Universidad Nacional de Ingeniería
Escuela Profesional de Física

Tópicos de Investigación II

Máquina de Boltzmann

Román Rafaele, Kevin Juan
20104043B

10 de diciembre de 2019

Índice general

1. Introducción **3**

2. Red de Hopfield **4**
2.1. Estructura básica 4
2.2. Aprendizaje Hebbiano 6
2.3. Energía de la red 8
2.4. Pros y contras 9

3. Máquina de Boltzmann **10**
3.1. Convergencia al equilibrio 11
3.2. Proceso de aprendizaje 13
3.3. Unidades Ocultas 16
3.4. Máquina de Boltzmann Restringida 19
3.5. Máquina de Boltzmann como modelo generativo 20

4. Modelo de Ising **21**
4.1. Aproximación de campo medio en el módelo de Ising fundamental . . . 22
4.2. Extensión del modelo de Ising 24

5. Campo Medio en la Máquina de Boltzmann **28**
5.1. Equivalencia entre el modelo de Ising y la red de Hopfield 28
5.2. Teoría de campo medio en la dinámica neuronal 29
5.3. Teoría de campo medio en la dinámica sináptica 30

6. Prueba de aprendizaje sobre MNIST y Discusiones **32**
6.1. Detalles de la prueba 32
6.2. Discusión de los resultados 33
6.3. Sugerencias 36

7. Conclusiones **38**

A. Modélo de Respuesta Neuronal **39**
A.1. Modelo de Integra-y-Dispara con Fuga 39
A.2. Plasticidad dependiente de Coincidencia Temporal 42

Capítulo 1

Introducción

Las redes neuronales artificiales son una de las herramientas de clasificación y regresión más estudiadas en *Machine Learning*, tanto por el potencial de su aplicación como por la teoría en la que se basa. En el caso de aprendizaje no supervizado, las conexiones de las unidades neuronales de estas redes intentan replicar las mismas relaciones estadisticas que presentan cada porción mínima de la *data* contenida en un conjunto objetivo sin esperar que la estructura estadistica del conjunto este explicitamente descrita, como en el caso de aprendizaje supervizado. A partir del formalismo presentado inicialmente por Hopfield, quedo en evidencia la posibilidad de enmarcar el problema de aprendizaje no supervizado (el cual se ha tratado como uno de mera optimización) como uno de busqueda de equilibrios, sumado al isomorfismo que tienen estas redes con las *lattices* de Ising. A partir de esto ultimo, Geofrey y Sejnowski adaptaron el formalismo de Hopfield a uno estocástico, completando el márco termodinámico en el que se estudia este tipo de redes. En este trabajo se describe esta red, denominada Máquina de Boltzmann, asimismo se propone su estudio siguiendo un enfoque físico, mostrando su equivalencia con una red de espines del modelo de Ising, lo que nos permite utilizar el método del campo medio, utilizado en el modelo de Ising, para mejorar el aprendizaje de una máquina de Boltzmann.

Capítulo 2

Red de Hopfield

Una red neuronal artifical es un sistema inspirado de las redes neuronales biológicas las cuales son capces de gestionar información, tanto en su almacenado como transmisión. Una red neuronal .ªprende.ª realizar cierto proceso con el fin de resolver un problema específico siguiendo un algoritmo de aprendizaje definido. Una red neuronal esta conformada por neuronas, sinapsis y una regla o función de activación. Las neuronas son las unidades minimas de computación en la red, estas reciben información transmitida por otras neuronas y la procesa definiendo un estado interno, el cual según su valor decidirá la información que sera transmitida a las demás neuronas. Esta información se transmite a travez de conexiones entras las neuronas llamadas sinapsis, y la eficacia de la transmisión de esta información depende de los valores los pesos sinápticos, los cuales indican la fuerza de transmisión que presentan las sinapsis. Asimismo, para poder mediar como los estados guardados en las neuronas causarian o no la transmisión de información hacia otras neuronas hace falta un criterio de activación, siendo indicada por medio de una función de activación especifica.

2.1. Estructura básica

Sea $x \in \mathbb{R}^N$ el conjunto de estados de cada neurona que componen una red neuronal artifical, es decir, estan conectadas y transmiten información entre ellas. Sea x_i un estado de la neurona i, el cual guarda y transmite a las demas neuronas. El estado de esta neurona esta definido mediante la activación producida por las demás neuronas la cual podemos expresar mediante

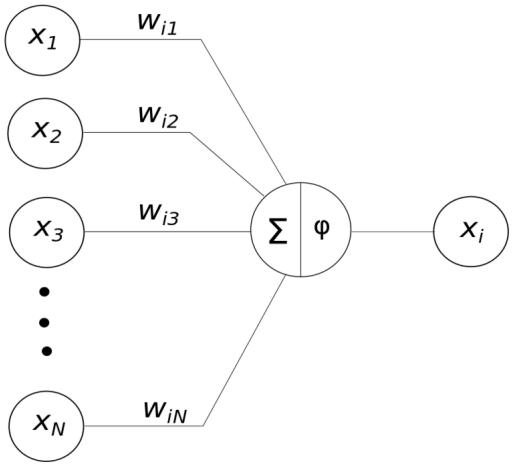

Figura 2.1: Esquema del perceptrón

La activación de una neurona i esta determinada por los valores de entrada que recibe de otras neuronas x, mediadas por los pesos sinápticos w, y finalmente por la función de activación ϕ

$$x_i = \phi(\sum_j w_{ij} x_j) \tag{2.1}$$

Siendo w_{ij} el peso sináptico de la conexión entre la neurona i y la neurona j. Esto permite definir el conjunto de pesos como una matriz cuadrada $w \in \mathbb{R}^{nxn}$. Cabe resaltar que esta matriz no es necesariamente simétrica, y aunque suele definirsela con los elementos de la diagonal iguales a cero (no existe autoinducción de activación) podria definirse un modelo el cual acepte elementos diferentes de cero, ya sea como un proceso de retroalimentación o de decaimiento. Asimismo, tenemos que la activación de la neurona va a regirse según la función de activación ϕ.

Dado el criterio biológico de "disparo"del cual se inspira este modelo, suele restringirse las activaciones a valores binarios. Esto implica que se utilice una función de activación de este tipo

$$\phi(z) = \begin{cases} 1 & z > \theta_i \\ -1 & z \leqslant \theta i \end{cases} \tag{2.2}$$

Siendo θ_i el umbral de la función de activación para la neurona i. Cabe recalcar que este es un ejemplo de como se puede definir una función de activación determinista para estados binarios.

2.2. Aprendizaje Hebbiano

Aprender, en el contexto de redes neuronales, es el proceso por el cual una red adquiere la estructura estadśitica de un conjunto de estados dados. Para entender esto, se puede ver a la red como una función parametrizada (por los pesos de conexión) la cual, mediante un algoritmo de regresión, se aproxima a una distribución que describe un conjunto de estados especifico. Se dice que una configuración de estados es estable dada una configuración fija de pesos si es que los estados de la neurona no cambian frente a activaciones sucesivas. Entonces, dada una configuración de estados especifica s_d, el aprender implica calibrar las conexiones neuronales para encontrar la configuración de pesos el cual tenga como estable la configuración s_d. Asimismo, sea un conjunto D de configuraciones de una red neuronal, entonces el aprender D equivale a lograr que las configuraciones dentro de ese conjunto sean estables para la red.

Hebb observo que las neuronas biológicas usan un criterio de correlación temporal de disparos para alterar sus sinapsis, es decir, dependiendo de que tan antes o despues se dan los disparos postsinápticos frente a los presinápticos la neurona decide potenciar si encuentra causalidad entre disparos o deprimir si no es así, de ahi su lema *neurons that fire together, wire together*[3][1].

A partir de ese principio, se ha inspirado un algoritmo de aprendizaje para las redes neuronales artificiales. Se define entonces la ecuación diferencial de aprendizaje de Hebbs como

$$\frac{dw_{ij}}{dt} = \eta x_i x_j \tag{2.3}$$

o, en su versión discreta

$$\Delta w_{ij} = \eta x_i x_j \tag{2.4}$$

donde η es la taza o ratio de aprendizaje. Se encuentra la capacidad de aprendizaje de esta regla reemplazando en 2.3 la ecuación 2.1. Tomando la función de activación, por simplicidad, igual a la función identidad obtenemos:

[1] Traducido literalmente quiere decir "neuronas que disparan juntas, se unen juntas", describiendo de forma breve y elocuente como las neuronas potencian su conectividad en base al grado de causalidad entre los disparos que hacen estas

$$\frac{dw_{ij}}{dt} = \eta x_i \sum_k w_{jk} x_k \approx \eta \sum_k w_{jk} < x_i x_k >= \eta \sum_k w_{jk} c_{ik} \tag{2.5}$$

Siendo $c_{ik} =< x_i x_k >$ la correlación entre las activaciónes de las neurones i y k. En notación matricial se observa que esta ecuación es igual

$$\frac{dw}{dt} = \eta c w \tag{2.6}$$

Siendo c la matriz de correlación de los estados de la neurona y que w es una matriz simétrica ya que las conexiones lo son. Haciendo $w = w_o \exp(\lambda t)$, siendo w_o nuestra configuración inicial de pesos, tenemos

$$\frac{d}{dt}(w_o \exp(\lambda t) = \lambda w_o \exp(\lambda t) = c w_o \exp(\lambda t) \tag{2.7}$$

despejando w_o queda

$$(c - \lambda \mathbb{I}) w_o = 0 \tag{2.8}$$

siendo $\mathbb{I}$ la matriz identidad. Se observa que este es un problema de autovalores, del cual podemos deducir que, para un autovalor λ tenemos su respectivo autovector u_λ y por ende se puede expresar nuestros pesos como una combinación lineal de estos

$$w = \sum_\lambda a_\lambda u_\lambda \exp(\lambda t) \tag{2.9}$$

siendo a_λ coeficientes de esta combinación lineal que cumplen la condición inicial

$$w_o = \sum_\lambda a_\lambda u_\lambda \tag{2.10}$$

Se ve que cuando $t \to \infty$, $w \to a_\lambda u_\lambda \exp(\lambda t)$, es decir, los pesos van adquiriendo el valor proporcional a la componente principal [10]. Por teoriía de analisis de datos, la componente principal indica la dirección de mayor dispersi]'on en un conjunto de datos vectorizados. En este caso, si se tiene una estructura estadiśitica (es decir, una distribución) que define un conjunto de datos, entonces la perturbación ajena a esa estructura presente en los datos será considerado ruido, y por lo tanto la componente principal será un indicador o representante de esa estructura-patrón definida.

2.3. Energía de la red

Sea una red con las siguientes caracteristicas: la matriz de pesos es simétrica, y la función de activación es definida como en 2.2. Con estas caracteristicas, se define la energía de esta red de la siguiente manera

$$E = -\sum_{i>j} w_{ij} x_i x_j + \sum_i \theta_i x_i \tag{2.11}$$

siento θ_i el umbral de activación de la neurona i. Se denomina a esta cantidad .enˆergía"dado que, según nuestro proceso de activación, esta cantidad ira disminuyendo[4]. Sea ΔE_i la variación que se produce en la energía debido a la variación del estado en la neurona i producida por la activación en esta

$$\Delta E_i = -\Delta x_i (\sum_{i \neq j} w_{ij} x_j - \theta_i) \tag{2.12}$$

Siendo Δx_i la variación del estado de la neurona i. Si el segundo factor $(\sum_{i \neq j} w_{ij} x_j - \theta_i)$ es mayor que cero, significa que la activación posterior de la neurona i es $+1$, lo que significa que su variación o es positiva o nula, dejando la variación de la energía negativa o nula, y visceversa, si el segundo factor es negativo, la varciación de la activación sera negativa o nula y la variación de la energía sera decreciente. En cualquiera de ambos casos, el proceso de activación neuronal disminuye dinámicamente el valor de la función energía propuesta. Asimismo, considerando la ecuación de aprendizaje 2.4, podemos definir la variación de la energia ΔE_{ij} producida por el cambio del peso que conectan las neuronas i y j de la siguiente manera

$$\Delta E_{ij} = -\Delta w_{ij} x_i x_j = -\eta x_i^2 x_j^2 \tag{2.13}$$

Se observa aquí como la energía disminuye frente a la dinámica de los pesos segun Hebbs, lo cual refuerza la intuición sobre la elección de función energía[7].

Una red definida con estas características, a la cual se le asigna una función energía tiene el nombre de red de Hopfield, la cual tiene como objetivo final de aprendizaje el reducir su valor de energía para una o varias configuraciones especificas de estados neuronales. Se define entonces la ecuación hamiltoniana de aprendizaje

$$\frac{dw_{ij}}{dt} = -\eta \frac{\partial E}{\partial w_{ij}} \tag{2.14}$$

Lo cual indica como la variacón de pesos implica una disminución de la energía con una taza de disminuciíon η equivalente a la taza de aprendizaje en la regla de Hebbs original.

2.4. Pros y contras

La regla de Hebbs sigue un paradigma de aprendizaje denominado aprendizaje no supervizado, es decir, no requiere que los estados a aprender esten etiquetados o preclasificados para su reconocimiento posterior, la red es capaz de hacer ese reconocimiento por si misma como parte del proceso de aprendizaje. Esto ofrece una ventaja frente a otros modelos de aprendizaje supervizado dado que estos otros modelos requeririan una mayor información de los datos de entrenamiento, los cuales no necesariamente son accesibles por el usuario. Sin embargo, esto ofrece un mayor nivel de dificultad frente a datos de alta dimensionalidad, lo cual implica que la red pueda capturar el patrón equivocado y reforzar ese error en el proceso de aprendizaje, esto equivale a que en el transcurso de la disminución de la energía la red caiga en un minimo local[7], uno del cual ya no pueda salir (esto por responsabilidad del valor de la taza de aprendizaje η)[5]. Una forma de solventar ese problema es usar el ruido dentro de los datos de entrenamiento como perturbación para que la red sea capaz de escapar de los minimos locales, algo fuera del marco de una regla de aprendizaje deteminista como la expuesta aquí para la red de Hopfield.

Capítulo 3

Máquina de Boltzmann

La máquina de Boltzmann es una red neuronal de Hopfield, es decir, una red neuronal con una función de energía asignada[7]

$$E(x) = -\left(\sum_{i \neq j} w_{ij} x_i x_j + \sum_i b_i x_j\right) \tag{3.1}$$

Siendo $x \in \{1, -1\}^N$ el vector que contiene los valores contenidos en las neuronas de la red, w los pesos de conexión entre los neuronas y b los sesgos de cada neurona.

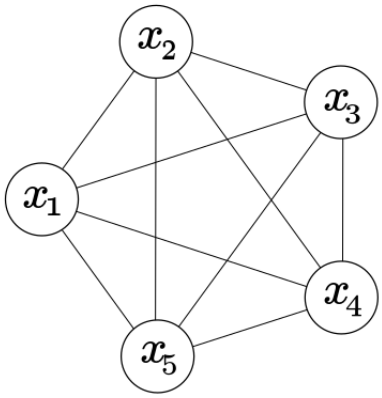

Figura 3.1: Red de Hopfield

Adicionalmente a la función energía, también presenta una distribución de probabilidad, la cual sigue la estadística de Boltzmann

$$p(x) = \frac{\exp(-\beta E(x))}{Z} \tag{3.2}$$

Siendo β la constante termodinámica de la distribución del sistema, y Z es la función de partición

$$Z = \sum_x \exp(-\beta E(x)) \tag{3.3}$$

Este sistema presenta una dinámica estocástica, presentando ventajas frente a su antecesor determinista (red de Hopfield) tanto en capturar la estadística de datos con alta dimensionalidad. El parametro termodinámico β determina que tan rapida será la dinámica de estados de las neuronas en la busqueda del equilibrio.

3.1. Convergencia al equilibrio

La dinámica que siguen los nodos de la red bajo valores de pesos(y sesgos) fijos persigue el equilibrio térmico, el cual se consigue bajando el valor de la energía siguiendo trancisiones estocasticas que conllevan a eso. Para esquematizar este proceso elijamos arbitrariamente una neurona x_i de nuestra red, el cual puede tomar el valor de 1 o -1. Para ambos se tendran los siguientes valores de energía:

$$\begin{aligned} E(x_i = 1) &= -\textstyle\sum_{i\neq j} w_{ij}x_j - b_i - \left(\sum_{j\neq k\neq i} w_{jk}x_j x_k + \sum_{j\neq i} b_j x_j\right) \\ E(x_i = -1) &= \textstyle\sum_{i\neq j} w_{ij}x_j + b_i - \left(\sum_{j\neq k\neq i} w_{jk}x_j x_k + \sum_{j\neq i} b_j x_j\right) \end{aligned} \tag{3.4}$$

Las cuales nos permite describir la energía de transición

$$\Delta E_i = E(x_i = -1) - E(x_i = 1) = 2\left(\sum_{i\neq j} w_{ij}x_j + b_i\right) \tag{3.5}$$

Usando la expresión de la distribución de probabilidad, podemos reexpresarla

$$E(x) = -\frac{1}{\beta}\left(\log p(x) + \log Z\right) \tag{3.6}$$

y asi redefinir 3.4

$$\begin{array}{rcl} E(x_i = 1) & = & -\frac{1}{\beta}\left(\log p(x_i = +1) + \log Z\right) \\ E(x_i = -1) & = & -\frac{1}{\beta}\left(\log p(x_i = -1) + \log Z\right) \end{array} \tag{3.7}$$

Reemplazamos en nuestra igualdad sobre la energía de transición

$$-\frac{1}{\beta}\log\left(\frac{p(x_i = -1)}{p(x_i = +1)}\right) = 2\left(\sum_{i \neq j} w_{ij}x_j + b_i\right) \tag{3.8}$$

Dado que x_i solo puede tomar valores binarios, $p(x_i = -1) = 1 - p(x_i = +1)$

$$\log\left(\frac{1 - p(x_i = +1)}{p(x_i = +1)}\right) = -2\beta\left(\sum_{i \neq j} w_{ij}x_j + b_i\right) \tag{3.9}$$

Despejando $p(i = +1)$ tenemos la probabilidad de estado para un solo nodo

$$p(x_i = +1) = \frac{1}{1 + \exp(-2\beta\Delta E_i)} \tag{3.10}$$

La relación aquí encontrada define la activación estocastica de las otras neuronas sobre la neurona x_i mediante una función logística. Dejando β fijo, hacemos que el sistema busque minimizar su energía mediante transiciones estocasticas usando nuestra relación encontrada[7]. La elección del valor del beta termodinámico determinaria tanto la rapidez como la precisión de la busqueda del equilibrio.

Algorithm 1: Método de Montecarlo en convergencia de equilibrio térmico en una máquina de Boltzmann simple para un valor de temperatura fijo

Se inicializa aleatoriamente array de valores $x \in \{1, -1\}^N$;
Se fija un valor de temperatura T;
Para una configuración dada de pesos w y sesgos b se define una función de energía $E_{w,b}$;
for $t >> 1$ **do**
    Se elige aleatoriamente un nodo x_i;
    Se calcula $p_i = 1/\left(1 + \exp(-\frac{\Delta E_i}{T})\right)$;
    Se genera aleatoriamente un número r entre $[0, 1]$;
    if $p_i > r$ **then**
        $x_i = 1$;
    else
        $x_i = 0$
    end
end

Para una red de neuronas binarias x con conexiones w y sesgos b, este algoritmo elige una de esas neuronas de forma aleatoria, luego calcula la probabilidad de que esta se active segun el valor contenido en las demas neuronas. A continuación, esta probabilidad es comparada con un número entre 0 y 1 generado aleatoriamente usando una distribución uniforme, y según lo que esa comparación indique se pasa a activar o no la neurona elegida. Se repite este proceso durante varios pasos temporales dentro de una función recursiva **for**. Esto consigue llevar a la red a un estado estable al disminuir su energía.

Usando este algoritmo dentro de uno más grande (ej. algoritmo de recocido simulado) se puede de encontrar el estado de equilibrio adecuado del sistema con mayor eficiencia.

3.2. Proceso de aprendizaje

Dado un conjunto de vectores de estados binarios $D \subset \{1, -1\}^N$ la cual representa los datos de entrenamiento, el proceso de aprendizaje consiste en ajustar los parametros de la red (pesos y sesgos) de tal manera que la configuraciones de nodos de la red con mayores probabilidades de incidencia (o sea, en estado de equilibrio) sean las que estan dentro de mi conjunto D. En otras palabras, se buscara resolver el sistema

$$\frac{\partial \log p}{\partial w_{ij}}(x \in D) = 0 \tag{3.11}$$

Esto es, se desea maximizar la probabilidad de obtener como estado de equilibrio del sistema uno cualquiera que pertenezca al conjunto D equivalentemente. Esto es equivalente a intentar minimizar la siguiente función perdida

$$L_D = -\frac{1}{N}\sum_{x \in D} \log p(x) = - < \log p(x) >_{x \in D} \tag{3.12}$$

Esta cantidad coincide con la definición de entropía muestral, por lo que la intuición del proceso de minimización es más completa: al minimizar la entropía muestral sobre el conjunto D, la distribución estará cada vez más sesgada en generar los elemenots del conjunto aleatoriamente. A partir de aquí podemos definir nuestra equación de aprendizaje

$$\frac{dw_{ij}}{dt} = -\gamma \frac{\partial L_D}{\partial w_{ij}} \tag{3.13}$$

Siendo γ nuestra taza de aprendizaje. La formulación es equivalente para los sesgos

$$\frac{db_i}{dt} = -\gamma \frac{\partial L_D}{\partial b_i} \tag{3.14}$$

Desarrollamos entonces la derivada de nuestra función de perdida

$$\frac{\partial L_D}{\partial w_{ij}} = - < \frac{\partial \log p}{\partial w_{ij}}(x) >_{x \in D} \tag{3.15}$$

La derivada ingresa linealmente dentro de la media sobre el conjunto D dado que su distribución es constante (en realidad, es esta la distribución que la red intenta regresionar). Se empieza a resolver la derivada dentro de los braquets

$$\frac{\partial \log p}{\partial w_{ij}} = \frac{1}{p(x)} \frac{\partial p}{\partial w_{ij}}(x) \tag{3.16}$$

Resolvemos aparte la derivada de la probabilidad

$$\frac{\partial p}{\partial w_{ij}}(x) = -\beta \frac{\exp(-\beta E(x))}{Z} \frac{\partial E}{\partial w_{ij}}(x) - \frac{\exp(-\beta E(x))}{Z^2} \frac{\partial Z}{\partial w_{ij}} \tag{3.17}$$

Esta expresión contiene la derivada de la función energía descrita en 3.1 respecto a un peso, cuya resolución se da por

$$\frac{\partial E}{\partial w_{ij}} = -\sum_{k<l} \frac{w_{kl}}{w_{ij}} x_k x_l = -\sum_{k<l} \delta_{ki} \delta_{lj} x_k x_l \tag{3.18}$$

Los deltas de Kronecker dentro de la sumatoria aislan un solo termino dejando como resultado

$$\frac{\partial E}{\partial w_{ij}} = -x_i x_j \tag{3.19}$$

Asimismo, debemos resolver la derivada de la función de partición

$$\frac{\partial Z}{\partial w_{ij}} = -\beta \sum_x \exp(-\beta E(x)) \frac{\partial E}{\partial w_{ij}}(x) = \beta Z \sum_x p(x) x_i x_j \tag{3.20}$$

En otras palabras

$$\frac{1}{Z} \frac{\partial Z}{\partial w_{ij}} = \beta < x_i x_j >_p \tag{3.21}$$

Reemplazando en 3.17, la derivada de la probabilidad de la siguiente manera

$$\frac{\partial p}{\partial w_{ij}}(x) = \beta p(x) \left(x_i x_j - < x_i x_j >_p \right) \tag{3.22}$$

Teniendo ya la expresión para la derivada de la probabilidad, se reemplaza en 3.15, quedando asi la derivada de la función perdida

$$\frac{\partial L_D}{\partial w_{ij}} = -\beta \left(< x_i x_j >_{x \in D} - < x_i x_j >_p \right) \tag{3.23}$$

Con esto la ecuación de aprendizaje queda

$$\frac{dw_{ij}}{dt} = \gamma\beta\left(< x_i x_j >_{x\in D} - < x_i x_j >_p\right) \tag{3.24}$$

El primer termino de la ecuación de aprenziaje expresa la correlación de los nodos conectados por el peso para valores dentro de las configuraciones de la data, el peso se potenciara en el tiempo en la medida que la correlación entre ambos nodos sea fuerte. El segundo termino en cambio expresa la correlación de las neuronas siguiendo configuraciones que nos resulta del modelo estadístico (siguiendo la distribucion p de ese instante), este término ayuda a que el sistema pueda escapar de mínimos locales (lo que uno busca es el mínimo global como objetivo optimo) al hacer que la busqueda pueda oponerse a la dirección que indica el primer termino. Asimismo, permite que los pesos no crezcan indefinidamente sino que converjan a medida que la distribución que define la red se acerque a la que presenta la data[2].

3.3. Unidades Ocultas

Esta red, como ya se comentó, busca encontrar las relaciónes estadísticas entre los miembros de un conjunto de entrenamiento capturando esas características en los parametros de la red, los cuales para la red descrita, se presenta en un número del orden $(N-1)^2$[7], esto parte del hecho de que la red presenta una cantidad de neuronas equivalente al número de componentes de los vectores de entremiento. Estos parametors pueden para casos prácticos no bastar en cantidad para la cantidad de relaciones presentes entre los elementos del conjunto D, por lo que se requiere más parametros de los disponibles en la red, lo que requiere entonces agregar más nodos en nuestra red. Estas unidades neuronales extra no visibles no capturan la configuración de las configuraciones de entrenamiento sino que solamente permiten transmitir información mediante conexiones que codifican otros tipos de correlaciones que las conexiones directas entre los nodos visibles no se dan abasto de capturar[5]. Entonces se pueden clasificar las neuronas x_i en visibles v_i y ocultas h_i. La dependencia de la energía sería descrita de forma más especifica

$$E(v,h) = \sum_{i<j} W_{ij}^{(vv)} v_i v_j + \sum_{i<j} W_{ij}^{(hh)} h_i h_j + \sum_{i,j} W_{ij}^{(vh)} v_i h_j + \sum_i b_i^{(v)} v_i + \sum_i b_i^{(h)} h_i \tag{3.25}$$

Asimismo con la distribución toma la forma

$$p(v,h) = \frac{\exp(-\beta E(v,h))}{Z} \tag{3.26}$$

$$Z = \sum_{v,h} \exp(-\beta E(v,h)) \tag{3.27}$$

Para esta nueva versión de la red, se requiere una nueva función de perdida. Antes de formularse, se tendrá en cuenta que la distribución para una configuración de neuronas visibles es

$$p(v) = \sum_{h} p(v,h) \tag{3.28}$$

La misma intuición para el caso simple se tiene para este caso con nodos ocultos: se busca ajustar los parámetros de la red de manera que las configuraciones de nodos visibles con mayor probabilidad sean los que estan dentro del conjunto D, para cualquier configuración de nodos ocultos dentro de la condición de equilibrio. Teniendo en cuenta que la distribución que da como mayor probabilidad de configuración cualquiera dentro de D es la distribución de estos elementos p_D, entonces se puede reformular el objetivo al de hacer que la distribución de nuestra red converja a la distribución de la *data*. Se formula este objetivo como un problema de minimización sobre la nueva función de perdida

$$L_D = -\frac{1}{N} \sum_{v \in D} \log(\frac{p_D(v)}{p(v)}) = - < \log(\frac{p_D(v)}{p(v)}) >_D \tag{3.29}$$

La cual es la diferencia de información que contiene el conjunto de datos D, usando su distribución real p_D, respecto a la que que se tendría si se asume a priori que D sigue la distribución de nuestra red p. Se pasa entonces a desarrollar

$$\frac{\partial L_D}{\partial w_{ij}} = 0 \tag{3.30}$$

Desarrollando esto de forma más explicita

$$\frac{\partial L_D}{\partial w_{ij}} = - < \frac{\partial}{\partial w_{ij}} \log(\frac{p_D(v)}{p(v)}) >_D = < \frac{\partial \log p(v)}{\partial w_{ij}} >_D \tag{3.31}$$

Pero teniendo en cuenta que , dado 3.28

$$\frac{\partial \log p(v)}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \log \sum_h p(v,h) = \frac{1}{p(v)} \sum_h \frac{\partial p(v,h)}{\partial w_{ij}} \tag{3.32}$$

Derivando 3.26, se tiene que

$$\frac{\partial p(v,h)}{\partial w_{ij}} = \beta p(v,h)(x_i x_j - < x_j x_j >_{p(v,h)}) \tag{3.33}$$

Reemplazando en en 3.32 se obtiene

$$\frac{\partial \log p(v,h)}{\partial w_{ij}} = \beta \frac{1}{p(v)} \sum_h p(v,h)(x_i x_j - < x_i x_j >_{p(v,h)}) \tag{3.34}$$

Reemplazando nuestro resultado en 3.31 Por lo que la derivada de la función de perdida queda

$$\frac{\partial L_D}{\partial w_{ij}} = \beta < \frac{1}{p(v)} \sum_h p(v,h) \left(x_i x_j - < x_i x_j >_{p(v,h)} \right) >_D \tag{3.35}$$

Por definición de probabilidad condicional

$$\frac{p(v,h)}{p(v)} = p(h|v) \tag{3.36}$$

$$\frac{\partial L_D}{\partial w_{ij}} = \beta \left(< \sum_h p(h|v) x_i x_j >_D - < x_i x_j >_p \right) \tag{3.37}$$

el primer termino de la derecha esta ecuación se desarrolla

$$< \sum_h p(h|v) x_i x_j >_D = \sum_v p(h|v) p_D(v) x_i x_j \tag{3.38}$$

Este ultima expresión es la media del producto $x_i x_j$ para todas las configuraciones admisibles en la red en equilibrio siguiendo su propia distribución pero fijando los nodos visibles al conjunto de data $v \in D$. Se puede entonces formular la derivada de nuestra función

$$\frac{L_D}{\partial w_{ij}} = -\beta \left(< x_i x_j >_{free} - < x_i x_j >_{fixed} \right) \tag{3.39}$$

Con $< x_i x_j >_{free} = < x_i x_j >_p$ la media del producto de nodos i y j siguiendo la distribucón p de la red de forma libre, y $< x_i x_j >_{fixed} = \sum_v p(h|v) p_D(v) x_i x_j$ restringidos a configuraciones de nodos visibles iguales a los contenidos en D. Nuestra ecuación de aprendizaje queda por ende

$$\frac{dw_{ij}}{dt} = \gamma\beta \left(< x_i x_j >_{free} - < x_i x_j >_{fixed} \right) \tag{3.40}$$

Esta expresión representa muy bien la generalización de nuestra formula para una red de solo nodos visibles, podemos por ende afirmar con seguridad la misma intuición usada para la anterior en esta.

3.4. Máquina de Boltzmann Restringida

Una aparente ventaja teórica que ofrece esta red es que es capaz de guardar una gran cantidad de memoria limitada por el tamaño de las configuraciones cuya distribución la red debe aprender; pero en realidad, a nivel implementativo, la complejidad del problema crece con el tamaño de los datos en la proporción $N(N-1)$. Esto sumado a la propagación de error no requerido vuelve el problema de aprendizaje usando una máquina de Boltzmann bastante costoso (en este caso, el coste es el gasto de tiempo que requiere el programa para poder terminar de aprender, asumiendo que lograra hacerlo). Una manera de enfrentar ese problema fue usando la siguiente arquitectura:

- Se elije plantear la red como una máquina de Boltzmann con unidades ocultas
- Se retiran las conexiones entre unidades visibles, asi como entre unidades ocultas, dejando solo las conexiones de las unidades visibles con las ocultas

A partir de esto, tenemos muchas menos conexiones, es decir, menos pesos sinápticos, almacenando las propiedades correlativas entre los unidades visibles en las conexiones que existen entre las unidades ocultas que conectan las unidades visibles mencionadas. Esta correlación, antes descrita explicitamente en los pesos que conectaban unidades visibles entre si, ahora son guardadas implicitamente en los pesos que conectan ambas unidades visibles con el mismo nodo oculto. Esta arquitectura aquí planteada para nuestra red no interfiere con los resultados generales descritos en las secciones anteriores, lo que implica que para redes con esta arquitectura las ecuaciones generales tanto de activación neuronal como plasticidad son validas.

3.5. Máquina de Boltzmann como modelo generativo

Un modelo generativo es una función que define una distribución de probabilidad condicional entre dos variables: un observable X y una variable .^{et}iqueta"Y. Esta probabilidad condicional se da como $p(X|y = Y)$. De esto se desprende que, si tenemos un modelo generativo f y queremos obtener un valor de variable cuya etiqueta sea Y, entonces f nos dara como valores de salidad o imagen X's que estén clasificados o etiquetados con Y. Un *autoencoder* es un ejemplo de modelo generativo el cual aprende configuraciones de un conjunto cuya distribución no es conocida[5]. El *autoencoder* aprende la distribución y entonces es capaz a partir de ahí de generar más miembros de la familia o clase de configuraciones que siguen esa misma distribución. Con las definiciones anteriores, se puede apreciar como la máquina de Boltzmann sera un modelo generativo competente frente a *autoencoders*, esto sin necesitar de una variable .^{et}iqueta"dado que su aprendizaje es hebbiano (por ende, no supervizado). Esto ultimo es lo que hace competente la máquina de Boltzmann frente a otros modelos generativos contemporaneos.

Capítulo 4

Modelo de Ising

El modelo de Ising consiste en una completa formulación matemática empleada para describir formalmente el ferromagnetismo mediante propiedades termodinámicas. El tipo de sistema descrito por este modelo es el un arreglo de varios dipolos de espines situados en una lattice graficamente periódica[6]. Cada valor dipolar de los espines s_i toma uno de dos valores $-1, +1$, y se dice que la configuración de los espines $\{s_i\}$ define el estado del sistema. La energía del sistema esta definida como

$$E(s) = -J \sum_{(i,j)} s_i s_j - H \sum_i s_i \tag{4.1}$$

Donde los coeficientes de interacción J y del campo magnético H son constantes. Notese el subdindice de la primera sumatoria (i, j), esto indicia que la sumatoria recoje pares de espines que esten adyacentes entre sí. Podemos entonces definir nuestra distribución de probabilidad en el equilibrio[8]

$$p(s) = -\frac{\exp(-\beta E(s))}{Z(H,T)} \tag{4.2}$$

con la función de particioón

$$Z(H,T) = \sum_s \exp(-\beta E(s)) \tag{4.3}$$

Nuestras funciones termodinámicas son obtenidas a partir de la función de partición. En este caso tenemos la energía libre de Hemholtz

$$F(H,T) = -\frac{1}{\beta}\log Z \tag{4.4}$$

una cantidad relevante es la magnetización

$$m = M/N \tag{4.5}$$

siendo N el número de dipolos y M el momento magnético medio

$$M = -\frac{\partial F}{\partial H} =< \sum_i s_i >= \sum_i < s_i > \tag{4.6}$$

La cual expresa la alineación esperada media de los espines. Si el sistema presenta $m \neq 0$ entonces se denomina al sistema ferromagnético.

4.1. Aproximación de campo medio en el módelo de Ising fundamental

Dado los grados de libertad en el modelo de Ising, así como el termino correlativo de la interacción entre dipolos y el quiebre de simetría de inversión que genera la presencia del campo magnético externo hacen complicada la resolución de este módelo. El planteo del problema que ofrece la teoría de campo medio es uno de los tantos propuestos para obtener resultados aproximados usando calculo variacional[6].

Sea la energía que aporta al total un solo espín frente a los demás

$$E_i = -(J\sum_{j\leftarrow i} s_j + H)s_i \tag{4.7}$$

De esta ecuación podemos inferir un campo neto sobre el espín i

$$H_i' = J\sum_{j\leftarrow i} s_j + H \tag{4.8}$$

dado que las interacciones y el campo son constantes, la red de Ising presenta simetría espacial respecto a desplazamientos. Eso tiene como consecuencia que $m = m_i$ para cualquier espín i en la red. Con esto podemos reformular nuestra ecuación 4.8

$$H_i' = \gamma J m + H + J \sum_{j \leftarrow i} (s_j - m) \tag{4.9}$$

Siendo γ el número de espínes que rodean directamente un solo espín de la red. La aproximación de campo medio en este caso resulta al despreciar las variaciones de los valores de los espínes frente al valor medio $(s_j - m)$, teniendo así nuestro campo medio

$$H' = \gamma J m + H \tag{4.10}$$

Con nuestra aproximación, reformulamos nuestra función energía como

$$E'(s) = -H' \sum_i s_i = \sum_i E_i' \tag{4.11}$$

con

$$E_i' = H' s_i \tag{4.12}$$

Dado que con esta aproximación desaparecen las interacciones explicitamente, podemos reformular la función de particion de la red a partir de una función de particion independiente de cada espín.

$$Z_i' = \sum_{s_i = +1, -1} \exp(-\beta E_i') = \exp(\beta H') + \exp(-\beta H') = 2\cosh(\beta H') \tag{4.13}$$

es decir

$$Z_i' = 2\cosh(\beta(\gamma J m + H)) \tag{4.14}$$

y por ende, la función de partición del sistema quedara

$$Z' = \prod_i Z_i' = 2^N \cosh^N(\beta(\gamma J m + H)) \tag{4.15}$$

La función de energía libre de Hemholtz bajo nuestra aproximación queda segun 4.4

$$F' = -\frac{1}{\beta}\log Z' \tag{4.16}$$

Con esto podemos formular la ecuación para nuestra magnetización

$$m = \frac{M}{N} = -\frac{1}{N}\frac{\partial F'}{\partial H} = \frac{1}{N\beta}\frac{\partial \log Z'}{\partial H} = \frac{1}{N\beta}\frac{1}{Z'}\frac{\partial Z'}{\partial H} \tag{4.17}$$

Desarrollando la derivada de la función de partición 4.15

$$\frac{\partial Z'}{\partial H} = 2^N N\beta \cosh^{N-1}(\beta(\gamma Jm + H)) \sinh(\beta(\gamma Jm + H)) \tag{4.18}$$

reemplazando este resultado en 4.17 se tiene

$$m = \frac{1}{N\beta}\frac{2^N N\beta \cosh^{N-1}(\beta(\gamma Jm + H)) \sinh(\beta(\gamma Jm + H))}{2^N \cosh^{N}(\beta(\gamma Jm + H))} \tag{4.19}$$

Esto se reduce a

$$m = \tanh(\beta(\gamma Jm + H)) \tag{4.20}$$

La resolución de esta ecuación trigonométrica da lugar al valor de la magnetización crítica frente a los estados ferromagnético y paramagnético del arreglo de espines. Cabe recalcar que este valor es el mismo para cada espín de la red, dada la simetría mencionada inicialmente producto de la propiedad de los campos de ser constantes.

4.2. Extensión del modelo de Ising

El método de aproximación de campo medio es uno muy poderoso para encontrar propiedades más subyacentes que no estan al alcance por métodos exactos convencionales (aún). Dado que se quiere resolver un módelo más general como es el de la máquina de Boltzmann, se requiere reformular las cantidades descritias inicialmente en la sección con la generalidad requerida.

Sea entonces la defínción de energía

$$E(s) = -\sum_{i>j} J_{ij} s_i s_j - \sum_i H_i S_i \tag{4.21}$$

En este caso, estamos asumiendo que existe posibles interacciones no triviales entre cada espín en la red de Ising, asimismo que esas interacciones no son equivalentes entre si, de las misma manera para la influencia que tiene el campo magnético sobre cada espín.

Las formulaciones de la distribución de probabilidad, la función de partición y la energía libre de Hemholtz siguen siendo las mismas a 5.2, 5.3 y 4.4, solo que ahora se tiene

$$H \quad = \quad \{H_i\}_{i=1,..,N} \tag{4.22}$$

A partir de esto se puede desprender que la magnetización puntual para cada espín i es

$$m_i = -\frac{\partial F}{\partial H_i} =< s_i > \tag{4.23}$$

Habiendo formulado el modelo de Ising de forma más general, se pasa a introducirse la aproximación de campo medio. Como en el caso anterior, se define la energía de aporte de un solo espín i

$$E_i = -(\sum_{j \neq i} J_{ij} s_j + H_i) s_i \tag{4.24}$$

y de esto inferir un campo neto sobre el espín i

$$H_i' = \sum_{j \neq i} J_{ij} s_j + H_i \tag{4.25}$$

Se logra apreciar que en este caso, la influencia magnética que ejerce la red sobre el espín i proviene exactamente en general de todos los demás espines que componen la red. Se pasa a reformular entonces 4.25

$$H_i' = \sum_{j \neq i} J_j m_j + H_i + \sum_{j \neq i} J_j (s_j - m_j) \tag{4.26}$$

Haciendo despreciable cada desviación de la media $(s_j - m_j)$ para cada espín j de la red, se tiene nuestra aproximación de campo medio para nuestro caso mas general

$$H'_i = \sum_{j \neq i} J_j m_j + H_i \tag{4.27}$$

y con este resultado se pasa a reformular las demás cantidades, como en el anterior caso

$$E'_i = -H'_i s_i \tag{4.28}$$

$$E'(s) = -\sum_i H'_i s_i = \sum_i E'_i \tag{4.29}$$

De la misma manera que en el caso anterior, en la formulación de campo medio desaparecen las interacciones de forma explicita, haciendo posible formular una función de partición para cada espín

$$Z'_i = \sum_{s_i=+1,-1} \exp(-\beta E'_i) = \exp(\beta H'_i) + \exp(-\beta H'_i) = 2\cosh(\beta H'_i) \tag{4.30}$$

Es decir

$$Z'_i = 2\cosh(\beta(\sum_{j \neq i} J_{ij} m_j + H_i)) \tag{4.31}$$

Por ende, la función de partición del sistema quedaría

$$Z' = \prod_i Z'_i = 2^N \prod_i \cosh(\beta(\sum_{j \neq i} J_j m_j + H_i)) \tag{4.32}$$

a partir de aqui se pasa a encontrar la ecuación para la magnetización crítica reemplazando 4.32 en 4.23 teniendo en cuenta la fórmula de la energía libre de Hemholts 4.16

$$m_i = -\frac{\partial F'}{\partial H_i} = \frac{1}{\beta}\frac{\partial \log Z'}{\partial H_i} = \frac{1}{\beta}\frac{1}{Z'}\frac{\partial Z'}{\partial H_i} \tag{4.33}$$

La derivada de la función de partición respecto al campo magnético sería

$$\frac{\partial Z'}{\partial H_i} = 2^N \sinh(\beta(\sum_{k\neq i} J_{ik}m_k + H_i)) \prod_{j\neq i} \cosh(\beta(\sum_{l\neq j} J_{jl}m_l + H_j)) \tag{4.34}$$

Reemplazando este resultado en 4.33 se obtiene

$$m = \frac{2^N \sinh(\beta(\sum_{k\neq i} J_{ik}m_k + H_i)) \prod_{j\neq i} \cosh(\beta(\sum_{l\neq j} J_{jl}m_l + H_j))}{2^N \prod_i \cosh(\beta(\sum_{j\neq i} J_j m_j + H_i))} \tag{4.35}$$

Esto se reduce a

$$m_i = \tanh(\beta(\sum_{j\neq i} J_{ij}m_j + H_i)) \tag{4.36}$$

Este ultimo resultado define aproximadamente el valor de la magnetización puntual crítica para cada espín, de la misma manera que lo hace el caso anterior más simple.

Capítulo 5

Campo Medio en la Máquina de Boltzmann

Dado ya el formalismo de campo medio para el modelo de Ising y a partir de la equivalencia de este modelo con el de la red de Hopfield, se puede pasar a desarrollar este formalismo sobre la máquina de Boltzmann y describir la ventajas sobre este enfoque.

5.1. Equivalencia entre el modelo de Ising y la red de Hopfield

Los desarrollos presentados hacen evidente la semejanza entre la energía entre ambas estructuras, así como el que ambas esten descritas estadísticamente con la misma distribución de probabilidad. Por ende podemos reescribir las definiciones y resultados encontrados a partir de nuestro modelo general de Ising bajo el contexto de nuestra red neuronal. Se reitera entonces que la energía de la máquina de Boltzmann se describe

$$E(s) = -\sum_{i>j} w_{ij} s_i s_j - \sum_i b_i s_i \tag{5.1}$$

así como su distribución

$$p(s) = \frac{\exp(-\beta E(s)}{Z} \tag{5.2}$$

$$Z = \sum_{s} \exp(-\beta E(s)) \tag{5.3}$$

Bajo estas definiciones, se puede reescribir los resultados de campo medio bajo este nuevo contexto

$$E'(s) = -\sum_{i} (\sum_{j \neq i} w_{ij} m_j + b_i) s_i \tag{5.4}$$

$$m_i = \tanh(\beta(\sum_{j \neq i} w_{ij} m_j + b_i)) \tag{5.5}$$

Siendo $m_i =< s_i >$ el valor medio de la activación de la neurona i. Notese como los terminos que se han cambiado para adaptarse a nuestro modelo de red son los de interacción y los de campo magnético por los de pesos sinápticos y los de sesgos respectivamente.

5.2. Teoría de campo medio en la dinámica neuronal

Teniendo en cuenta que la probabilidad de activación de una neurona i se describió

$$p(s_i = +1) = \frac{1}{1 + \exp(-\beta \Delta E_i)} \tag{5.6}$$

Siendo

$$\Delta E_i = E(s_i = -1) - E(s_i = +1) \tag{5.7}$$

Reemplazando 5.4 en 5.7 tenemos

$$\Delta E_i' = 2(\sum_{j \neq i} w_{ij} m_j + b_i) \tag{5.8}$$

La probabilidad de activación de una neurona queda definida por la aproximación

$$p(s_i = +1) = \frac{1}{1 + \exp(-2\beta(\sum_{j \neq i} w_{ij} m_j + b_i))} \tag{5.9}$$

Reemplazando en esta expresion la formulación de activación media 5.5 se tiene que

$$p(s_i = +1) = \frac{1 - m_i}{2} \tag{5.10}$$

Dado que m_i es una constante, y según 5.5, depende deterministicamente de la activación media de las demas neuronas, la probabilidad de activación al poder ser descrita de esta manera, podra cambiarse por una función determinista de activación equivalente a 5.5. Con nuestro resultado se pasa a definir nuestro algoritmo de activación neuronal para nuestra red

Algorithm 2: Método Determinista en convergencia de equilibrio térmico en una máquina de Boltzmann simple para un valor de temperatura fijo

Se inicializa aleatoriamente array de valores de activación "medios" $m \in \{1, -1\}^N$;
Se fija un valor de temperatura T;
Para una configuración dada de pesos w y sesgos b se define una función de energía $E_{w,b}$;
for $t >> 1$ **do**
    Se elige aleatoriamente un nodo m_i;
    Se actualiza el nodo elegido $m_i : \tanh(\beta(\sum_{j \neq i} w_{ij} m_j + b_i))$;
end

Se ha de notar que este método es determinista: no recurre a una función generadora de números randomicos. Asimismo, cada neurona guarda un valor de activación evaluado entre $[-1, +1]$, a diferencia de la versión original de la red la cual guardaba en sus neuronas valores binarios[1].

5.3. Teoría de campo medio en la dinámica sináptica

El efecto que tiene esta aproximación sobre los productos cruzados de las activaciones de las neuronas parte de entender lo siguiente

$$\begin{aligned} < s_i s_j > &= < (m_i + (s_i - m_i))(m_j + (s_j - m_j)) > \\ &= m_i m_j + < m_i(s_j - m_j > + < (s_i - m_i) m_j > \end{aligned} \tag{5.11}$$

Los terminos que presentan diferencias entre los valores puntuales y medios de las activaciones se hacen cero por la aproximación, dejando[2]

$$< s_i s_j > \approx m_i m_j \tag{5.12}$$

Este resultado modifica la regla de aprendizaje dejandola en su totalidad no un proceso estocástico en si mismo, sino que el conjunto de operaciones involucradas en la red son ahora deterministas (al reemplazar la estocasticidad por no linealidad). La regla de aprendizaje entonces queda

$$\frac{dw_{ij}}{dt} = \gamma\beta((m_i m_j)_D - (m_i m_j)_p) \tag{5.13}$$

Los subíndices D y p indican que esas medias que siguen la *data* y la distribución de la red respectivamente. El algoritmo de plasticidad solamente es el siguiente

Algorithm 3: Método Determinista de aprendizaje en BM usando método de campo medio para aproximar las correlaciones de las activaciones neuronales

Se tiene un conjunto $D \subset -1, 1^N$ cuya distribución se quiere aprender;
Se tiene un array de pesos $w_{ij_{i,j=1...N}}$ que parametrizan mi BM;
Asimismo, se tiene una configuración media de activaciones neuronales $m \in [-1, 1]^N$ representando un estado de equilibrio según mis parametros de pesos;
Se establece un valor de taza de aprendizaje γ;
for i *in* $(1, ..., N)$ *and* j *in* $(1, ..., N)$ *and* $i \neq j$ **do**
    Se elijen m_i y m_j de m, asi como se calcula promedia todos los s_i y s_j de todos los $s \in D$;
end

Capítulo 6

Prueba de aprendizaje sobre MNIST y Discusiones

6.1. Detalles de la prueba

Se uso para este experimento de prueba una red restringida de Boltzmann con aproximación de campo medio. La base de datos objetivo que se uso fue el MNIST, la cual esta conformada por 60000 imagenes de 28×28 pixeles de números del 0 al 9 escritos a mano, en escala de grises, como se muestra en la figura 6.1

La arquitectura de la red se describe con 784 unidades visibles y 100 ocultas. Las unidades visibles se definieron con valores admisibles en el rango real de $[-1, 1]$, mientras que las unidades ocultas se dejaron binarias. La temperatura de la red se fijo en $T = 1$ por convención, tanto los pesos sinápticos con los sesgos se inicializaron aleatoriamente de forma uniforme en el rango de $[-0{,}5, 0{,}5]$. Asimismo se hizo más pruebas con diferentes valores de temperatura $\{0{,}1, 0{,}5, 2, 5\}$ y se midio el cambio de la función de error cuadrático medio para cada valor de temperatura

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n}(s_t - \hat{s}_t) \tag{6.1}$$

Siendo n el número de muestras usadas para el calculo de error, s una configuración generada por la red y $\hat{s}$ la configuración objetivo.

Por ultimo, se ejecuto una red de Boltzmann restringida de solo unidades binarias, con 784×8 unidades visibles y 144 unidades ocultas. Se uso una version binarizada

del MNIST: cada pixel fue extendido un un array de 8 elementos, los cuales son el reusltado de convertir el valor del pixel (en escala de grises) al sistema binario (ej. 50 $\rightarrow$ 00110010).

La red se ejecuto en un ordenador con distribución Ubuntu 16.04 de arquitectura x64, con una CPU Intel® Core™ i5-7300HQ CPU @ 2.50GHz $\times$ 4.

Se puso a aprender a la Máquina de Bolzman, solicitándole que luego de su aprendizaje escriba aleatoriamente un conjunto de los números aprendidos, los cuales esta generó graficamente y están representados en 6.2. El código en Python usado esta en el repositorio https://github.com/PoppinElo/meanFieldRBM.

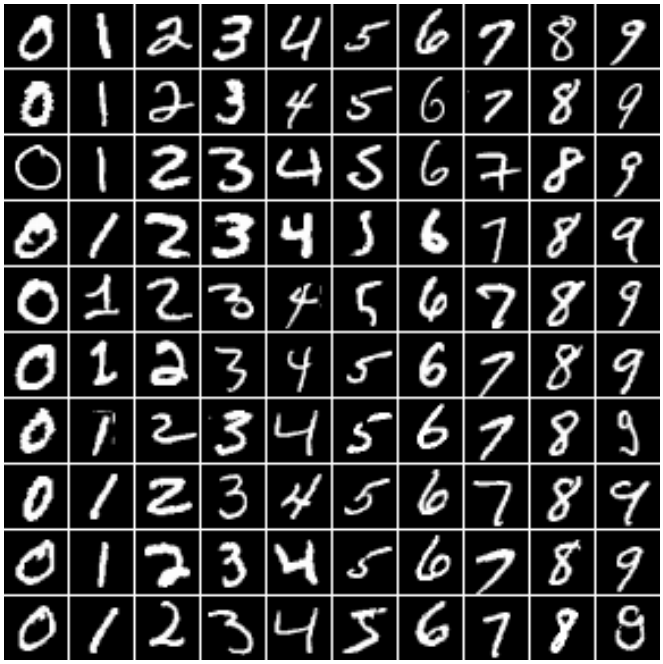

Figura 6.1: Imagenes de la base de datos del MNIST

6.2. Discusión de los resultados

Se logra apreciar que la red a sido capaz de aprender la distribución de las imagenes, notandose en la representación de esa distribución aprendida por la red en las imagenes generadas por esta, así como en la caida de la función de perdida (basada en error cuadrático medio). Sin embargo, el proceso de aprendizaje para obtener ese resultado tomó 4min20seg de tiempo transcurrido. También el hecho de que existe ruido en las

imagenes generadas, como se puede apreciar en la figura 6.2.

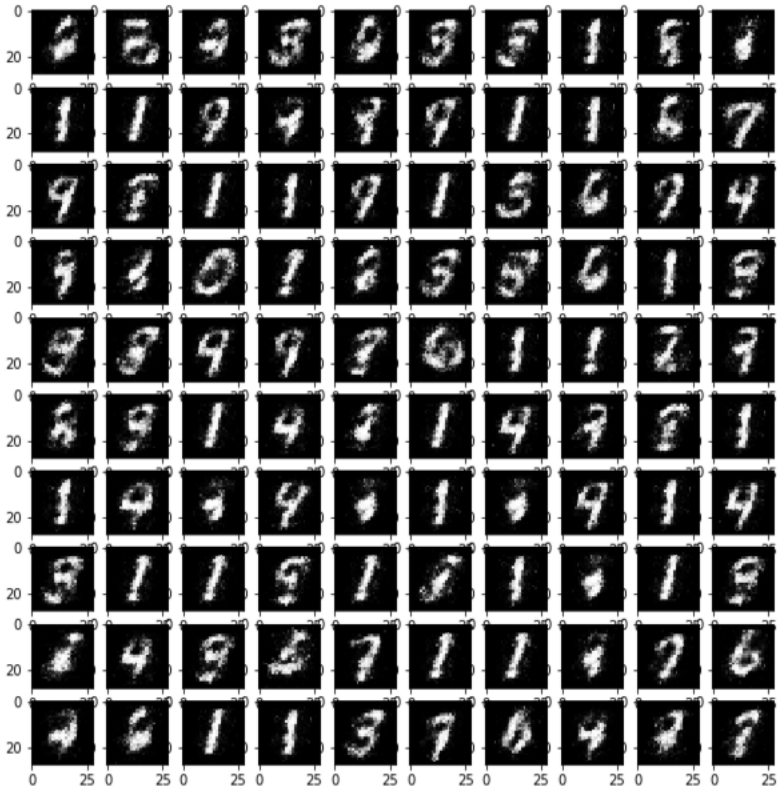

Figura 6.2: Imagenes de números generadas por la red

Según la gráfica 6.3, durante los primeros pasos temporales, el error presenta valores altos para temperaturas bajas y visceversa, pero decaen rapidamente en todos los casos, para luego tender a un promedio estable de 50 para temperaturas bajas, a diferencia del caso de temperatura más alta donde el error no decae a valores tan bajos como el resto sino además la varianza es tambien alta. Esto ocurre ya que la temperatura alta produce tazas de cambio m'as grandes por paso temporal, facilitando la minimización de funciones objetivo durante primeros pasos temporales, pero luego impidiendo la convergencia haciendo que el sistema oscile alrededor del punto de equlibrio. Es pertinente aclarar que un bajo valor en el error cuadrático medio no es un indicador completo de aprendizaje: como se observa en 6.4, al presentar una baja temperatura, la red le ha de costar mucho más tiempo en converger a un equilibrio energético coordinadamente con el proceso de aprendizaje. Por otro lado, una temperatura alta permite una convergencia energética más rapida pero equivocada dado el mismo conflicto con el proceso de

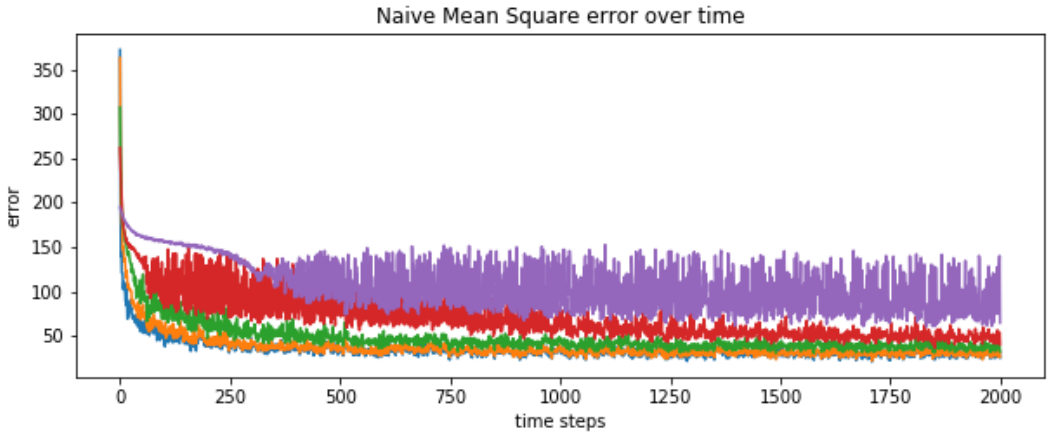

Figura 6.3: Cambio del error calculado para diferentes valores de temperatura
Las lineas de colores azul, anaranjado, verde, rojo y violeta corresponden a las temperaturas 0.1, 0.5, 1, 2 y 5 respectivamente

aprendizaje mencionado para el caso de temperatura baja

La prueba con la versión binaria tomó más de 5 horas en tiempo de ejecución. Esto demuestra la superioridad en economía de computación de la red de Boltzmann con aproximación de campo medio frente a una red de unidades binarias.

Para máquinas de Boltzman vainilla [1], se requiere que se binarice los datos de entrada antes de usarlos para entrenar. Si estos datos son reales, al binarizarlos se estaría contando con 32-64 valores de los que se tenian antes de binarizar, y se requeririan 32x32-64x64 más conexiones sinápticas requeridas en una red de Boltzmann como la usada. Esto implica menos variables a usar. Asimismo, una función que genera números aleatorios esta realizando por debajo decenas de operaciones para lograr una correlación nula entre el valor de entrada y salida; por ende, si esta es invocada muchas veces en un programa, eso implicara un coste alto en tiempo de ejecución. Al aplicar la aproximación de campo medio, ya no dependemos de ninguna función aleatoria al trabajar directamente con las supuestas medias y volver las operaciones deterministas, por lo que se reduce el tiempo de ejecución en comparación con una red binaria. Juntas ambas propiedades de nuestra red permite que esta pueda usarse para aprender de datos no binarios sino reales.

[1] Una red vainilla de un tipo es la que tiene la estructura más simple y fundamental en su definición. En este caso, una máquina de Boltzmann binaria estocástica

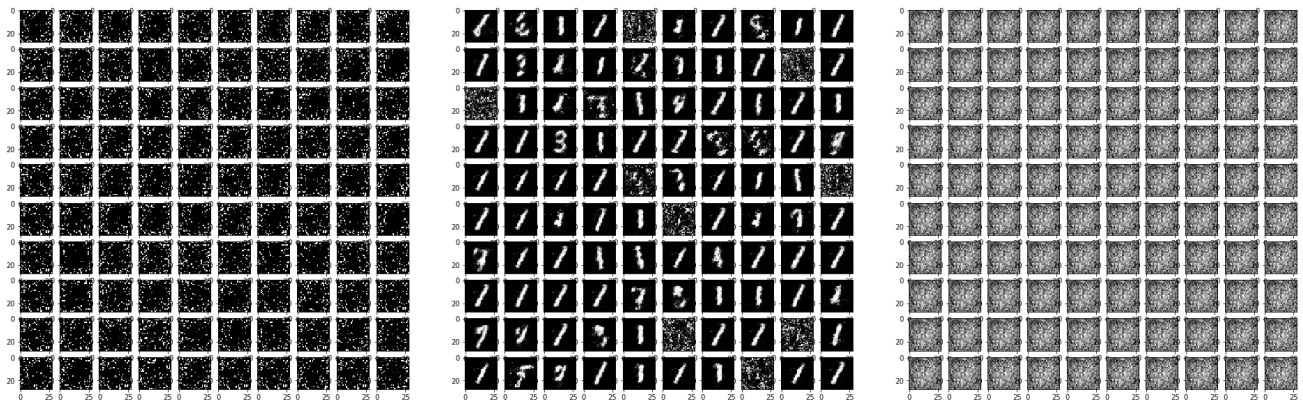

Figura 6.4: Imagenes generadas a diferentes temperaturas

La imagen de la izquierda es generada por la red con temperatura 0.1, la imagen del medio con temperatura 2 y la de la derecha con temperatura 5.

6.3. Sugerencias

A pesar de la optimización matemática y estructural descrita de la red, el tiempo de ejecución y la claridad de las imagenes generadas aun estan debajo de los estandares modernos en deep learning, pero puede hacerse más corto implementando encima estas posibles alternativas, tanto computacionales como físico-matemáticas.

- Computacionales:
  - Usar contraste de divergencia, es decir, para cada inicio de ciclo de aprendizaje usar la ultima configuración a la que se llego al equilibrio con los valores de pesos anteriores, esto daría un poco menos de tiempo de ejecución.
  - Imponer, recurriendo al teorema del límite central, que las distribuciones condicionales no son sigmoides sino gaussianas, esto permitiría que se filtre el ruido[5].
  - Implementar el código no en Python sino en C, o mejor, C CUDA para ejecutrar operaciones matriciales con los pesos en paralelo.
- Físico matemáticas:
  - Usar valores de temperatura fuera de la convención, ya sea constantes o que vayan variando a medida que pasan los ciclos de aprendizaje (por ejemplo, el recocido simulado).
  - Definir de forma distinta la energía, ya sea que incluya fielmente las corre-

laciones entre neuronas de la red, de forma directa o indirecta (por ejemplo, un modelo de energía basado en la regla hebbiana de Oja)[10].

- Definir de forma distinta la función de perdida de la cual se extraera su gradiente (podría complementarse la entropía con un termino que dependa de forma más contrastante la función de partición, o definir una función de energía libre especialmente para el aprendizaje)[5].
- Definir un método de campo medio que considere un grado de exactitud de un grado, de esta manera la covarianza entre neuronas (de la cual depende la regla de aprendizaje) no sera nulificable mediante esta aproximación y podra ajustarse mejor a la dirección de la componente principal.
- Definir una arquitectura sobre la cual se pueda descubrir una propiedad de simétria que permita una simplificación de la dinámica de activación o de plasticidad, ya sea definiendo un *gauge* y asi poder definir una energía libre para la función perdida o una simplificación de las ecuaciones de activación y aprendizaje ya conocidas.

Capítulo 7

Conclusiones

Se concluye que, a partir de nuestros resultados basados en teoría termodinámica de Ising, que la teoría de campo medio aplicada en la máquina de Boltzmann mejora su rendimiento tanto en capacidad de gestión de información así como del aumento de rango de problemas que una red de Boltzmann es capaz de enfrentar, todo esto reduciendo la cantidad de operaciones por unidad de tiempo con respecto al que el modelo fundamental de la red le sería necesario usar.

Asimismo, frente a los resultados obtenidos por nuestro método recogido del paradigma de la física, se recoge la alta posibilidad de obtener mejores resultados en teoría de implementación de la máquina de Boltzmann usando otros métodos de cálculo basados en termodinámica estadística.

Apéndice A

Modélo de Respuesta Neuronal

El perceptrón es uno de los primeros por no decir el primer modelo de neurona artifical inspirado de la biología. A pesar de su simpleza, la eficacia que ofrece como unidad mínimica de computación[1] se puede intuir de la eficacia con las que las neuronas biológicas, de las que se inspira el concepto de perceptrón, computan información.

A.1. Modelo de Integra-y-Dispara con Fuga

Los modelos del tipo Integra-y-Dispara intentan describir el comportamiento de la membrana celular de una neurona ante estimulos externos, ya sea aquellos provienientes de otras neuronas o de algun agente ajeno al sistema biológico.

Basicamente, el modelo simple de Integra-y-Dispara con Fuga indica que la membrana celular de la neurona conduce las señales eléctricas que inciden sobre ella como si fuera un circuito pasa-baja.

Sea u_i el potencial de membrana de la neurona, I_i es la corriente que incide sobre la membrana, R y C la resistencia y la capacitancia de la membrana respectivamente. La ecuacion de Integra-y-Dispara con Fuga es[10]

$$\tau_m \frac{du_i}{dt}(t) = -(u_i(t) - u_o) + RI_i(t) \tag{A.1}$$

[1] Entiendase computación como el proceso de gestionar información; esto es el almacenar, transimitir y manipular información

Con las condiciones de ligadura

$$u_i(t = t^-_{disparo}) = u_{disparo} \text{ cuando } u_i(t = t^+_{disparo}) = u_{umbral} \tag{A.2}$$

Aquí, $\tau_m = RC$ es el tiempo de membrana de la neurona; esto es, la constante temporal de decaimiento del potencial de membrana cuando no recibe estimulo alguno. u_{umbral} y $u_{disparo}$ son los potenciales de umbral y de disparo respectivamente; el potencial de umbral es el valor de corte el cual provoca que la neurona dispare cuando el potencial de membrana alcanza ese valor, y el potencial de disparo es el valor al cual la membrana se eleva bruscamente al disparar.

Cuando la corriente que incide sobre la membrana proviene de las otras neuronas entonces podemos definir I como

$$I_i(t) = \sum_j w_{ij} S_j(t) \tag{A.3}$$

Siendo S_j la serie de pulsos electricos que dispara la neurona j (sobre la neurona i) en el tiempo, y w_{ij} la fuerza de la sinápsis entre las neuronas i y j.

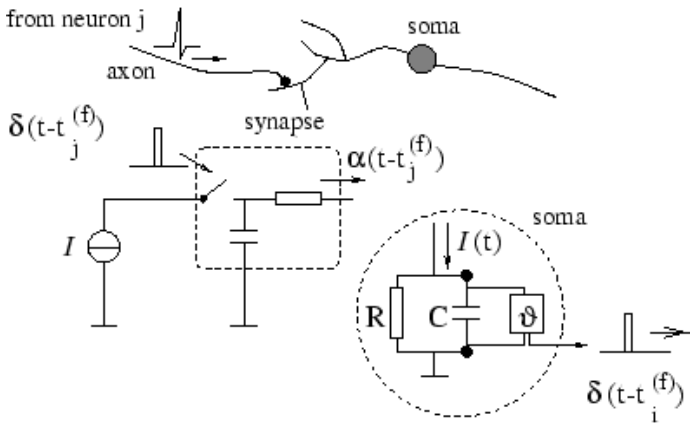

Figura A.1: Descripción de la neurona según el model Integra-y-Dispara[10]

La membrana actua como un filtro pasa baja, regulando el potencial con la que se carga la neurona, de la cual depende la activación de la neurona. La activación, definida por las condiciones de ligadura, se decide en el soma de la neurona.

Teniendo en cuenta A.3, al resolver la ecuación A.1 se obtiene

$$u_i(t) = u_o + \sum_j w_{ij} \frac{R}{\tau_m} \int_{s=0}^{\infty} g(t,s) S_j(s) ds \tag{A.4}$$

Siendo g la función de green de la ecuación A.1

$$g(t,s) = \exp(-\frac{t-s}{\tau})\Theta(t-s) \tag{A.5}$$

Θ es la función escalón

$$\Theta(z) = \begin{cases} 1 & z > 0 \\ 0 & z \leqslant 0 \end{cases} \tag{A.6}$$

Se puede interpretar la integral de green como una media sobre el tiempo de decaimiento

$$< x > (t) = \frac{1}{\tau} \int_{s=0}^{\infty} g(t,s) x(s) ds \tag{A.7}$$

Al hacer que los pesos sinápticos absorvan el valor de la resistencia de la membrana y haciendo $u_o = 0$, la ecuación A.4 queda

$$u_i(t) = \sum_j w_{ij} < S_j > (t) \tag{A.8}$$

Este ultimo resultado, junto con las condiciones de ligadura A.2, es equivalente con la definición de perceptrón descrita en 2.2.

Dado que a partir de un módelo más realista de la dinámica de activación neuronal nos da, en promedio, una descripción del fenomeno de transmisión de señales semejante al perceptrón, la eficiencia del perceptrón como modelo de computación es tambien semejante a la de una neurona biológica para un rango específico de escalas temporales.

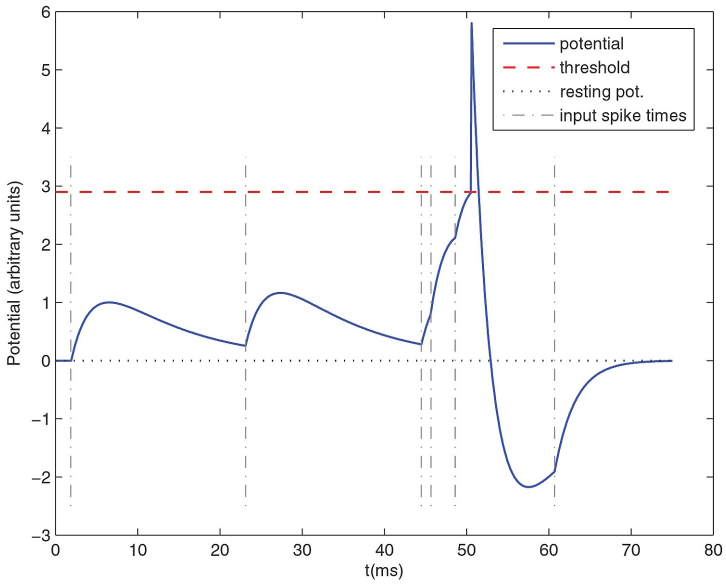

Figura A.2: Gráfica del potencial de membrana según el modelo de Integra-y-Dispara[9]

Las lineas verticales entrecortadas indican los instantes donde aparece una señal de estimulo, la linea entrecortada horizontal roja representa el potencial de umbral. Se aprecia como en presencia de estimulo el potencial se incrementa y como en ausencia el potencial decae. Asimismo, cuando el potencial alcanza el valor de umbral, genera un disparo, siguiendole de un decaimiento de potencial inferior al valor base.

A.2. Plasticidad dependiente de Coincidencia Temporal

Dado que, segun Hebb, la plasticidad sináptica sigue una regla de correlación, es facil entender como, para una secuencia de estimulos en el tiempo, los pesos van a cambiar según las diferencias temporales entre los instantes en que se dan los disparos presinápticos y postsinápticos. La ecuación diferencial del modelo de Plasticidad dependiente de Coincidencia Temporal es la siguiente[10]

$$\frac{dw_{ij}}{dt}(t) = S_j(t)\int_0^\infty A_-(w_{ij})g_-(t,s)S_i(s)ds + S_i(t)\int_0^\infty A_+(w_{ij})g_+(t,s)S_j(s)ds \quad \text{(A.9)}$$

Siendo $A_\pm$, $g_\pm$ las amplitudes y las funciones de green respectivamente. Los subíndices $+$ y $-$ indican potenciación y depresión respectivamente. Se pasa a definir ahora una función semejante a la correlación segun las escalas temporales

$$< xy >_{\pm} = \int \int g_{\pm}(t,s)x(t)y(s)dtds \tag{A.10}$$

Con esta definición de correlación, y asumiendo que el valor de las amplitudes varían muy lentamente en general, la evolución de los pesos sinápticos respecto al tiempo se escribe

$$\Delta w_{ij}(t) = w_{ij}(t) - w_{ij}(0) \approx A_{-} < S_i S_j >_{-} (t) + A_{+} < S_i S_j >_{+} (t) \tag{A.11}$$

Se aprecia como la regla clásica de Hebb 2.4 es semejante a esta ultima ecuación, solo que en este caso se da bajo dos escalas temporales distintas.

Bibliografía

[1] J. R. Anderson C. Peterson. "A Mean Field Theory Learning Algorithm for Neural Networks". En: *Complex Systems* 1 (1987), págs. 995-1019.

[2] D. H. Ackley G. E. Hinton T. J. Sejnowski. "A learning algorithm for Boltzmann Machines". En: *Cognitive Science* 9 (1985), págs. 147-169.

[3] D. O. Hebb. *The Organization of Behavior: A neurophychological theory*. Wiley Press, 1949.

[4] J. J. Hopfield. "Neural networks and physical systems with emergent collective computational abilitiies". En: *Proc. Natl. Acad. Sci.* 79 (1982), págs. 2554-2558.

[5] A. Courville I. Goodfellow Y. Bengio. *Deep Learning*. MIT Press, 2016.

[6] G. Parisi. *Statistical Field Theory*. Addison-Wesley, 1988.

[7] J. Feldman R. Rojas. *Neural Networks - A Systematic Introduction*. Springer Press, 1996.

[8] F. Reif. *Fundamentals of Statistical and thermal physics*. McGraw-Hill, 1965.

[9] S. J. Thorpe T. Masqueller R. Guyonneau. "Spike Timing Dependent Plasticity Finds the Start of Repeating Pattern in Continuous Spike Trains". En: *PLoS ONE* (2008).

[10] R. Naud W. Gerstner W. M. Kisler y L. Paninski. *Neural Dynamics*. Cambridge University Press, 2014.