Teoría Matemática de Redes Neuronales

Notas de Clase

Miguel Arturo Ballesteros Montero

Facultad de Ciencias, UNAM

Noviembre, 2020

Índice general

Prólogo

Durante miles de años, hemos intentado comprender como pensamos, la historia nos ha regalado grandes pensadores que han tratado de estructurar de diversas formas el complejo organismo que somos, la inteligencia artificial intenta ir un paso más allá, intenta no solo comprender si no también construir entidades inteligentes. La inteligencia Artificial es uno de los campos más nuevos de la ciencia y la ingeniería que encontró su origen poco después de la Segunda Guerra Mundial.

Inteligencia Artificial a traves del tiempo:

- **1943-1955:** La gestación de la inteligencia artificial.
    - Modelo de neuronas artificiales de Warren McCulloch y Walter Pitts en 1943.
    - Turing dio conferencias sobre IA desde 1947.
- **1956:** El nacimiento de la inteligencia artificial.
    - John MaCarthy convenció a Marvin Minsky, Claude Shannon y Nathaniel Rochester para que lo ayudaran a reunir investigadores estadounidenses interesados en la teoría de autómatas, las redes neuronales y el estudio de la inteligencia en un taller de dos meses en Dartmouth.
- **1952-1969:** Entusiasmo inicial, grandes exprectativas.
    - Primeros solucionadores de problemas, jugadores de juegos, probadores de teoremas.
    - John McCarthy se refirió a este periodo como el "¡Mira mama, sin manos!"
    - Creación de LISP
    - *Perceptron* de Frank Rosenblatt en 1958.
    - Adalines (*Adaptive Linear Neuron*) de Bernie Widrow y Marcian Hoff en 1960.
- **1966-1973** Una dosis de realidad.

- Prueba y error: explosión combinatoria.
- Falta de recursos computacionales.

- **1969-1979:** Sistemas basados en el conocimiento ¿La clave del poder?
  - Algoritmos que utilizan conocimientos específicos de dominio en lugar de solucionadores de propósito general.
  - Sistemas expertos para diagnóstico médico.
  - Incorporación de incertidumbre.

- **1980-presente:** La IA se convierte en industria.
  - Optimización de la logística.
  - Auge repentino, pero solo unos pocos proyectos estuvieron a la altura de las expectativas.
  - Invierno IA.

- **1986-presente:** El regreso de las redes neuronales:
  - El algoritmo de retropropagación para entrenar redes neuronales se reiventó en "Learning Representation by Back-Propagating Errors."Por David E. Rumelhart, Geoffrey E. Hinton y Ronald J. Williams.

- **1987-presente:** La IA adopta el método científico.
  - Modelos ocultos de Markov.
  - Redes Bayesianas.

- **1995-presente**
  - Internet impulsa el desarrollo de agentes inteligentes, por ejemplo:
    - Chatbots.
    - Sistemas de recomendación.
    - Recomendaciones de amistades.
  - Acceso a recursos de computación a suficiente velocidad.
  - La era del *Big Data:* Gran cantidad de *label training data*, por ejemplo:
    - Diccionarios.
    - Wordnets.
    - Wikipedia.
    - Google.

- Los fundadores de la IA descontentos con su estado actual:
  - La IA debería volver a sus raíces de luchar por, en palabras de Herbert Simon, “Maquinas que piensan, que aprenden y crean”.
- La IA en la cultura popular.
  - Lucha contra el spam.
  - Reconocimiento de voz: Siri, Alexa, Cortana.
  - Reconocimiento facial: Facebook, Apple Photos, Google Photos.
  - *Deep Blue vs Garry Kasparov*.
  - Planificación y programación autónoma: Mars rover de la NASA.
  - Vehículos robóticos: EL automóvil autónomo de Tesla.
  - Traducción automática: Traductor de Google.

- **Hoy** Tu empiezas a leer estas notas.

1 | ¿Que es la Inteligencia Artificial?

1.1. Definiciones Históricas

1.1.1. Actuar Humanamente: El enfoque de la prueba de Turing

> "El arte de crear maquinas que realizen funciones que requieren inteligencia cuando las personas lo ejecuten"
>
> Kurzweil, 1990

> "El estudio de hacer a las computadoras hacer cosas en las que de momento, somos mejores."
>
> Rich and Knight, 1991

En 1950 Turing ideo una prueba para proporcionar una definición operativa satisfactoria de inteligencia; Una computadora pasa la prueba si un interrogador humano, después de formular algunas preguntas escritas, no puede decir si las respuestas escritas provienen de una persona o una computadora.

La computadora debe poseer las siguientes características:

- **Procesamiento del lenguaje natural:** para comunicarse, por ejemplo en inglés.
- **Representación del conocimiento:** para almacenar información
- **Razonamiento automatizado:** utilizar la información almacenada, responder preguntas y sacar conclusiones
- **Aprendizaje automático:** adaptarse a nuevas circunstancias, extrapolar y detectar patrones.

La prueba de Turing sigue siendo relevante incluso hoy, pero menos desde la ingeniería y más desde la postura filosófica, en ejemplo parecido sería:

- La busqueda del "vuelo artificial" tuvo éxito cuando los hermanos Wright y otros dejaron de imitar a las aves y comenzaron a usar túneles de viento y aprender sobre aerodinámica.
- Los textos de ingeniería aeronáutica no definen el objetivo de su campo como fabricar "Máquinas que vuelan tan exactamente como las palomas que pueden engañar incluso a otras palomas"

1.1.2. Pensar Humanamente: El enfoque de modelado cognitivo

> "El nuevo y emocionante esfuerzo de hacer que las computadoras piensen [...] maquinas con mentes, en el sentido completo y literal."
>
> Haugeland, 1985

> "[La automatización de] actividades que asociamos con el pensamiento humano, actividades como la toma de decisiones, la resolución de problemas, el aprendizaje [...]"
>
> Bellman, 1978

Para saber si un programa "piensa como un ser humano", debemos aprender qué es el pensamiento humano:

- Observa nuestros pensamientos a medida que pasan
- Observar a las personas durante una tarea
- Experimentos psicológicos/neurocientíficos

Sin embargo, para nuestro esfuerzo, será una buena práctica mantener separados los campos como la ciencia cognitiva, la neurociencia, la psicología y la filosofía el mayor tiempo posible.

1.1.3. Actuar Racionalmente: El enfoque del agente racional

> "La Inteligencia Computacional es el estudio del diseño de agentes inteligentes"
>
> Poole et al., 1998

> "IA [...] se preocupa por el comportamiento inteligente de los artefactos"
>
> Nilsson, 1998

Crear agentes (por ejemplo, programas informáticos) que operen de forma autónoma, perciban su entorno, persistan durante un período de tiempo prolongado, se adapten al cambio, creen, persigan metas. Hacer inferencias es el caso extremo de ser un agente racional.

En muchas ocasiones, sin embargo, no es posible realizar inferencias correctas, por ejemplo:

- Comprensión insuficiente del medio ambiente
- No hay suficientes datos de entrada para basar una decisión

Un agente racional es aquel que actúa para lograr el mejor resultado ó, cuando hay incertidumbre, el mejor resultado esperado.

De esta forma, el estándar de racionalidad está matemáticamente bien definido y es completamente general (como mostraremos en la siguiente sección con el enfoque de "Las leyes del pensamiento"), por otro lado el comportamiento humano, esta bien adaptado para un entorno específico y se define a si mismo por la suma total de todas las cosas que hacen los humanos.

1.1.4. Pensar Racionalmente: El enfoque de las leyes del pensamiento

Un uso de razonamiento lógico y argumentación, por ejemplo:

- **Deducción:** Una regla general aplicada a un caso particular implica un resultado trivial.

Entrada	**Implicación**	**Ejemplo**
REGLA		En un planeta, su sol sale todos los días
CASO		Estamos en un planeta
	RESULTADO	El sol sale todos los dias

  El reino de las matemáticas.

- **Inducción** de un resultado trivial en un caso particular esperamos inferir la regla general.

Entrada	**Implicación**	**Ejemplo**
RESULTADO		El sol salía todos los días
CASO		Estamos en un planeta
	REGLA	En un planeta, su sol sale todos los días

  El reino de la ciencia.

- **Secuestro:** De una regla general y un resultado trivial esperamos inferir el caso particular

Entrada	Implicación	Ejemplo
REGLA		En un planeta, su sol sale todos los días
RESULTADO		El sol salía todos los días
	CASO	Estamos en un planeta

Más raramente utilizado.

1.2. ¿Que es el Machine Learning?

> "[Machine Learning] brinda a las computadoras la capacidad de aprender sin estar programada explicitamente."
>
> Samuel, 1959

Dicho de otra manera, el Machine Learning es un sub-campo de la Inteligencia Artificial en donde se buscan algoritmos "suaves"que, hasta cierto punto, pueden adaptarse a un cierto tipo de tarea en lugar de consistir simplemente en una lógica codificada. que busca manejar grandes cantidades de datos y realizar por ejemplo:

- Estructuración de datos
- Encontrar correlaciones
- Clasificar datos
- Reconocer patrones
- Comprender datos
- Tomar decisiones basadas en datos
- Adaptar tareas a datos
- Extrapolación/ predicción

1.2.1. Supervised Learning

Algoritmos "suaves"que infieren la tarea designada mediante la inspección de los datos de entrenamiento apropiados.

Ejemplos:

- Clasificación: Predicción de clases discretas, por ejemplo:

- El correo electrónico es spam o no
- Identificar imagenes borrosas

◦ Regresión: Predicción de parámetros continuos, por ejemplo:

- Consumo de energia de acuerdo con el comportamiento del usuario aprendido
- Predicción de una tendencia según una historia determinada.

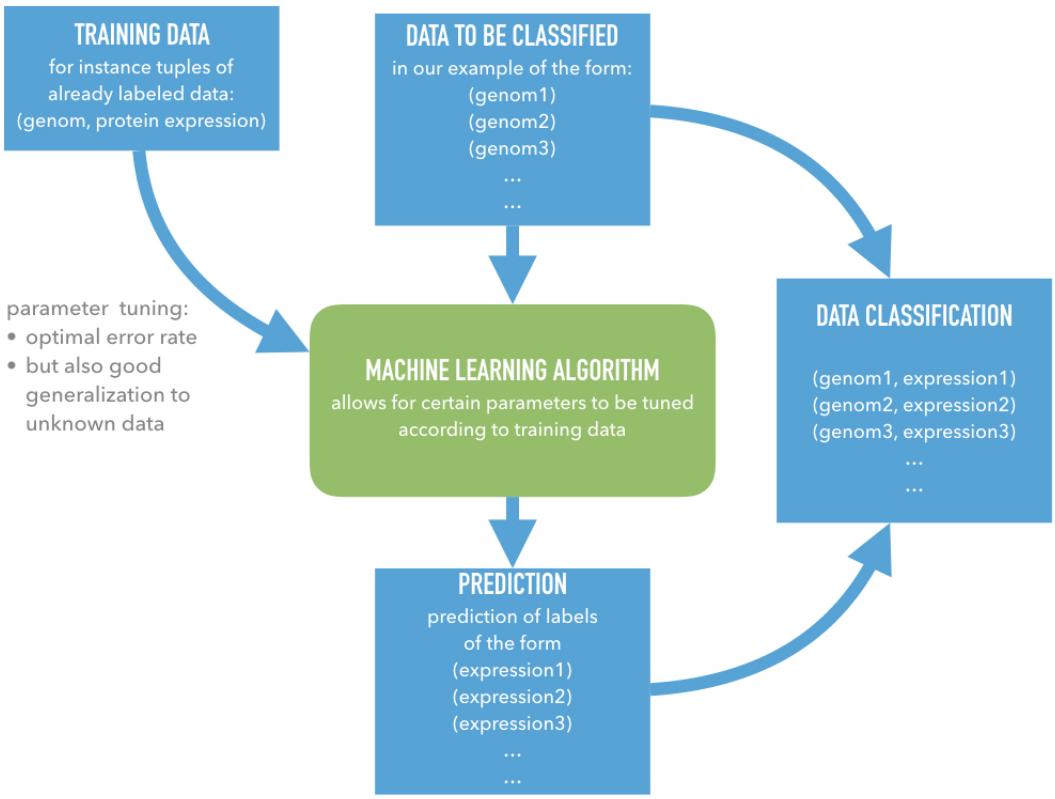

Figura 1.1: Algoritmo Machine Learning

1.2.2. Unsupervised Learning

Estructurar los datos en grupos sin conocimientos previos detallados.

Ejemplos:

◦ Wordnets: relaciones entre palabras de un leguaje natural.

◦ Referencias cruzadas entre documentos.

◦ Comprensión de datos y reducción de dimensionalidad.

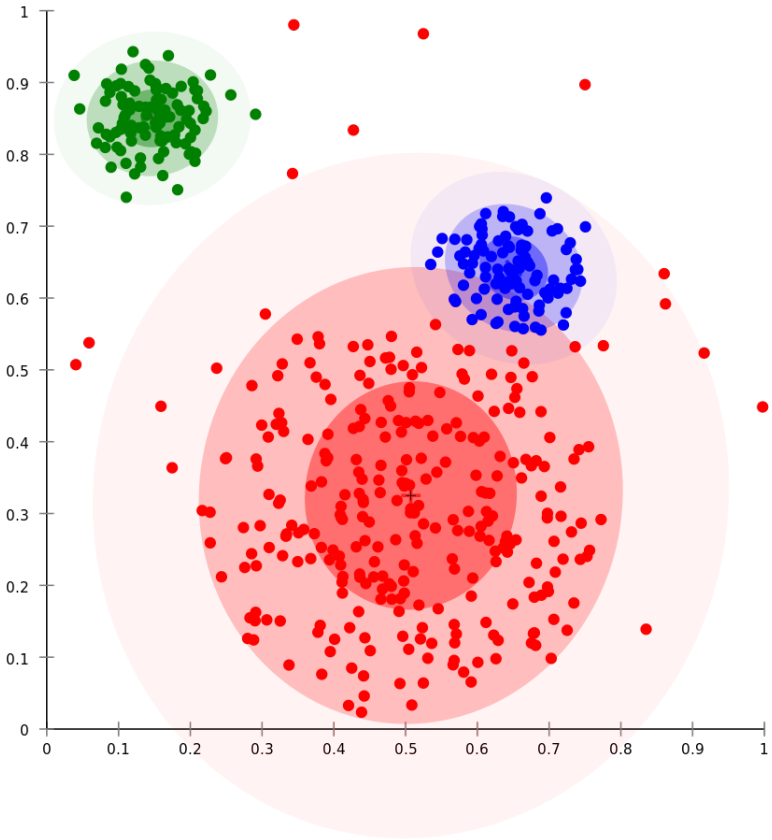

Figura 1.2: Ejemplo ficticio de puntos de datos 2D. El color indica una relación entre los puntos de datos. A partir de estas relaciones, las regiones sombreadas pueden inferirse mediante un algoritmo de aprendizaje automático no supervisado. Esto puede ser util cuando se buscan propiedades estructuradas burdas de un conjunto de datos

1.2.3. Reinforcement Learning

Un agente (Machine Learning program + artifact) aprende a realizar una determinada tarea mediante, por ejemplo, prueba y erros. El aprendizaje se ve facilitado por la capacidad de observar el entorno y recibir retroalimentación en función de las acciones.
Ejemplos:

- Movimiento de un robot en terreno desconocido o bajo condiciones variables
- Obteniendo puntueciones altas en juegos de Atari como *Google depmind*

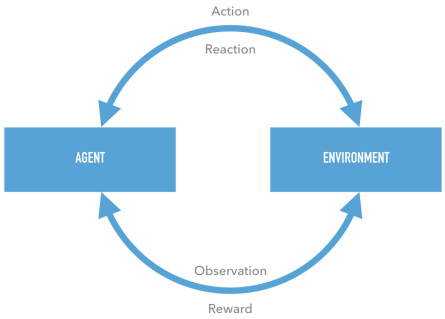

Figura 1.3: a

Parte I

Clases

2 | Neuronas y Redes Neuronales

En este curso estudiaremos objetos matemáticos llamados redes neuronales. Estos son elementos centrales en IA (Inteligencia Artificial).

La primera neurona en el contexto del cerebro fue propuesta por McCulloch y Pitts en 1943.

$$\mathbb{R}^d \ni \widehat{x} \to \chi_{\mathbb{R}^+}\left(\sum w_i x_i - \theta\right), \tag{2.1}$$

en donde:

- $d \in \mathbb{N}$.
- Para todo conjunto A.
$$\chi_A(y) = \begin{cases} 1 & \text{si } y \in A, \\ 0 & \text{si } y \notin A. \end{cases}$$
- $\mathbb{R}^+ = [0, \infty]$.
- $w_i \in \mathbb{R}, \forall i :=$ Pesos
- $\theta \in \mathbb{R}:=$ Umbral
- $\chi_{\mathbb{R}^+} :=$ Función de activación.
- $\widehat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$

Si la combinación lineal $\sum x_i w_i$ es mayor que θ, entonces la neurona se activa, de lo contrario no lo hace.

La función:

$$\widehat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \to \sum x_i w_i - \theta \tag{2.2}$$

es una transformación afín (una transformación lineal compuesta con una traslación) y la función $\chi_{\mathbb{R}^+}$ es una función no lineal.

De forma mas general, definimos una neurona como una transformación afin compuesta con una función no lineal, es decir:

Definición 2.0.1: Neurona

Sea una transformación afin $T : \mathbb{R}^m \to \mathbb{R}^n$ y una función no lineal $\rho : \mathbb{R} \to \mathbb{R}$. Definimos una **neurona** como la composición:

$$\mathbb{R}^m \ni \widehat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \to \rho(T(\widehat{x})) := \begin{pmatrix} \rho(T(\widehat{x})_1) \\ \vdots \\ \rho(T(\widehat{x})_n) \end{pmatrix} \tag{2.3}$$

De donde:

- $T(\widehat{x}) = \begin{pmatrix} T(\widehat{x})_1 \\ \vdots \\ T(\widehat{x})_n \end{pmatrix} \in \mathbb{R}^n$
- $T(\widehat{x})_i$ es la i-ésima componente de $T(\widehat{x})$

La función ρ es la función de activación y la transformación afin T contiene a los pesos y al umbral (de forma abstracta).

Una red neuronal es una composición de neuronas, en nuestro caso consideramos una única función de activación para todas las neuronas de la red.

Definición 2.0.2: Red Neuronal

Una red neuronal consiste en L transformaciones afines, $T_1, T_2, ..., T_L$ y una función de activación $\rho : \mathbb{R} \to \mathbb{R}$, definimos la red neuronal como la composición:

$$F = T_L \circ \rho \circ T_{L-1} \circ \rho \circ ... \circ \rho \circ T_1 \tag{2.4}$$

En donde estamos identificando a la función ρ con funciones $\rho : \mathbb{R}^n \to \mathbb{R}^n$ como:

$$\rho \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} := \begin{pmatrix} \rho(Y_1) \\ \vdots \\ \rho(Y_n) \end{pmatrix}$$

(Esto es un abuso de notación que siempre se usa)

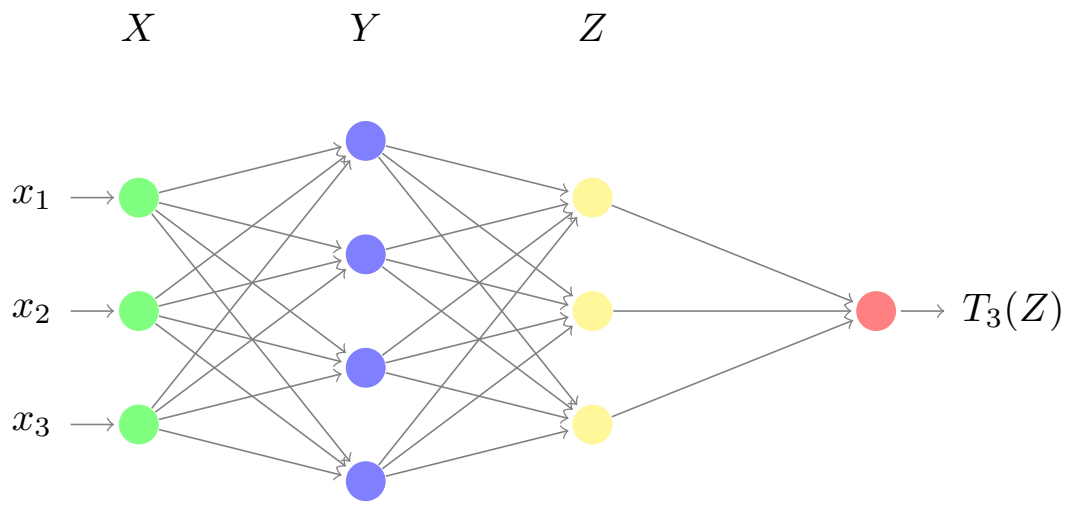

Figura 2.1: Ejemplo de Red Neuronal

En donde:

- $X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$
- $Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \rho(T_1(X))$
- $Z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \rho(T_2(Y))$

A la función F se le nombra un **"Multilayer Perceptron"(MLP)** de dimensión d, L niveles y función de activación ρ. donde el dominio de T_1 es $\mathbb{R}^d$.

3 | Resultados de Análisis Funcional

3.1. Espacios de Banach y Espacios de Hilbert

Definición 3.1.1: Norma

Sea V un espacio vectorial, una norma en V es una asignación $V \ni \widehat{x} \to \|\widehat{x}\|$, tal que:

1. $\|\widehat{x}\| \geq 0, \forall \widehat{x}$
2. $\|\widehat{x}\| = 0 \Leftrightarrow \widehat{x} = \widehat{0}$
3. $\|\alpha \cdot \widehat{x}\| = |\alpha| \|\widehat{x}\|, \forall \alpha \in \mathbb{R}$
4. $\|\widehat{x} + \widehat{y}\| \leq \|\widehat{x}\| + \|\widehat{y}\|$

Definición 3.1.2: Producto Interior

Sea V un espacio vectorial. un producto interior es una asignación $V \ni \widehat{x}, \widehat{y} \to \langle \widehat{x}, \widehat{y} \rangle$, tal que:

1. $\langle \widehat{u} + \widehat{v}, \widehat{w} \rangle = \langle \widehat{u}, \widehat{w} \rangle + \langle \widehat{v}, \widehat{w} \rangle$
2. $\langle \lambda \widehat{v}, \widehat{w} \rangle = \lambda \langle \widehat{v}, \widehat{w} \rangle, \forall \lambda \in \mathbb{R}$
3. $\langle \widehat{u}, \widehat{v} \rangle = \langle \widehat{v}, \widehat{u} \rangle$
4. $\langle \widehat{v}, \widehat{v} \rangle \geq 0, \langle \widehat{v}, \widehat{v} \rangle = 0 \Leftrightarrow \widehat{v} = 0$

Notación 3.1.1

Todo producto interior induce una norma de la forma:

$$\|\widehat{x}\| = \sqrt{\langle \widehat{x}, \widehat{x} \rangle}$$

Definición 3.1.3: Completitud

Decimos que un espacio vectorial normado V, es completo si toda sucesión de Cauchy converge, es decir:

Sea $(\widehat{v}_n)_{n\in\mathbb{N}}$ una sucesión en V tal que $\forall\epsilon > 0, \exists N \in \mathbb{N}$, tal que $\forall n, m \geq N, \|\widehat{v}_n - \widehat{v}_m\| < \epsilon$

Entonces existe:

$$\widehat{v} = \lim_{n\to\infty} \widehat{v}_n$$

Definición 3.1.4: Espacios de Banach y de Hilbert

- Un espacio de Banach es un espacio vectorial normado y completo.
- Un espacio de Hilbert es una espacio vectorial con producto interior y completo.

Notación 3.1.2

Sea V un espacio vectorial normado:

1. Bola: Sea $\widehat{x} \in V$ y $r > 0$, denotamos por $B_(\widehat{x}, r) := \{y \in V \mid \|\widehat{y} - \widehat{x}\| < r\}$.

2. Conjunto Abierto: Sea $A \subset V$. Decimos que A es abierto si $\forall \widehat{a} \in A, \exists r > 0$ tal que $B(\widehat{a}, r) \subset A$.

3. Conjunto Cerrado: $B \subset V$. Decimos que B es cerrado si B^c es abierto o equivalentemente si se cumple lo siguiente.

   Si $\widehat{y} \in V$ es tal que $\forall\epsilon > 0$ tal que $B(\widehat{y}, \epsilon) \cap B \neq \emptyset$, entonces $\widehat{y} \in B$.

4. Cerradura: Sea $C \subset V$ la cerradura es el conjunto cerrado mas chico que contiene a C. Denotamos con $\bar{C}$ a la cerradura de C. El conjunto $\bar{C}$ se caracteriza con la siguiente propiedad.

   $$\widehat{y} \in \overline{C} \Leftrightarrow \forall\epsilon > 0, B(\widehat{y}, \epsilon) \cap C \neq \emptyset.$$

   Notemos que C es cerrado $\Leftrightarrow C = \overline{C}$.

5. Interior: El interior de un conjunto A es el conjunto abierto mas grande contenido en A y lo denotamos con A°.

Definición 3.1.5: Densidad

Sea V un espacio vectorial normado y $M \subset V$. Decimos que M es denso si se cumple alguno de los siguientes puntos:

1. $\overline{M} = V$.
2. $\forall \widehat{v} \in V, \forall \epsilon > 0$ tal que $B(\widehat{v}, \epsilon) \cap M \neq \emptyset$.

Definición 3.1.6: Norma de una transformación lineal continua

Sean V y W espacios vectoriales normados y $\Lambda : V \to W$ una transformación lineal, definimos:

$$\|\Lambda\| := \sup\{\|\Lambda(\widehat{v})\| \mid \widehat{v} \in V \ y \ \|\widehat{v}\| \leq 1\}$$

Nota: Λ es continua $\Leftrightarrow \sup\{\|\Lambda(\widehat{v})\| \mid \widehat{v} \in V \ y \ \|\widehat{v}\| \leq 1\} < \infty$

Teorema 3.1.1: Hahn-Banach

Sea V un espacio vectorial normado y $M \subset V$ un sub-espacio vectorial, supongamos $h : M \to \mathbb{R}$ es continua y lineal. entonces existe una extensión $g : V \to \mathbb{R}$ (lineal y continua) tal que:

1. $g(m) = h(m), \ \forall m \in M$
2. $\|h\| = \|g\|$

Demostración. Se deja para el lector. □

Corolario 3.1.1

Sea V un espacio vectorial normado y $M \subset V$ un sub-espacio vectorial, suponemos que M no es denso, (i.e.$\overline{M} \neq V$, entonces existe una función lineal $g : V \to \mathbb{R}$ continua tal que:

1. $g|_{\overline{M}} = 0$
2. $g \neq 0$

Demostración. Como $\overline{M} \neq V$, entonces existe $\widehat{x} \in V \setminus \overline{M}$, sea W:= Espacio vectorial generado por $\overline{M}$ y X, entonces todo elemento de W se puede escribir de manera única de la siguiente manera:

$$W \ni \widehat{w} = \alpha\widehat{x} + \widehat{m}$$

en donde $\alpha \in \mathbb{R}$, m$\in \overline{M}$, es decir:

$$W = \{\alpha\widehat{x} + \widehat{m} \mid \alpha \in \mathbb{R}, \widehat{m} \in \overline{M}\}$$

Veamos que los elementos de W se pueden escribir de forma única como se describe antes:

Supongamos que existen $\alpha_1, \alpha_2, \widehat{m}_1, \widehat{m}_2$:

$\Rightarrow \widehat{w} = \alpha_1\widehat{x} + \widehat{m}_1 = \alpha_2\widehat{x} + \widehat{m}_2$
$\Rightarrow (\alpha_1 - \alpha_2)\widehat{x} = \widehat{m}_2 - \widehat{m}_1 \in \overline{M}$

(Pues como M es espacio vectorial, entonces $\overline{M}$, también lo es). Concluimos que $(\alpha_1, \alpha_2)\widehat{x} \in \overline{M}$

pues si $\alpha_1 - \alpha_2 \neq 0$, entonces tendriamos que existe $\widehat{y} \in \overline{M}$ tal que $\widehat{x} = (\frac{1}{\alpha_1 - \alpha_2})\widehat{y} \in \overline{M}$, lo que contradice que $\widehat{x} \notin \overline{M}$

Entonces $\alpha_1 - \alpha_1 = 0$ por lo que $\alpha_1 = \alpha_2$ como $(\alpha_1 - \alpha_2)\widehat{x} = m_2 - m_1$, concluimos que $m_1 = m_2$

Definimos a $h : W \to \mathbb{R}$ como $h(\alpha\widehat{x} + \widehat{m}) := \alpha\|x\|$, la cual es lineal, pues:

$h(\lambda(\alpha_1\widehat{x} + \widehat{m}) + \alpha_2\widehat{x} + \widehat{m}_2) = h(\lambda\alpha_1 + \alpha_2)\widehat{x} + \lambda\widehat{m}_1 + \widehat{m}_2$ (en donde $\lambda\widehat{m}_1 + \widehat{m}_2 \in \overline{M}$)
$= (\lambda\alpha_1 + \alpha_2)\|x\| = \lambda h(\alpha_1\widehat{x} + \widehat{m}_1) + h(\alpha_2\widehat{x} + \widehat{m}_2)$

Para toda $\lambda \in \mathbb{R}, \alpha_1, \alpha_2 \in \mathbb{R}$, $\widehat{m}_1, \widehat{m}_2 \in \overline{M}$, Además $h(m) = 0$, $\forall m \in \overline{M}$, Veamos que $h : W \to \mathbb{R}$ es continua para lo cual verificamos que:

$$\|h\| = \sup\{|h(\widehat{w}) \mid \widehat{w} \in W, \|\widehat{w}\| \leq 1|\} \leq \infty$$

Como $\widehat{x} \in V \setminus \overline{M}$:

$\Rightarrow \exists\epsilon > 0$ tal que $B(\widehat{x}, \epsilon) \cap \overline{M}$.
$\Rightarrow \exists\epsilon > 0$ tal que $B(\widehat{x}, \epsilon) \cap \overline{M} = \emptyset$

Sea $\widehat{x} = \alpha\widehat{x} + \widehat{m} \in W$ tal que $\|\widehat{w}\| \leq 1$, tenemos que:

$$\|\alpha\widehat{x} + \widehat{m}\| = |\alpha|\|\widehat{x} + \tfrac{1}{\alpha}\widehat{m}\| \leq 1$$

Además como $\frac{1}{\alpha}\widehat{m} \in \overline{M}$, se obtiene que:

$$\|\widehat{x} - (-\tfrac{1}{\alpha}\widehat{m})\| = \|\widehat{x} + \tfrac{1}{\alpha}\widehat{m}\| \geq \epsilon$$

Concluimos que:

$$|\alpha|\|\widehat{x}+\tfrac{1}{\alpha}\widehat{m}\|\leq 1 \Rightarrow |\alpha|\leq \frac{1}{\|\widehat{x}+\frac{1}{\alpha}\widehat{m}\|}\leq \frac{1}{\epsilon}$$

Obtenemos que:

$$|\alpha|\leq \frac{1}{\epsilon}, \forall \widehat{w}=\alpha\widehat{x}+\widehat{m}\in W \text{ tal que } \|\widehat{w}\|\leq 1$$

De esto obtenemos que:

$$|h(\widehat{w})|=|\alpha|\|\widehat{x}\|\leq \frac{1}{\epsilon}\|\widehat{x}\|$$

$\forall \widehat{w}$, con $\|\widehat{w}\|\leq 1$, con $w=\alpha\widehat{x}+\widehat{m}$

Conluimos que $sup\{\|h(\widehat{w})\| \mid \widehat{w}\in W, |\widehat{w}|\leq 1\}=|h|\leq \frac{1}{\epsilon}\|\widehat{x}\|$ y por lo tanto es continua.

Obtenemos entonces que h es una función continua tal que $h\neq 0$, puesto que $h(\widehat{x})=\|x\|$ y además $h(\widehat{m})=0$, puesto que $h(\widehat{x})=\|\widehat{x}\|$ y además $h(\widehat{m})=0$, $\forall \widehat{m}\in \overline{M}$

Por el Teorema de Hahn-Banach, existe g lineal y continua tal que:

- $g(\widehat{w})=h(\widehat{w}), \forall \widehat{w}\in W$
- $\|g\|=\|h\|$, (en particular g es continua)

g es la función que buscamos, pues:

- $g\neq 0$ (porque $h\neq 0$)
- $g(\widehat{m})=h(\widehat{m})=0, \forall \widehat{m}\in \overline{M}$

□

3.2. El Dual de las Funciones Continuas e Integrables

Definición 3.2.1: Dual de un espacio vectorial

Sea V un espacio vectorial normado, Denotamos por $V^\bullet$ al conjunto de funciones lineales y continuas de V en $\mathbb{R}$

Definición 3.2.2

Sea $K \subset \mathbb{R}$ un conjunto compacto (Cerrado y Acotado), denotamos por:

$$C(K) := \{f : K \to \mathbb{R} : \text{f es continua } \}$$

Al espacio vectorial de las funciones continuas en K con valores en $\mathbb{R}$, denotado de la norma.

$$\|f\|_\infty := sup\{|f(\widehat{x}| \ : \ \widehat{x} \in K\}$$

Recordemos que $C(k)^\bullet$ es el dual de C(K) el cual es el conjunto de funciones lineales y continuas de C(K) en $\mathbb{R}$ y que la función lineal $h : C(K) \to \mathbb{R}$ es continua $\Leftrightarrow \|h\| := sup\{|h(\widehat{x}| : \widehat{x} \in C(K)\} < \infty$

Definición 3.2.3: Teorema de Representación de Riez

Decimos que una función f acotada, $f : K \to \mathbb{R}$ es integrable respecto a $h \in C(K)$, si existe una sucesión de funciones $\{f_n\}_{n\in\mathbb{N}}$ en C(K) tal que:

1. Existe $c \in \mathbb{R}$ tal que:
$$|f_n(\widehat{x})| \leq c \ \forall \widehat{x} \ , \forall n$$
2. $$\lim_{n\to\infty} f_n(\widehat{x}) = f(\widehat{x}), \ \forall \widehat{x} \in K$$

Notación 3.2.1

Denotaremos por $L_a^1(h)$ al espacio de funciones integrables y acotadas con respecto a h.

Notemos que en está definición $L_a^1(h)$ no depende de h, pero mantenemos h en nuestra notación porque se usa adelante.

Teorema 3.2.1: Convergencia Dominada de Lebesgue 1

Sea $f \in L_a^1(h)$ para algún $h \in C(K)^\bullet$ y $\{f_n\}_{n\in\mathbb{N}}$ como en 3.2.3, entonces $\lim_{n\to\infty} h(f_n)$, existe y si $\{g_n\}_{n\in\mathbb{N}}$ es otra sucesión que cumple 1 y 2 de la Def 3, entonces:

$$\lim_{n\to} h(f_n) = \lim_{n\to\infty} h(g_n)$$

Es fácil ver que $L_a^1(h)$ es un espacio vectorial y que la función $f \to \int f \, dh$ es lineal

Teorema 3.2.2: Convergencia Dominada de Lebesgue 2

Sea $h \in C(K)^\bullet$, supongamos que $g \in L_a^1(h)$ es tal que existe una $\{g_n\}_{n\in\mathbb{N}}$ en $L_a^1(h)$, tal que:

1. Existe $G \in L_a^1(h)$ tal que:

$$|g_n(\widehat{x})| \leq G(\widehat{x}),\ \forall \widehat{x} \in K,\ \forall n \in \mathbb{N}$$

2. $\lim_{n\to\infty} g_n(\widehat{x}) = g(\widehat{x}),\ \forall \widehat{x} \in K$

Entonces:

$$\lim_{n\to\infty} \int g_n\, dh = \int g\, dh$$

Definición 3.2.4

Sea $K \subset \mathbb{R}^d$ compacto. Dado $x \in \mathbb{R}^d$ y $a \in \mathbb{R}^d$, denotamos por:

$$\langle \widehat{a}, \widehat{x} \rangle = \sum_{i=1}^{d} \widehat{a}_i \widehat{x}_i \tag{3.1}$$

En donde:

$$\widehat{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_d \end{pmatrix}, \widehat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

Lema 3.2.1

Dados $b_1, b_2 \in \mathbb{R}$ con $b_1 < b_2$ y $f \in C(K)$, se cumple que:

$$X_{[b_1,b_2]} \circ f \in L_a^1(h)$$

Notemos que $f : K \to \mathbb{R}, K \subset \mathbb{R}^d$ y $X_{[b_1,b_2]} : \mathbb{R} \to \mathbb{R}$ es tal que:

$$X_{[b_1,b_2]}(y) = \begin{cases} 1 & y \in [b_1, b_2] \\ 0 & y \notin [b_1, b_2] \end{cases}$$

Teorema 3.2.3

Sea $h \in C(K)^\bullet$, dado $a \in \mathbb{R}$ y f_a la función dada por $f_a = \langle a, x \rangle$, $\forall \widehat{x} \in \mathbb{R}^d$, si:

$$\int X_{[b_1,b_2]} \circ f_a\, dh = 0$$

$\forall a \in \mathbb{R}^d,\ \forall b_1, b_2 \in \mathbb{R}$ con b_1, b_2 entonces h=0

4 | Aproximación por medio de Redes Neuronales

4.1. Universalidad

Definición 4.1.1: MLP

Sea $\rho : \mathbb{R} \to \mathbb{R}$ continua, d,L$\in \mathbb{N}$ y $K \subset \mathbb{R}^d$ compacto. Denotamos por $MLP((\rho, d, L))$ al conjunto de MLP's con función de activación ρ, L niveles, dimensión d y tales que T_L toma valores en $\mathbb{R}$, es decir $\forall f \in MLP(\rho, d, L)$ tal que $f : \mathbb{R}^d \to \mathbb{R}$

Definición 4.1.2: Universalidad

Sea K un subconjunto compacto de $\mathbb{R}^d$ Decimos que $MLP(\rho, d, L)$ es universal si las restricciones de las funciones $f \in MLP(\rho, d, L)$ son densas en C(K), para toda K.

Normalmente se identifican las funciones $f : \mathbb{R}^d \to \mathbb{R}$ con sus restricciones a K, $f|_K \equiv f : K \to \mathbb{R}$

Una manera de interpretar que el conjunto de MLP'S sea denso en el conjunto de funciones continuas, es que para cada función continua, existe un MLP que lo aproxima.

Definición 4.1.3: Discriminante

Sea $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compacto. Una función $f : \mathbb{R} \to \mathbb{R}$ continua es llamada discriminante si dada $h \in C(K)^\bullet$:

$$h(t_{a,b}(f)) = 0 \ \forall \widehat{a} \in \mathbb{R}^d, b \in \mathbb{R} \Rightarrow h = 0$$

En donde denotamos por $t_{a,b}(f) \in C(K)$ a la función dada por $t_{a,b}(f)(\widehat{x}) = f(\langle \widehat{a}, \widehat{x} \rangle - b)$

Proposición 4.1.1

$MLP(\rho, d, 2)$ es un espacio vecotrial

Demostración. Sean $F, G \in MLP(\rho, d, 2)$ y $\alpha \in \mathbb{R}$, veamos que:

$$\alpha F + G \in MLP(\rho, d, 2)$$

Como $F, G \in MLP(\rho, d, 2)$, entonces existen transformaciones afines T_1, T_2, V_1, V_2 tales que:

- $F = T_2 \circ \rho \circ T_1$
- $G = V_2 \circ \rho \circ V_2$

En donde T_1 *y* V_1 tiene dominio $\mathbb{R}^d$ y T_2, V_2 toman valores en $\mathbb{R}$ definimos:

$$U_1(\widehat{x}) = \begin{pmatrix} T_1(\widehat{x}) \\ V_1(\widehat{x}) \end{pmatrix}, \forall \widehat{x} \in \mathbb{R}$$

Entonces U_1 es una transformación afín.

Dados y, z con $y \in Dom\, T_2$, $z \in Dom\, V_2$ definimos:

$$U_2 \begin{pmatrix} y \\ z \end{pmatrix} = \alpha T_2(y) + V_2(z)$$

Entonces tenemos que:

$$\begin{aligned} U_2 \circ \rho \circ U_1(\widehat{x}) &= U_2\ \rho \begin{pmatrix} T_1(\widehat{x}) \\ V_1(\widehat{x}) \end{pmatrix} \\ &= \alpha T_2(\rho(T_1(\widehat{x})) + V_2(\rho(V_1(\widehat{x}))) \\ &= \alpha F(\widehat{x}) + G(\widehat{x}) \end{aligned}$$

Entonces $\alpha F + G = U_2 \circ \rho \circ U_1 \in MLP(\rho, d, 2)$, por lo que $MLP(\rho, d, 2)$ es un espacio vectorial. □

Teorema 4.1.1: Aproximación Universal, Gybenko 1989

Sea $d \in \mathbb{N}, K \subset \mathbb{R}^d$ compacto y $\rho : \mathbb{R} \to \mathbb{R}$ una función discriminante. Entonces $MLP(\rho, d, 2)$ es denso en C(K)

Demostración. Procederemos por contradicción:

Supongamos que $MLP(\rho, d, 2)$ no es denso, entonces $\overline{MLP(\rho, d, 2)} \subsetneq C(K)$ por lo que existe $f \in C(K) \setminus \overline{MLP(\rho, d, 2)}$ por el Corolario 2.1.1. existe una función $h \in C(K)^\bullet$ tal que:

- $h \neq 0$

- $h|_{\overline{MLP(\rho,d,2)}}$

Sea $T_{\widehat{a},b}$ la función afín dada por $T_{\widehat{a},b} : \mathbb{R}^d \to \mathbb{R},\ T_{\widehat{a},b} = \langle \widehat{a}, \widehat{x} \rangle - b$, en donde $a \in \mathbb{R}^d,\ b \in \mathbb{R}$, entonces:

$$\rho \circ T_{\widehat{a},b} = Id \circ \rho \circ T_{\widehat{a},b} \in MLP(\rho, d, 2)$$

En donde $Id : \mathbb{R} \to \mathbb{R}$ es la identidad

Como $h|_{\overline{MLP(\rho,d,2)}} = 0$, concluimos que:

$$h(\rho \circ T_{\widehat{a},b}) = 0,\ \forall \widehat{a} \in \mathbb{R}^d, \forall b \in \mathbb{R}$$

Notemos que:

$$\rho \circ T_{\widehat{a},b}(\widehat{x}) = \rho(\langle \widehat{a}, \widehat{x} \rangle - b) = t_{\widehat{a},b}(\rho)(\widehat{x})$$

Entonces tenemos que:

$$h(t_{\widehat{a},b}(\rho)) = 0,\ \forall \widehat{a} \in \mathbb{R}^d,\ \forall b \in \mathbb{R}$$

Como suponemos que ρ es discriminante, esto implica que $h = 0$

Esto es una contradicción, pues $h \neq 0$ la contradicción se obtiene por suponer que $MLP(\rho, d, 2)$ no es denso por lo que **concluimos que *MLP*$(\rho, d, 2)$ es denso en C(K)** □

Definición 4.1.4: Sigmoidal

Una función $f : \mathbb{R} \to \mathbb{R}$ continua tal que:

$$\lim_{r \to \infty} f(r) = 1,\ \lim_{r \to -\infty} f(r) = 0$$

Se llama sigmoidal

Proposición 4.1.2

Sea $d \in \mathbb{N}, K \subset \mathbb{R}^d$ compacto, entonces toda función sigmoidal $\rho : \mathbb{R} \to \mathbb{R}$ es discriminante (y por lo tanto $\overline{MLP(\rho, d, 2)}$ es denso en C(K), por el Teorema 3.1.1)

Demostración. Sea ρ sigmoidal, entonces dados $\lambda, b, \theta \in \mathbb{R},\ \widehat{a}, \widehat{x} \in \mathbb{R}^d$, tenemos que:

$$lim_{\lambda \to \infty} \rho(\lambda(\langle \widehat{a}, \widehat{x} \rangle - b) + \theta) = \begin{cases} 1, \text{ si } \langle \widehat{a}, \widehat{x} \rangle - b > 0 \\ \rho(\theta), \text{ si } \langle \widehat{a}, \widehat{x} \rangle - b = 0 \\ 0, \text{ si } \langle \widehat{a}, \widehat{x} \rangle - b < 0 \end{cases}$$

Sea $h \in C(K)^\bullet$ y sean:

$$H_{\widehat{a},b,>} := \{\widehat{x} \in K \ : \ \langle \widehat{a}, \widehat{x} \rangle - b > 0\}$$
$$H_{\widehat{a},b,=} := \{\widehat{x} \in K \ : \ \langle \widehat{a}, \widehat{x} \rangle - b = 0\}$$
$$H_{\widehat{a},b,<} := \{\widehat{x} \in K \ : \ \langle \widehat{a}, \widehat{x} \rangle - b < 0\}$$

Sea $g_{\lambda,\widehat{a},b,\theta}(\widehat{x}) = \rho(\lambda(\langle \widehat{a}, \widehat{x} \rangle - b + \theta))$, entonces:

$$lim_{\lambda \to \infty} g_{\lambda,\widehat{a},b,\theta}(\widehat{x}) = X_{H\widehat{a},b,>}(\widehat{x}) + \rho(\theta) X_{H\widehat{a},b,=}(\widehat{x})$$

Ahora veamos que ρ es discriminadora, es decir si $h \in C(K)^\bullet$ tal que $h(t_{\widehat{a},b}(\rho)) = 0, \forall \widehat{a}, \ \forall b$, únicamente tenemos que ver h=0.

Notemos que:

$$g_{\lambda,\widehat{a},b,\theta} = t_{\lambda\widehat{a},\lambda b - \theta}(\rho)$$

Entonces $h(g_{\lambda,\widehat{a},b,\theta}) = 0, \ \forall \lambda, \ \forall \widehat{a}, \forall b, \forall \theta$

Por el teorema de convergencia dominada de Lebesgue 2, obtenemos que:

$$\begin{aligned} 0 = lim_{\lambda \to \infty} h(g_{\lambda,\widehat{a},b,\theta}) &= \int lim_{\lambda \to \infty} g_{\lambda,\widehat{a},b,\theta} \, dh \\ &= \int (X_{H\widehat{a},b,>} + \rho(\theta) X_{H\widehat{a},b,=}) \, dh \\ &= \int X_{H\widehat{a},b,>} \, dh + \rho(\theta) \int X_{H\widehat{a},b,=} \, dh \end{aligned}$$

Obtenemos que:

$$\int X_{H\widehat{a},b,>} \, dh + \rho(\theta) \int X_{H\widehat{a},b,=} \, dh = 0, \ \forall \widehat{a}, \forall b, \forall \theta$$

1. Haciendo θ tender a ∞ obtenemos:

$$\int X_{H\widehat{a},b,>} \, dh + \int X_{H\widehat{a},b,=} \, dh = 0$$

2. Haciendo θ tender a $-\infty$ obtenemos:

$$\int X_{H\widehat{a},b,>} \, dh = 0$$

Sea $f_a(\widehat{x}) := \langle \widehat{a}, \widehat{x} \rangle, \ \forall \widehat{x} \in K$

$$H_{a,b,>} = \{\widehat{x} \in K \ : \ \langle \widehat{a}, \widehat{x} \rangle - b > 0\} = f_a^{-1}((b, \infty))$$
$$H_{\widehat{a},b,=} = f^{-1}(b)$$
$$X_{H\widehat{a},b,>} + X_{H\widehat{a},b,=} = X_{f_a^{-1}([b,\infty))}$$

Dados $b_1, b_2 \in \mathbb{R}$ con $b_1 < b_2$, obtenemos que:

$$X_{H\widehat{a},b_1,>} + X_{H\widehat{a},b,=} - X_{H\widehat{a},b_2,>} = X_{f_a^{-1}([b_1,b_2])}$$

de (1) y (2) obtenemos que:

$$\int X_{f_a^{-1}([b_1,b_2])}\, dh = 0,\ \forall b_1 < b_2$$

Notemos que:

$$X_{f_a^{-1}}([b_1, b_2]) = X_{[b_1,b_2]} \circ f_a$$

Entonces:

$$\int X_{[b_1,b_2]} \circ f_a\, dh = 0, \forall\ b_1 < b_2 \in \mathbb{R}, \forall \widehat{a} \in \mathbb{R}^d$$

Usamos el Teorema 2.2.3, para concluir que h=0, por lo tanto ρ es discriminante. □

4.2. Redes Neuronales

Recordemos que las transformaciones afines consisten de transformaciones afines consisten de transformaciones lineales compuestas con traslaciones. Dada una transformación afín T, existe una matriz A y vector $\widehat{b}$ tales que:

$$T(\widehat{x}) = A\widehat{x} + \widehat{b}$$

De manera que identificamos:

$$T \equiv (A, \widehat{b})$$

Dada un red neuronal.

$$F = T_L \circ \rho \circ T_{L-1} \circ \rho \circ \cdots \circ \rho \circ T_1$$

en done $T_1, T_2, \cdots, T_L$ son afines, identificamos:

$$F \equiv (T_1, T_2, \cdots, T_L)$$

Regularmente a $(T_1, T_2, \cdots, T_L)$ se le nombra red neuronal y a F se le lla,a su realización, se denota en general:

$$\phi = (T_1, T_2, \cdots, T_L)$$

La red neuronal o bien abreviada con **NN (Neural Network)** para denotar la realización, utilizamos:

$$F = R_\rho(\phi) \equiv R(\phi), \text{ (R se refiere a realización)}$$

Otra notación que se usa es: dado $\widehat{x} \in \mathbb{R}^d$ ponemos:

$$x_0 = \widehat{x},\ x_1 = \rho \circ T_1 \circ x_0, \cdots, x_p = \rho \circ T_p \circ x_{p-1}, \cdots, x_L = T_L \circ x_{L-1}$$

Definición 4.2.1

Supongamos que:

$$T_p : \mathbb{R}^{N_{p-1}} \to \mathbb{R}^{N_p},\ N_0 = d$$

Definimos los siguientes conceptos:

1. Dimensión de entrada:=d
2. Número de niveles o capas:=$L = L(\phi)$
3. Número de neuronas:= $d + \sum_{j=1}^{L} N_j$
4. Dada una matriz A y un vector $\widehat{b}$, denotamos por:

$$\|A\|_0 := \text{Número de entradas no cero de A.}$$
$$\|\widehat{b}\|_0 := \text{Número de entradas no cero de } \widehat{b}$$

5. Dada una transformación afín $T \equiv (A, \widehat{b})$, definimos $\|T\|_0 := \|A\|_0 + \|\widehat{b}\|_0$
6. Denotamos por:

$$M_j(\phi) := \|T_i\|_0$$

   En donde $\Phi = (T_1, T_2, \cdots, T_L)$

7. Denotamos por:

$$M(\phi) = \sum_{j=1}^{L} M_j(\phi)$$

4.3. Operaciones Básicas con Redes Neuronales

Sean ϕ_1, ϕ_2 redes neuronales

$$\phi_1 = (T_1, T_2, \cdots, T_L)$$
$$\phi_2 = (V_1, V_2, \cdots, V_m)$$

Definimos:

$$\phi_1 \bullet \phi_2 := (V_1, V_2, \cdots, V_{m-1}, T_1 \circ V_m, T_2, \cdots, T_L)$$

La cual es una red neuronal de L+m-1 niveles, es directo ver que:

$$R(\phi_1 \bullet \phi_2) = \phi_1 \circ \phi_2$$

Definición 4.3.1

Dados T y V transformaciones afines, definimos:

1. Si T y V actúan en el mismo espacio, definimos:

   $$\begin{pmatrix} \mathrm{T} \\ \mathrm{V} \end{pmatrix}(\widehat{x}) = \begin{pmatrix} T(\widehat{x}) \\ V(\widehat{x}) \end{pmatrix}$$

   $\begin{pmatrix} \mathrm{T} \\ \mathrm{V} \end{pmatrix}$ es una transformación afín.

2. $T \oplus V \begin{pmatrix} \widehat{x} \\ \widehat{y} \end{pmatrix} = \begin{pmatrix} T(\widehat{x}) \\ V(\widehat{y}) \end{pmatrix}$

   $T \oplus V$ es una transformación afín.

Definición 4.3.2

Dadas redes neuronales

$$\phi^1 = (T_1, T_2, \cdots, T_L)$$
$$\phi^2 = (V_1, V_2, \cdots, V_L)$$

Definimos:

1. $P(\phi^1, \phi^2) = \left(\begin{pmatrix} T_1 \\ V_1 \end{pmatrix}, T_2 \oplus V_2, \cdots, T_L \oplus V_L \right)$

   En el caso en el que las dimensiones de entrada de ϕ^1 y ϕ^2 sean las mismas.

2. $FP(\phi^1, \phi^2) = (T_1 \oplus V_1, T_2 \oplus V_2, \cdots, T_L \oplus V_L)$

   $P(\phi^1, \phi^2)$ se le llama **Paralelización con entrada compartida** de ϕ^1, ϕ^2
   $FP(\phi^1, \phi^2)$ se le llama **Paralelización con entradas no compartidas** de ϕ^1, ϕ^2

Proposición 4.3.1

1. $M(P(\phi^1,\phi^2)) = M(FP(\phi^1,\phi^2)) = M(\phi^1) + M(\phi^2)$
2. $R(P(\phi^1,\phi^2))(\widehat{x}) = \begin{pmatrix} R(\phi^1(\widehat{x})) \\ R(\phi^2(\widehat{x})) \end{pmatrix},\ \forall \widehat{x} \in \mathbb{R}^d$

Proposición 4.3.2

Sea $d \in \mathbb{N}$ y $K \subseteq \mathbb{R}^d$ compacto, supongammos que $\rho : \mathbb{R} \to \mathbb{R}$ es no constante y diferenciable en un abierto. Entonces $\forall \epsilon > 0$ existe una red neuronal ϕ tal que:

$$\phi = (T_1, T_2)$$

con $T_1, T_2 : \mathbb{R}^d \to \mathbb{R}^d,\ M(\phi) \leq 4d,\ y\ |R(\phi)(\widehat{x}) - \widehat{x}| < \epsilon,\ \forall \widehat{x} \in K$

Demostración. Sea d=1, $x^\bullet \in \mathbb{R}$ tal que ρ es diferenciable en una vecindad de $x^\bullet$ y suponemos que:

$$\rho'(x^\bullet) = \theta \neq 0$$

Dada $\lambda > 0$ definimos:

$$T_1(x) = \frac{1}{\lambda}x + x^\bullet$$
$$T_2(x) = \frac{\lambda}{\theta}x - \lambda\rho\frac{(x^\bullet)}{\theta}$$

Entonces obtenemos:

$$\underbrace{|\ R(\Phi)(x) - x\ |}_{T_2(\rho \circ T_1(x))} = |\frac{\lambda}{\theta}(\overbrace{\ \rho(\tfrac{x}{\lambda} + x^\bullet\ )}^{T_1(x)}) - \rho(x^\bullet)) - x|$$

Es claro que $|R(\phi)(0) - 0| = 0$ y para $x \neq 0$, tomamos a $x \in K$ fijo:

$$|R(\phi)(x) - x| = \frac{|x|}{|\theta|}|\frac{\rho(\frac{x}{\lambda} + x^\bullet) - \rho(x^\bullet) - \theta}{\frac{x}{\lambda}}|$$
$$= \frac{|x|}{\theta}|\frac{\rho(x^\bullet + h_\lambda) - \rho(x^\bullet)}{h_\lambda} - \rho'(x^\bullet)|$$

Obtenemos:

$$|R(\phi)(x) - x| = \frac{|x|}{|\theta|}|\frac{\rho(x^\bullet + h_\lambda) - \rho(x^\bullet)}{h_\lambda} - \rho'(x^\bullet)|$$

Sea $\alpha = max\{|x| \ :\ x \in K\}$, como ρ es diferenciable en un intervalo y que contiene a $x^\bullet$ entonces dada $\epsilon > 0$ existe $\delta > 0$ tal que si $|y| < d$, entonces:

$$|\frac{\rho(x^\bullet + y) - \rho(x^\bullet)}{y} - \rho'(x^\bullet)| < \frac{|\theta|}{M}\epsilon$$

Como $h_\lambda = \frac{x}{\lambda}$, tomamos λ tal que $\frac{\alpha}{\lambda} < \delta$, entonces:

$$|h_\lambda| = \frac{|x|}{\lambda} < \delta$$

En este caso obtenemos que $\forall x \in K$

$$|R(\phi)(x) - x| = \frac{|x|}{|\theta|}|\frac{\rho(x^\bullet + y) - \rho(x^\bullet)}{y} - \rho'(x^\bullet)| < \frac{|\theta|}{M}\frac{|x|}{|\theta|}\epsilon \leq \epsilon$$

Concluimos que:

$$|R(\phi)(x) - x| \leq \epsilon, \quad \forall x \in K$$

□

Teorema 4.3.1: Universalidad, Caso $L \geq 2$

Sea $d \in \mathbb{N}, K \subset \mathbb{R}^d$ compacto y $\rho : \mathbb{R} \to \mathbb{R}$ una función discriminante tal que es diferenciable y no constante en un abierto. Entonces $MLP(\rho, d, L)$ es universal para $L \in \mathbb{N},\ L \geq 2$, es decir, $MLP(\rho, d, L)$ es denso en C(K)

Demostración. Tomamos $f \in C(K)$, encontraremos una sucesión de redes neuronales $(\phi_n)_{n\in\mathbb{N}}$ tal que:

$$\lim_{n\to\infty} \phi_n = f$$

con $\phi_n \in MLP(\rho, d, L),\ \forall n$

Por el teorema de universalidad para el caso L=2, existe φ_1 red neuronal con profundidad 2 tal que $R(\varphi^n) \in MLP(\rho, d, 2)$, satisface:

$$\|R(\varphi^n) - f\| < \frac{1}{n},\ \forall n \in \mathbb{N} \tag{4.1}$$

Como f es continua, entonces f(k) es compacto (Cerrado y Acotado).

Sea $\tilde{k} = \overline{B(0; r)} \subseteq \mathbb{R}$ tal que $r > L$ y $f(k) \subseteq B(0; r - L)$, usamos la proposición 2 de este capítulo, para encontrar redes neuronales ψ^n tales que:

$$|R(\psi^n)(y) - y| < \frac{1}{n},\ \forall y \in \tilde{K} \subseteq \mathbb{R} \tag{4.2}$$

y la profundidad de $R(\psi^n)$ es 2

Definimos:

$$\phi_n = \underbrace{\psi^n \bullet \psi^n \bullet \cdots \bullet \psi^n \bullet \varphi^n}_{L-2\ veces}$$

Entonces ϕ_n tiene profundidad L.

Recordemos que:

$$R(\phi_n) = R(\psi^n) \circ R(\psi^n) \circ \cdots \circ R(\psi^n) \circ R(\varphi^n)$$

Calculamos:

$$\begin{aligned} \mid\ R(\phi_n)(x) - f(x)\ \mid \leq \mid\ & R(\psi^n[R(\psi^n) \circ \cdots \circ R(\varphi^n)(x)] - \overbrace{R(\psi^n) \circ \cdots \circ R(\varphi^n)(x)\ )}^{n-1\,veces} \mid \\ & + \mid R(\psi^n)[\overbrace{\ R(\psi^n) \circ \cdots \circ R(\varphi^n)(x)\ ]}^{n-2\,veces} - \overbrace{R(\psi^n) \circ \cdots \circ R(\varphi^n)(x)\ ]}^{n-2\,veces} \mid \\ & + \cdots + \mid R(\phi^n)[R(\varphi^n)(x)] - R(\varphi^n(x))\mid + \mid R(\varphi^n)(x) - f(x)\mid \\ & <_2 \overbrace{\tfrac{1}{n} + \tfrac{1}{n} + \cdots + \tfrac{1}{n}}^{L-2\,veces} + \mid R(\varphi^n)(x) - f(x)\mid <_1 (L-1)\frac{1}{n} \end{aligned}$$

Obtenemos entonces que $\|R(\phi_n) - f\| \leq (L-1)\frac{1}{n},\ \forall\, n$,por lo tanto.

$$\lim_{n\to\infty} R(\phi_n) = f$$

□

Esto último demuestra el teorema, sin embargo usamos que:

$$\underbrace{R(\phi^n) \circ \cdots \circ R(\phi^n)(x)}_{m\,terminos} \in \tilde{k}$$

Esto para todo $x \in K$ y para todo $m \leq L-2$ para emplear **??**, veamos por que pasa esto:

Demostración. Notemos que por **??** tenemos que:

$$|R(\varphi^n)(x) - f(x)| < \frac{1}{n}$$

Por lo que la distancia

$$(R(\varphi^n)(x), F(k)) \leq \frac{1}{n}$$

Ahora usamos **??**:

$$|R(\psi^n) \circ R(\varphi^n)(x) - R(\varphi^n(x))| < \frac{1}{n}$$

Por lo que:

$$dist(R(\psi^n) \circ \varphi(\varphi^n)(x), (R \circ \varphi^n)(k)) \leq \frac{1}{n}$$

Esto implica que:

$$dist(R(\psi^n) \circ R(\varphi^n)(x), F(K)) \leq \frac{2}{n}$$

De esta forma concluimos que:

$$dist[R(\psi^n) \circ \cdots \circ R(\varphi^n)(x), f(k)] \leq \frac{L}{n}\ \ \forall x \in K$$

Esto implica que $R(\psi^n) \circ \cdots \circ R(\varphi^n)(x) \in \tilde{k},\ \ \forall x \in k$ (Esto por la definicion de $\tilde{k}$) □

4.4. Reaproximación de Diccionarios

Definición 4.4.1

Sea H un espacio normado y $(A_n)_{n\in\mathbb{N}}$ una colección anidada de subconjuntos $A_n \subseteq A_{n+1}$, $\forall n$, dado $C \subseteq H$, definimos:

$$\sigma(A_n, C) := \sup_{f\in C} \inf_{g\in A_n} \|f \cdot g\|$$

Sea $h : \mathbb{N} \to \mathbb{R}^+$, si se cumple que:

$$\sigma(A_n, C) = \mathcal{O}(h(n)), \textit{ cuando } n \to \infty$$

Entonces decimos que $(A_n)_{n\in\mathbb{N}}$ tiene una tasa de aproximación h para C.

Recordemos que en este caso $\sigma(A_n, C) = \mathcal{O}(h(n))$ quiere decir que existe una constante α tal que:

$$\sigma(A_n, C) \leq \alpha \cdot h(n), \ \forall n$$

Definición 4.4.2

Dadas funciones f,g decimos que:

$$f \lesssim g$$

Si existe una constante α tal que $f \leq \alpha \cdot g$ puntualmente.

En lo anterior tenemos que:

$$\sigma(A_n, C) = \mathcal{O}(h(n)) \Leftrightarrow \sigma(A_n, C) \lesssim h(n)$$

Definición 4.4.3: Diccionario

Sea $D = (f_i)_{i=1}^{\infty} \subseteq M$ una sucesión (A la cual llamaremos Diccionario), definimos los espacios:

$$A_N := \{\sum_{i=1}^{\infty} C_i f_i \ : \ C_i \in \mathbb{R} \, y \, \|C\|_0 \leq N\}$$

En donde $\|C\|_0 = \#\{i \in \mathbb{N} \ : \ C_i \neq 0\}$, es claro que $A_N \subseteq A_{N+1}$, $\forall N$

Teorema 4.4.1

Sea $d \in \mathbb{N}$, $H \subseteq \{f : K \subseteq \mathbb{R}^d \to \mathbb{R}\}$ un espacio normado, $\rho : \mathbb{R} \to \mathbb{R}$ y $D := (f_i)_{i=1}^{\infty} \subseteq H$ un diccionario.

Supongamos que existen $L, C \in \mathbb{N}$ tales que $\forall i \in \mathbb{N}, \forall \epsilon > 0$, existe una red neuronal ϕ_i^ϵ tal que:

$$L(\phi_i^\epsilon) = L$$
$$M(\phi_i^\epsilon) \leq C$$
$$\|R(\phi_i^\epsilon) - f_i\|_H \leq \epsilon$$

Para todo $C \subseteq H$, definimos A_N como en la definición 4.4.3. y definimos:

$$B_N := \{R(\phi) \ : \ \phi \, es \, una \, red \, neuronal, \, L(\phi) = L, \, M(\phi) \leq N\}$$

Entonces, $\forall \zeta \subseteq H$

$$\sigma(B_{CN}, \zeta) \leq \sigma(A_N, \zeta)$$

Demostración. Sea $a \in A_N$, entonces:

$$a = \sum_{j=1}^{n} C_{i(j)} f_{i(j)}$$

Sea $\epsilon > 0$, entonces por hipótesis existen redes neuronales $(\phi)_{j=1}^n$ tales que:

- $L(\phi_j) = L$
- $M(\phi_j) \leq C$
- $\|R(\phi_j) - f_{i(j)}\| \leq \dfrac{\epsilon}{N \cdot \text{máx}(\{|C_{i(j)}|\})}$

Definimos:

$$\phi^{(a,\epsilon)} = \phi^c \cdot P[\phi_1, \phi_2, \cdots, \phi_N]$$

en donde $\phi^c = ([C_{i(1)}, C_{i(2)}, \cdots, C_{i(N)}], 0)$, entonces:

$$R(\phi^{(a,\epsilon)}) = \sum_{j=1}^{n} C_{i(j)} R(\phi_j)$$

Por lo que:

$$\begin{aligned}\|R(\phi^{(a,\epsilon)}) - a\| &= \|\sum_{j=1}^{n} C_{i(j)} R(\phi_j) - \sum_{j=1}^{n} C_{i(j)} f_{i(j)}\| \\ &\leq \sum_{j=1}^{n} |C_{i(j)}| \|R(\phi_j) - f_{i(j)}\| \\ &\leq N \cdot \text{máx}(\{|C_{i(j)}|\}) \frac{\epsilon}{N \cdot \text{máx}(\{|C_{i(j)}|\})} \\ &\leq \epsilon\end{aligned}$$

Entonces $\forall \epsilon$, $\forall a \in A_N$, existe $\phi^{(a,\epsilon}$ tal que:

- $L(\phi^{(a,\epsilon)}) = L$
- $M(\phi^{(a,\epsilon)} \leq C \cdot N$
- $\|R(\phi^{(a,\epsilon)}) - a\| \leq \epsilon$

En otras palabras, $\forall \epsilon > 0$, $\forall a \in A_N$, existe $R(\phi^{(a,\epsilon)}) \in B_{CN}$ tal que:

$$\|R(\phi^{(a,\epsilon)}) - a\| \leq \epsilon \tag{4.3}$$

Recordemos que:

$$\sigma(A_N, \zeta) = \sup_{f \in \zeta} \inf_{a \in A_N} \|f - a\|$$
$$\sigma(B_{CN}, \zeta) = \sup_{f \in \zeta} \inf_{b \in B_{CN}} \|f - b\|$$

Sea $f \in \zeta$ y $a \in A_N$, por (4.3.), encontramos $b_\epsilon \in B_{CN}$ tal que $\|a - b_\epsilon\| \leq \epsilon$, entonces:

$$\begin{aligned}\inf_{b \in B_{CN}} \|f - b\| \leq \|f - b_\epsilon\| &= \|f - b_\epsilon + a - a\| \\ &\leq \|f - a\| + \|a - b_\epsilon\| \\ &\leq \|f - a\| + \epsilon\end{aligned}$$

esto para toda $a \in A_N$, por lo que haciendo tender $\epsilon \to 0$, obtenemos:

$$\inf_{b \in B_{CN}} \|f - b\| \leq \inf_{a \in A_N} \|f - a\|, \ \forall f \in \zeta$$

Tomando el supremo en la desigualdad, obtenemos:

$$\sigma(B_{CN}, \zeta) = \sup_{f \in \zeta} \inf_{b \in B_{CN}} \|f - b\| \leq \sup_{f \in \zeta} \inf_{a \in A_N} \|f - a\| = \sigma(A_N, \zeta)$$

Por lo tanto concluimos que:

$$\sigma(B_{CN}, \zeta) \leq \sigma(A_N, \zeta)$$

□

4.5. Aproximación de funciones suaves

Definición 4.5.1: B-Spline

Definimos el B-Spline de orden $k \in \mathbb{N}$ de la siguiente manera:

$$\mathcal{N}_k := \frac{1}{(k-1)!} \sum_{s=0}^{k} (-1)^s \binom{k}{s} (x-s)_+^{k-1},\ x \in \mathbb{R}$$

en donde $0^0 = 0,\ 0! = 1$ y $y_+ = \begin{cases} y,\ si\ y \geq 0 \\ 0,\ si\ y \leq 0 \end{cases}$.

Definición 4.5.2: B-Spline Multivariado

Para $t \in \mathbb{R}$ y $l \in \mathbb{N}$, definimos:

$$\mathcal{N}_{l,t,k} := \mathcal{N}_k(2^p(x-t))$$

Dada $d, l \in \mathbb{N}$ y $t = (t_1, t_2, \cdots, t_d) \in \mathbb{R}^d$, definimos el **B-Spline multivariado**:

$$\mathcal{N}_{l,t,k}^d(x) := \prod_{j=1}^{d} \mathcal{N}_{l,t_j,k}(x_j),\ con\, x = (x_i, \cdots, x_d) \in \mathbb{R}$$

Definición 4.5.3

Finalmente definimos:

$$\mathcal{B}^k := \{\mathcal{N}_{l,t_l,k}^d\ :\ l \in \mathbb{N}, t_l \in 2^{-l}\mathbb{Z}^d\}$$

Teorema 4.5.1

Sea $d, k \in \mathbb{N},\ 0 < s < k$. Entonces existe $c > 0$ tal que para toda $f \in C^s([0,1]^d)$, tenemos que para toda $\delta > 0$ y todo $N \in \mathbb{N}$, existen constantes $C_e,\ e \in \{1, 2, \cdots N\}$, y $B_e \in \mathcal{B}^k$, tales que:

$$|C_e| \leq C\|f\|$$
$$\|f - \sum_{i=1}^{N} C_i B_i\| \lesssim N^{\frac{\delta - s}{d}} \|f\|_{C^s([0,1]^d)}$$

En donde $a \lesssim b \Leftrightarrow$ existe una constante α tal que $\alpha \cdot \beta$, uniformemente con respecto a los parámetros que definen a y b.

Además, $C^s([0,1]^d)$ son las funciones continuamente diferenciables hasta orden s, en nuestro caso tomamos $s \in \mathbb{N}$ y definimos:

$$\|f\|_{C^s([0,1]^d)} := \sum_{J_i + \cdots + J_d \leq s} \|\frac{\partial^{J_i}}{\partial x_i} \cdots \frac{\partial^{J_d}}{\partial x_d} f\|$$

Definición 4.5.4: Sigmoidal

Una función $\rho : \mathbb{R} \to \mathbb{R}$ es llamada sigmoidal de orden $q \in \mathbb{N}$ si $\rho \in C^{q-1}(\mathbb{R})$ y además:

1. $\frac{\rho(x)}{x^q} \to 0$, para $x \to -\infty$, $\frac{\rho(x)}{x^q} \to 1$, para $x \to \infty$
2. $|\rho(x)| \lesssim (1 + |x|)^q,\ \forall x \in \mathbb{R}$

Definición 4.5.5: Función Sigmoidal Estandar

El ejemplo estandar de función sigmoidal de orden q es:

$$\rho_s(x) := x_+^q,\ x \in \mathbb{R}$$

Lema 4.5.1

Dados $\epsilon > 0$ y $p \in \mathbb{N}$, existe una red neuronal ϕ tal que:

$$L(\phi) = \lceil \text{máx}\{\log_q(p-1), 0\} + 1 \rceil$$
$$\|R(\phi) - x_+^{p-1}\|_{[-z,z]} \leq \epsilon$$

Para todo intervalo de la forma $[-z, z]$ en donde $z > 0$ y ϕ depende de z.

Las redes neuronales se toman con función de activación $\rho_s = x_+^q$, aunque la demostración es válida para cualquier función sigmoidal de orden q.

Demostración. Sea $\lambda := \lceil \max(log_q(p-1), 0) \rceil$, entonces tenemos que:

$$\underbrace{\rho_s \circ \rho_s \circ \cdots \circ \rho_s(x)}_{\lambda \, veces} = x_+^{q^\lambda} = x_+^b$$

con $b = q^\lambda$ entonces tenemos que:

$$b = q^\lambda \geq q^{\log_q(p-1)} \geq p - 1$$

Denotamos por $g(x) = x_+^b$, tenemos que:

$$\rho_s \circ \rho_s \circ \cdots \circ \rho_s = g$$

Tomamos, dada $\delta \in \mathbb{R}$,

- $T_1^\delta(x) = \begin{pmatrix} x + \delta \\ x \end{pmatrix}, \ \forall x \in \mathbb{R}$
- $T_l^\delta \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}, \ \forall x, y \in \mathbb{R}, \ \forall l \in \{2, 3, \cdots, \lambda - 1\}$
- $T_{\lambda+1}^\delta \begin{pmatrix} x \\ y \end{pmatrix} = \frac{x-y}{\delta}, \ x, y \in \mathbb{R}$

Sea $\phi^\delta = (T_1^\delta, T_2^\delta, \cdots, T_{\lambda+1}^\delta)$, tenemos que:

$$\begin{aligned} R(\phi^\delta)(x) &:= T_{\lambda+1}^\delta \circ \rho_s \circ T_{\lambda-1}^\delta \circ \rho_s \circ \cdots \circ \rho_s \circ T_1(x) \\ &= T_{\lambda+1}^\delta \circ \rho_s \circ \cdots \circ T_2 \begin{pmatrix} \rho_s(x+\delta) \\ \rho_s(x) \end{pmatrix} \\ &= T_{\lambda+1}^\delta \underbrace{\begin{pmatrix} \rho_s \circ \rho_s \circ \cdots \circ \rho_s(x+\delta) \\ \rho_5 \circ \rho_s \circ \cdots \circ \rho_s(\ \ x\ \ ) \end{pmatrix}}_{\lambda \, veces} \\ &= T_\lambda^\delta \begin{pmatrix} g(x+\delta) \\ g(x) \end{pmatrix} \\ &= \frac{g(x+\delta) - g(x)}{\delta} \end{aligned}$$

Esto implica que $|R(\phi^\delta)(x) - g'(x)| \to 0$, cuando $\delta \to 0$ uniformemente en $[-z, z]$

Concluimos que dado $\epsilon > 0$ existe una red neuronal ϕ^δ tal que:

$$\|R(\phi^\delta) - g'\|_{[-z,z]} < \epsilon \ y \ L(\phi^\delta) = \lambda + 1$$

El siguiente paso es aproximar $g''(x) = b(b-1)x_+^{b-2}$ usando redes neuronales.

Tomamos $h \in (0,1)$ suficientemente chico tal que:

$$|\frac{g(x+h) - g(x) - h \cdot g'(x) - \frac{h^2}{3}g''(x)}{h^2}| < \epsilon, \ \forall x \in [-z, z]$$

(Usamos el Teorema de Taylor: $g(x+h) = g(x) + h \cdot g'(x) + \frac{h^2}{2}g''(x) + \mathcal{O}(h^3)$)

Encontramos $\delta > 0$ tal que:

$$|R(\phi^\delta)(x) - g'(x)| < h^2 \cdot \epsilon, \ uniformemente$$

Entonces tenemos que:

$$\begin{aligned} &|\frac{g(x+h) - g(x) - h \cdot R(\phi^d)(x) - \frac{h^2}{2}g''(x)}{h^2}| \\ \leq &|\frac{g'(x) - R(\phi^d)(x)}{h}| + |\frac{g(x+h) - g(x) - h \cdot g'(x) - \frac{h^2}{2}g''^{(x)}}{h^2}| \\ &< 2 \cdot \epsilon \end{aligned}$$

entonces converge unifromemente (1)

Recordemos que:

$$\phi^d = (T_1^\delta, \cdots, T_{\lambda+1}^\delta)$$

Tomamos ahora:

$$\phi^{\epsilon,\delta} = (T_1^{\epsilon,\delta}, T_2^{\epsilon,\delta}, \cdots, T_{\lambda+1}^{\epsilon,\delta})$$

En donde:

- $T_1^{\epsilon,\delta}(x) = \begin{pmatrix} x+h \\ x \\ T_1^\delta(x) \end{pmatrix}$
- $T_2^{\epsilon,\delta}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ T_2^\delta(y_3) \end{pmatrix}, \cdots, T_\lambda^{\epsilon,\delta}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ T_\lambda^\delta(y_3) \end{pmatrix}$
- $T_{\lambda+1}^{\epsilon,\delta}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \frac{y_1 - y_2 - h \cdot T_{\lambda+1}^\delta(y_3)}{h^2}$

Tenemos que:

$$\begin{aligned}
R(\phi^{\epsilon,\delta})(x) &= T_{\lambda+1} \circ \rho_s \circ \cdots \circ \rho_s \circ T_1(x) \\
&= T^{\epsilon,\delta}_{\lambda+1} \begin{pmatrix} \rho_s \circ \cdots \circ \rho_s(x+h) \\ \rho_s \circ \cdots \circ \rho_s(x) \\ \rho_s \circ T^{\delta}_{\lambda} \circ \cdots \circ \rho_s \circ T^{\delta}_1(x) \end{pmatrix} \\
&= T^{\epsilon,\delta}_{\lambda+1} \begin{pmatrix} g(x+h) \\ g(x) \\ \rho_s \circ T^{\delta}_{\lambda} \circ \cdots \circ \rho_s \circ T^{\delta}_1(x) \end{pmatrix} \\
&= \frac{g(x+h) - g(x) - T^{\delta}_{\lambda+1} \circ \cdots \circ T^{\delta}_1(x)}{h^2} \\
&= \frac{g(x+h) - g(x) - h \cdot R(\phi^{\delta})}{h^2}
\end{aligned}$$

Obtenemos que:

$$R(\phi^{\epsilon,\delta})(x) = \frac{g(x+h) - g(x) - h \cdot R(\phi^{\delta})}{h^2}$$

y de (1) obtenemos que:

$$|R(\phi^{\epsilon,\delta})(x) - \frac{1}{2} g''(x)| < 2 \cdot \epsilon$$

Por lo tanto converge uniformemente. □

Parte II

Ayudantias

5 | Regresión Lineal

La regresión lineal es un modelo matemático q, e consiste en encontrar la **ecuación lineal** que mejor se ajuste o aproxime a un conjunto de datos dado. Dicho de otra manera, se busca aproximar la relación de dependencia entre una variable dependiente **Y** y variables independientes x_i de la siguiente forma:

$$Y = \sum_{i=0}^{d} \widehat{\beta}_i x_i$$

En donde a los $\widehat{\beta}_i$ se les conoce como **estimadores o pesos**. Vamos a considerar $x_0 = 1$

Cuando nos referimos a la que mejor **aproxime** es en el sentido de aquella ecuación lineal que minimice una función llamada función de error o de costo, por lo que la regresión lineal se convierte en un problema de minimización

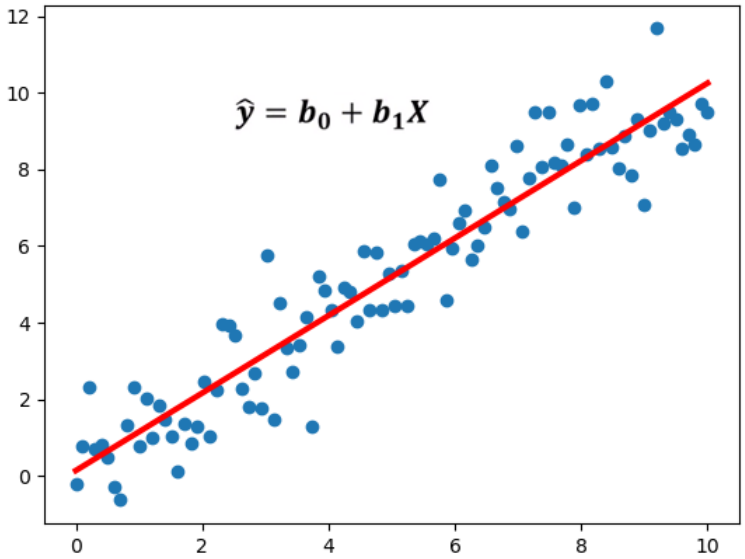

Figura 5.1: Arpoximación Lineal

Durante estas notas vamos a deducir la ecuación:

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

En donde X es el conjunto de datos, Y la variable dependiente, y $\widehat{\beta}$ es el vector de estimadores o pesos.

Definición 5.0.1

Sea $\widehat{w} \in \mathbb{R}^n$ y $\widehat{x} \in \mathbb{R}^m$, definimos a la derivada de $\widehat{x}$ con respecto a $\widehat{w}$ como:

$$\frac{dx}{dw} = \begin{pmatrix} \frac{\partial x_1}{\partial w_1} & \cdots & \frac{\partial x_1}{\partial w_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_m}{\partial w_1} & \cdots & \frac{\partial x_m}{\partial w_n} \end{pmatrix}$$

es decir $\left(\frac{dx}{dw}\right)_{i,j} = \frac{\partial x_i}{\partial w_j}$

Proposición 5.0.1

Sea $A \in M_{nxm}(\mathbb{R})$ y $w \in \mathbb{R}^m$ (o equivalentemente $w \in M_{mx1}(\mathbb{R})$). Si las entradas de A no dependen de las entradas de w entonces:

$$\frac{d(Aw)}{dw} = A$$

Demostración.

$$(Aw)_i = \sum_{k=1}^{n} A_{ik} w_k$$

Entoces:

$$\begin{aligned} \left(\frac{d(Aw)}{dw}\right)_{ij} &= \frac{\partial (Aw)_i}{\partial w_j} \\ &= \frac{\partial \left(\sum_{k=1}^{n} A_{ik} w_k\right)}{\partial w_j} \\ &= \sum_{k=1}^{n} \frac{\partial (A_{ik} w_k)}{\partial w_j} \\ &= \sum_{k=1}^{n} A_{ik} \frac{\partial w_k}{\partial w_j} \\ &= \sum_{k=1}^{n} A_{ik} \delta_{kj} \\ &= A_{ij} \end{aligned}$$

Por lo tanto

$$\left(\frac{d(Aw)}{dw}\right)_{ij} = A_{ij}$$

y se concluye que:

$$\frac{d(Aw)}{dw} = A$$

□

Para la regresión lineal vamos a utilizar la función RSS, la cual es:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Esta función se conoce normalmente como la "función de costo", "loss function.º "función de error". Nos da la suma del error cuadrático para cada una de las predicciones. Esta sería una función que depende de los estimadores $\hat{\beta}_i$ o el vector de estimadores $\hat{\beta}$.

Sea $X \in M_{nxp}(\mathbb{R})$ el conjunto de datos y $\hat{\beta} \in M_{px1}(\mathbb{R})$ el vector de estimadores, entonces:

$$\hat{y} = X\hat{\beta}$$

Definición 5.0.2: RSS

Podemos reescribir a RSS en forma matricial como:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (y - \hat{y})^T(y - \hat{y})$$

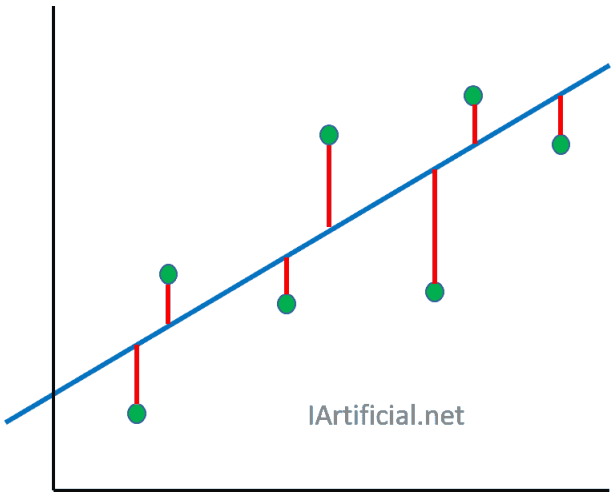

Figura 5.2

Lo que queremos hacer es obtener los estimadores o pesos que minimicen dicha función de error.

$$(AB)^T = B^T A^T$$

Proposición 5.0.2

El vector de estimadores que minimiza a RSS es:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

Demostración.

$$\begin{aligned} RSS &= (y-\hat{y})^T(y-\hat{y}) \\ &= y^Ty - y^T\hat{y} - \hat{y}^Ty + \hat{y}^T\hat{y} \\ &= y^Ty - y^TX\hat{\beta} - \hat{\beta}^TX^Ty + \hat{\beta}^TX^TX\hat{\beta} \end{aligned}$$

Derivamos la expresión anterior con respecto a $\hat{\beta}$ e igualamos la derivada a 0 para encontrar el vector de estimadores que minimiza el error RSS:

$$\frac{dRSS}{d\hat{\beta}} =^{Proposicion\,1} -y^TX + \hat{\beta}^TX^TX = 0$$

$$\begin{aligned} \Rightarrow & y^TX = \hat{\beta}^TX^TX \\ & \Rightarrow X^TX\hat{\beta} = X^Ty \\ & \therefore \hat{\beta} = (X^TX)^{-1}X^Ty \end{aligned}$$

□