

Máquinas de Boltzmann

Verónica E. Arriola-Rios

Facultad de Ciencias, UNAM

17 de mayo de 2023



Intro

- 1 Intro
- 2 Equilibrio térmico
- 3 Usos
- 4 Aprendizaje
- 5 Máquinas de Boltzmann Restringidas

Temas

1 Intro

- Limitaciones de Hopfield
- Definición

Limitaciones

- Las redes de Hopfield memorizan información en sus mínimos.
- Dado que la activación de sus neuronas es determinista, partiendo de la misma configuración inicial siempre se llega al mínimo local más cercano, es decir, a la misma memoria.
- Para encontrar otros mínimos de energía, más aún, el mínimo global, se puede agregar ruido a la activación, que le permita escapar.

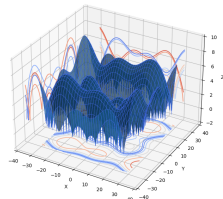


Figura: Como ilustración: función con muchos mínimos.

Temas

1 Intro

- Limitaciones de Hopfield
- Definición

Energía

- La probabilidad de que la máquina se encuentre en un cierto estado global depende de la energía asociada a cada configuración conjunta de sus unidades visibles \vec{v} e invisibles \vec{h} .

$$-E(\vec{s}) = \sum_i s_i b_i + \sum_{i < j} s_i s_j w_{ij} \quad (1)$$

$$\begin{aligned}
 -E(\vec{v}, \vec{h}) = & \underbrace{\sum_{i \in v} v_i b_i}_{\text{sesgo visibles}} + \underbrace{\sum_{k \in h} h_k b_k}_{\text{sesgo ocultas}} + \\
 & \underbrace{\sum_{i < h} v_i v_j w_{ij}}_{\text{conexión entre visibles}} + \underbrace{\sum_{i, k} v_i h_k w_{ik}}_{\text{visibles ocultas}} + \underbrace{\sum_{k < l} h_k h_l w_{kl}}_{\text{ocultas ocultas}} \quad (2)
 \end{aligned}$$

con b los sesgos y w los pesos.

Activación logística

- Para encontrar mínimos globales de la energía se usa **recocido simulado**.
- La función de activación es aleatoria.
- Para decidir si se activa ($s_i = 1$) o no ($s_i = 0$) una neurona, se utiliza una función de activación logística probabilista, dependiente de una temperatura T y el cambio de energía ΔE_i producido por la activación de la neurona i .

$$p(s_i = 1) = \frac{1}{1 + e^{-\frac{\Delta E_i}{T}}} \quad (3)$$

$$\Delta E_i = E(s_i = 0) - E(s_i = 1) = b_i + \sum_j s_j w_{ij} \quad (4)$$

- Para obtener una muestra se utiliza un valor aleatorio $x \in [0, 1]$, si $x < p(s_i = 1)$ la neurona se activa, si no, se apaga.

Probabilidad de activación

Entre mayor sea la temperatura T , mayor es el ruido que afecta al movimiento de las partículas y los niveles de energía E se encuentran más cerca unos de otros, lo cual facilita escapar de mínimos locales.

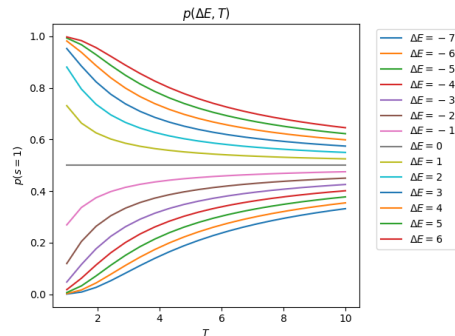


Figura: Entre más se reduzca la energía al activar la neurona, mayor probabilidad de activarla; entre más se reduzca al apagarla, mayor probabilidad de apagarla. Cuando T es alta, los cambios son más aleatorios.

Equilibrio térmico

- 1 Intro
- 2 Equilibrio térmico
- 3 Usos
- 4 Aprendizaje
- 5 Máquinas de Boltzmann Restringidas

Temas

2 Equilibrio térmico

- Definición
- Configuraciones
- Muestreo Monte Carlo

Equilibrio térmico

Definición

Se dice que la red se encuentra en *equilibrio térmico* a temperatura T si la distribución de probabilidad sobre los estados posibles ya no cambia con el tiempo.

También se puede entender en términos de ensembles:

- Imaginemos que se tienen varias copias de la misma red, con los pesos fijos, de tal modo que haya **más sistemas** que **configuraciones**.
- En el estado de equilibrio, se tendrá el mismo número de sistemas en cada configuración posible para todo tiempo subsecuente, aunque entre un tiempo y otro cada sistema cambie sus valores de activación.

Histograma

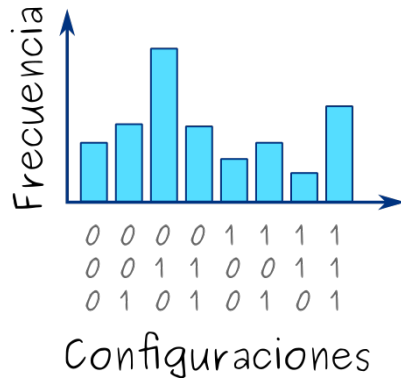


Figura: Número de ejemplares por configuración (o estado) en un ensemble de máquinas de Boltzmann.

Temas

2 Equilibrio térmico

- Definición
- Configuraciones
- Muestreo Monte Carlo

Configuraciones

- La probabilidad de que una máquina de Boltzmann en equilibrio térmico (suponiendo $T = 1$) se encuentre en un estado depende de su energía como:

$$p(\vec{v}, \vec{h}) \propto e^{-E(\vec{v}, \vec{h})} \quad (5)$$

concretamente:

$$p(\vec{v}, \vec{h}) = \frac{e^{-E(\vec{v}, \vec{h})}}{\sum_{\vec{u}, \vec{g}} e^{-E(\vec{u}, \vec{g})}} \quad (6)$$

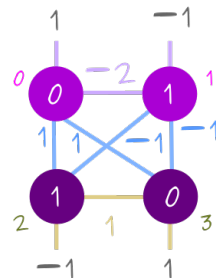
- La probabilidad de que un vector determinado haya sido generado se obtiene marginalizando sobre todos los valores posibles de activación de las neuronas ocultas:

$$p(\vec{v}) = \frac{\sum_{\vec{h}} e^{-E(\vec{v}, \vec{h})}}{\sum_{\vec{u}, \vec{g}} e^{-E(\vec{u}, \vec{g})}} \quad (7)$$

\vec{v}	\vec{h}	$-E$	e^{-E}	$p(\vec{s}) = p(\vec{v}, \vec{h})$	$p(\vec{v})$
00	00	0	1.00	0.011	0.0476
00	01	-1	0.37	0.004	
00	10	1	2.72	0.031	
00	11	-2	0.14	0.002	
01	00	1	2.72	0.031	0.2652
01	01	-1	0.37	0.004	
01	10	3	20.09	0.226	
01	11	-1	0.37	0.004	
10	00	-1	0.37	0.004	0.0359
10	01	-3	0.05	0.001	
10	10	1	2.72	0.031	
10	11	-3	0.05	0.001	
11	00	1	2.72	0.031	0.6514
11	01	-2	0.14	0.002	
11	10	4	54.60	0.615	
11	11	-1	0.37	0.004	

Ocultas

Visibles



Temas

- 2 Equilibrio térmico
 - Definición
 - Configuraciones
 - Muestreo Monte Carlo

Monte Carlo

- $p(v, h)$
 - Crear la red.
 - Actualizarla hasta que llegue a su equilibrio térmico.
 - Muestrear cuántos pasos permanece en cada una de sus configuraciones.
- $p(h|v)$ *a posteriori*
 - Se fijan los valores de las unidades visibles v .
 - Se actualizan las unidades ocultas hasta alcanzar el equilibrio térmico.
 - Muestrear.

Dificultades:

- No existe un método para saber, a priori, cuántas actualizaciones requiere una red antes de alcanzar el equilibrio térmico, podrían ser de cientos a miles.
- Al ir actualizando neurona por neurona el proceso para pasar de una región del espacio de estados a otra es muy lento.

Usos

- 1 Intro
- 2 Equilibrio térmico
- 3 Usos
- 4 Aprendizaje
- 5 Máquinas de Boltzmann Restringidas

Temas

3 Usos

- Modelaje probabilista
- Aplicaciones

Cada máquina es un modelo

- Una máquina de Boltzmann representa un **modelo** que asigna una **probabilidad** a cada **vector binario** posible. Por lo que es posible, dado un vector, evaluar la probabilidad de que haya sido generado por este modelo.
- Su entrenamiento tiene por objetivo ajustar el modelo de tal forma que asigne probabilidades acordes a un conjunto de entrenamiento de vectores binarios.
- Es posible entrenar varias máquinas para representar modelos distintos. Dado un vector, será posible determinar la probabilidad **a posteriori** de que haya sido generado por un modelo en particular utilizando la regla de Bayes:

$$p(\text{Modelo}_i|\text{datos}) = \frac{p(\text{datos}|\text{Modelo}_i)}{\sum_j p(\text{datos}|\text{Modelo}_j)} \quad (8)$$

Temas

3 Usos

- Modelaje probabilista
- Aplicaciones

Posibles aplicaciones

- Varias máquinas de Boltzmann entrenadas se pueden utilizar para resolver problemas de clasificación, pues modelan la distribución de probabilidad que asocia los ejemplares con la representación de sus clases:
 - Dado un documento y un vector de características binarias, extraídas de él, determinar qué tipo de documento es.
 - Dado un vector de características de audio, identificar la palabra que fue pronunciada.
- Una máquina de Boltzmann también puede identificar la presencia de datos atípicos, pues aprende a asociar probabilidades altas a estados conocidos y asignará una probabilidad baja a un estado que nunca antes había visto. Por ejemplo:
 - En un sistema de vigilancia, entrenado con datos recolectados en situaciones normales, detectar la aparición de situaciones atípicas cuando se presentan datos, cuya probabilidad de ocurrencia dado el modelo conocido es muy baja.

Aprendizaje

- 1 Intro
- 2 Equilibrio térmico
- 3 Usos
- 4 Aprendizaje**
- 5 Máquinas de Boltzmann Restringidas

Temas

- 4 Aprendizaje
 - Entrenamiento
 - Partículas

Aprendizaje I

- Se utiliza aprendizaje **no supervisado**.
- El objetivo es **maximizar** la probabilidad de obtener los vectores \vec{v}_i en el conjunto de entrenamiento a partir del modelo \mathcal{M} representado por la máquina, esto es la *verosimilitud* L .
Utilizando la definición de verosimilitud en Koller y Friedman 2009:

$$L(\mathcal{M} : V) = p(V|\mathcal{M}) = p(V)_{\text{Modelo}} \quad (9)$$

- Es decir, si se deja evolucionar a la máquina N veces, la cantidad de veces que el vector visible coincide con los datos de entrenamiento debe ser máximo.

Aprendizaje II

- Si asumimos la generación de cada ejemplar es independiente de los otros vectores, entonces esta probabilidad es el producto de la probabilidad de generar cada uno de estos vectores independientemente.

$$p(V|\mathcal{M}) = \prod_{\mathcal{M}} p(\vec{v}_i : \mathcal{M}) \quad (10)$$

$$p(V)_{\text{Modelo}} = \prod_i p(\vec{v}_i)_{\text{Modelo}} \quad (11)$$

- Esto es equivalente a maximizar la suma de las log-verosimilitudes.

Gradiente I

- Dada la energía E

$$\begin{aligned}
 -E(\vec{v}, \vec{h}) = & \underbrace{\sum_{i \in v} v_i b_i}_{\text{sesgo visibles}} + \underbrace{\sum_{k \in h} h_k b_k}_{\text{sesgo ocultas}} + \\
 & \underbrace{\sum_{i < h} v_i v_j w_{ij}}_{\text{conexión entre visibles}} + \underbrace{\sum_{i, k} v_i h_k w_{ik}}_{\text{visibles ocultas}} + \underbrace{\sum_{k < l} h_k h_l w_{kl}}_{\text{ocultas ocultas}}
 \end{aligned}$$

- Recordemos que:

$$p(\vec{v}) = \frac{\sum_{\vec{h}} e^{-E(\vec{v}, \vec{h})}}{\sum_{\vec{u}, \vec{g}} e^{-E(\vec{u}, \vec{g})}}$$

Gradiente II

- La probabilidad de observar un patrón, dado el modelo actual es:

$$p(v) \propto e^{-E(v,h)} \Rightarrow \log p(v) \propto -E(v,h) \quad (12)$$

- Si escribimos genéricamente como s a los valores de activación, independientemente de si son ocultas o visibles:

$$E(s) = - \sum_i s_i b_i - \sum_{i < j} s_i s_j w_{ij} \quad (13)$$

- Obsérvese que la derivada con respecto a un peso es:

$$-\frac{\partial E}{\partial w_{ij}} = s_i s_j \quad (14)$$

Aprendizaje Hebiano

Si disparan juntas, se conectan.

El proceso para llegar al equilibrio térmico realiza un trabajo análogo a la propagación hacia atrás, al transmitir la información acerca de los pesos entre las diferentes unidades.

- Dado que $p(\vec{v}, \vec{h}) \propto e^{-E(\vec{v}, \vec{h})}$ cuando se ha alcanzado el **equilibrio térmico**, para maximizar la verosimilitud se aplica la regla:

$$\frac{\partial \log p(\vec{v})}{\partial w_{ij}} = \langle s_i s_j \rangle_{\vec{v}} - \langle s_i s_j \rangle_{\text{modelo}} \quad (15)$$

donde $\langle \rangle$ denota el valor esperado (promedio pesado) y \vec{v} es el vector de valores para las neuronas visibles, para el ejemplar que se quiere memorizar.

Considerando todos los vectores de entrada \vec{v} (datos), los pesos de la red se ajustan según:

$$\Delta w_{ij} = \langle s_i s_j \rangle_{\text{datos}} - \langle s_i s_j \rangle_{\text{modelo}} \quad (16)$$

- El primer término es como el de aprendizaje de la red de Hopfield e indica la frecuencia con la que dos neuronas conectadas por una arista se activan juntas.
- El segundo término indica cómo remover los mínimos espúreos (*fase del sueño*).
- Para estimar estos valores esperados se utilizarán partículas.

“Fire together, wire together”

¡Aprenden correlaciones!

Temas

- 4 Aprendizaje
 - Entrenamiento
 - Partículas

Partículas y entrenamiento

Las partículas se utilizan para estimar las estadísticas necesarias para entrenar una máquina de Boltzmann, de tal modo que los mínimos de su energía correspondan con los patrones que se desea que memorice.

Partículas y partículas de fantasía

- Se debe tener una bolsa de *partículas* por cada **vector de entramiento** (patrón que se desea memorizar).

Definición

Las *partículas* (a secas) son vectores que se utilizan para estimar $\langle s_i s_j \rangle_{\text{datos}}$, fijando los valores de las **neuronas visibles** a los valores de los patrones que se desea memorizar y variando las neuronas ocultas.

- Una bolsa de partículas para la red completa, estas son las *partículas de fantasía*, de ellas se contabiliza $\langle s_i s_j \rangle_{\text{modelo}}$.

Definición

Las *partículas de fantasía* son vectores con el estado de la máquina de Boltzmann donde **todas** las neuronas pueden cambiar su valor.

Entrenamiento (Neal 1992)

- Las partículas se usan de acuerdo al siguiente algoritmo:
 - Se llevan al punto de equilibrio térmico, modificando el valor de activación de una neurona a la vez.
 - Fase *positiva*: Por cada caso de entrenamiento, actualizar secuencialmente las unidades ocultas de cada partícula en la bolsa correspondiente, manteniendo fijos los valores de las neuronas visibles. Para cada pareja de neuronas conectadas, promediar $\langle s_i s_j \rangle$.
 - Fase *negativa*: Actualizar secuencialmente todas las unidades de las partículas de fantasía. Para cada pareja de neuronas conectadas, promediar $\langle s_i s_j \rangle$ sobre todas las partículas de fantasía.
 - Se actualiza w_{ij} usando los estimados para $\langle s_i s_j \rangle_{\text{datos}}$ y $\langle s_i s_j \rangle_{\text{modelo}}$.

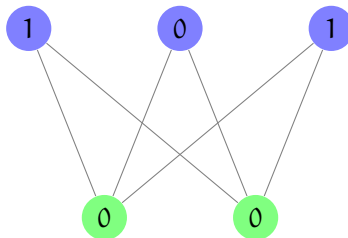
$$\Delta w_{ij} = \langle s_i s_j \rangle_{\text{datos}} - \langle s_i s_j \rangle_{\text{modelo}} \quad (17)$$

- Se retoman las partículas en los valores donde quedaron y se hacen evolucionar hasta alcanzar el nuevo equilibrio térmico, con los valores actualizados de los pesos.
- Necesita menos pasos para llegar nuevamente al equilibrio. Repetir.

Máquinas de Boltzmann Restringidas

- 1 Intro
- 2 Equilibrio térmico
- 3 Usos
- 4 Aprendizaje
- 5 Máquinas de Boltzmann Restringidas



Máquinas de Boltzmann Restringidas



- Si se fijan los valores \vec{v} se necesita un sólo paso para alcanzar el equilibrio térmico y calcular $\langle v_i h_i \rangle$, porque las neuronas ocultas son independientes entre sí, al igual que las visibles.

$$p(h_j = 1) = \frac{1}{1 + e^{-(b_j + \sum_{i \in \text{vis}} v_i w_{ij})}} \quad (18)$$

Referencias I

-  Hinton, Geoffrey E. (2007). *Boltzmann machine*. English. URL: http://www.scholarpedia.org/article/Boltzmann_machine.
-  Koller, Daphne y Nir Friedman (2009). *Probabilistic Graphical Models, Principles and Techniques*. MIT Press Cambridge.

Licencia

Creative Commons
Atribución-No Comercial-Compartir Igual

