIFAR10, MNIST, and Fashion MNIST.
- Models only classifying either CIFAR10, MNIST, or Fashion MNIST.

Single model classifying 30 classes : 76.41% in average.
Simple average of models classifying 30 classes : 82.35%
Our logifold formulation : 94.94%.