ENTROPIA INFERENCIAL DE UNA RED NEURONAL

E. Ferrán y R. Perazzo
Departamento de Física (TANDAR), Comisión Nacional de Energía Atómica, Av. Libertador 8250, 1429 Buenos Aires

En la presente comunicación se establece y analiza el concepto de entropía inferencial para redes neuronales sin retroalimentación. Además se muestra la correspondencia entre dicho concepto y la función entropía que se obtiene mediante un análisis mecánico-estadístico de la termodinámica del proceso de aprendizaje de la red.

Consideramos redes estructuradas en capas, con neuronas interconectadas sin retroalimentación. Las interconexiones están dadas por una matriz sináptica J, con la cual la red representza una dada función booleana entre sus entradas y salidas exteriores. Centramos nuestro análisis en la capacidad de generalización de redes que aprenden mediante ejemplos. Suponemos que los parámetros internos que definen la estructura de la red, a excepción de las eficacias sinápticas $J_{ij}^{\lambda\nu}$, permanecen invariables. Es decir, el aprendizaje se realiza buscando la matriz adecuada en el espacio J de todas las matrices sinápticas. Denotamos con L_k al conjunto de matrices sinápticas que representan una la función booleana F_k y con $H[L_k]$ a la cantidad de elementos de dicho conjunto. Asimismo, llamamos $H[J]$ a la cantidad de matrices sinápticas distintas que existen en el espacio J y $H[M_k^{j1,j2,...,jn}]$ a la cantidad de matrices que representan alguna de las funciones $F_k^{j1,j2,...,jn}$ coincidentes con F_k en los puntos correspondientes a los n vectores de entrada $E^{j1,j2,...,jn}$, $(1 \le j_k \le N_e, k = 1, 2, ..., n)$ siendo $M_k^{j1,j2,...,jn}$ el conjunto de dichas matrices. N_e es el número total de ejemplos y n es la cantidad de ejemplos utilizados durante el aprendizaje.

Si elegimos al azar alguna matriz sináptica de entre las $H[J]$ existentes en J tenemos una probabilidad:

$$p_M = \frac{H\left[M_k^{j_1 j_2 \cdots j_n}\right]}{H[J]} \qquad (1)$$

de haber elegido una de las matrices que representan alguna de las funciones $F_k^{j1,j2,...,jn}$. Por lo tanto p_M es la probabilidad de "memorizar" los ejemplos utilizados. A su vez, tenemos una probabilidad:

$$p_G = \frac{H[L_k]}{H[J]} \qquad (2)$$

de haber elegido una de las matrices que representan la función F_k. Es decir, p_G es la probabilidad de "generalizar" procediendo por elección al azar. Las probabilidades indicadas en (1) y (2) corresponden a la "memorización" y "generalización", respectivamente, en una red que aún no ha sido sometida a un proceso de aprendizaje. Si dicho proceso de aprendizaje es llevado a cabo exitosamente, esto significa que la matriz sináptica pertenece al conjunto $M_k^{j1,j2,...,jn}$, que incluye todas las matrices que representan una función $F_k^{j1,j2,...,jn}$. Por lo tanto, la probabilidad de que la red "generalice" condicionada a que previamente haya logrado "memorizar" estos n ejemplos, es $H[L_k] / H[M_k^{j1,j2,...,jn}]$. Esta probabilidad depende del conjunto particular de ejemplos utilizados en el aprendizaje.

El proceso de aprendizaje equivale a una reducción de los estados accesibles del sistema, en el espacio J de las matrices sinápticas.

Si, en el aprendizaje de una función booleana F_k dada, la red logra memorizar exitosamente n "ejemplos" particulares $(\bar{E}^{[j]}, \bar{S}_c^{[j]})$, $(1 < j_l < N_e, l = 1, 2, ..., n)$ el espacio de los estados accesibles se reduce de J a $M_k^{j1,j2,...,jn}$. Esta reducción corresponde a un incremento en la probabilidad de generalización $p_G^{j1,j2,...,jn}$ definida como:

$$p_G^{j_1 j_2 \cdots j_n}(F_k) = \frac{H[L_k]}{H\left[M_k^{j_1 j_2 \cdots j_n}\right]} \qquad (3)$$

ya que $M^{j_1}(F_k) \supset M^{j_1 j_2}(F_k) \supset \ldots \supset M_k^{j_1 j_2 \cdots j_n} \supset \ldots L_k$

Por lo tanto: $M^{j_1}(F_k) \supset M^{j_1 j_2}(F_k) \supset \ldots \supset M_k^{j_1 j_2 \cdots j_n} \supset \ldots L_k$

$$p_G^{j_1}(F_k) \le p_G^{j_1 j_2}(F_k) \le \ldots \qquad (4)$$
$$\ldots \le p_G^{j_1 j_2 \cdots j_n}(F_k) \le \ldots \le p_G^{1,2,\ldots,n}(F_k)$$

Una vez memorizados n ejemplos particulares $(\bar{E}^{[j]}, \bar{S}_c^{[j]})$ de la función F_k estamos seguros de que se trata de alguna función booleana perteneciente al conjunto $C^{j1,j2,...,jn}(F_k)$ que agrupa a todas las funciones $F_k^{j1,j2,...,jn}$. Si el evento que nos interesa es la generalización de una dada función booleana F_k, la información (en bits) que obtenemos al saber que la

red efectivamente generalizó dicha función [una vez que ha memorizado los n ejemplos $(\bar{E}^{[j]}, \bar{S}_c^{[j]})$] está dada por $\log_2 [1 / p_G^{j1, j2,..., jn}(F_k)]$. Por ejemplo, la información que obtendríamos si la red lograra generalizar una función irrepresentable ($L_k = \emptyset \rightarrow p_G^{j1, j2,..., jn}(F_k) = 0$) sería infinita y la que corresponde a la generalización de una función representable luego de producida la memorización de los N_e ejemplos que definen dicha función [$p_G^{1, 2,..., Ne}(F_k) = 1$ si $L_k \neq 0$] es nula.

La red, una vez memorizados los n ejemplos $(\bar{E}^{[j]}, \bar{S}_c^{[j]})$ de F_k, puede representar cualquier función F_n perteneciente a $C^{j1, j2,..., jn}(F_k)$. La probabilidad con que ello ocurre es precisamente $C^{j1, j2,*, jn}(F_k)$ ya que, como todas estas funciones F_n pertenecen a $C^{j1, j2,..., jn}(F_k)$, en el cálculo dado por (3) el denominador es el mismo que el utilizado para calcular $p_G^{j1, j2,..., jn}(F_k)$. Por lo tanto, podemos caracterizar a la red mediante una entropía inferencial definida de la siguiente manera:

$$S^{j_1 j_2,...,j_n}(F_k) = \quad (5)$$
$$= -\sum_{F \in C^{j_1 j_2,...,j_n}(F_k)} p_G^{j_1 j_2 - j_n}(F) \log_2 p_G^{j_1 j_2 - j_n}(F),$$

donde hemos utilizado logaritmos en base 2 para expresar la información en bits.

La entropía de la red neuronal es una medida de la información promedio que se obtiene al conocer cuál es la función finalmente representada, una vez que ha concluido la etapa de aprendizaje de los n ejemplos $(\bar{E}^{[j]}, \bar{S}^{[j]})$. Es, por lo tanto, una medida de la incerteza promedio que existe con respecto a cuál regla ha inferido la red.

Para el caso en que $n = N_e$, si F_k es representable, la sumatoria de (5) se restringe sólo a $\{F_k\}$ y, como $p_G^{1,2,..,Ne}(F_k) = 1$ resulta $S^{1,2,..,Ne}(F_k)$. Si F_k no es representable, la sumatoria no incluye ningún término y también es $S^{1,2,..., Ne}(F_k) = 0$. Por otra parte, si n = 0, como $C(F_k) = F$ y $M(F) = J \; \forall \; F$, entonces se obtiene $p_G(F) = H[L(F)] / H[J] \; \forall \; F$ y (5) queda:

$$S(F_k) = -\sum_{F \in F} \frac{H[L(F)]}{H[J]} \log_2 \frac{H[L(F)]}{H[J]} \quad (6)$$

La variación de entropía inferencial entre los casos n = 0 (sin aprendizaje) y $n = N_e$ (aprendizaje perfecto) de la función F_k es:

$$\Delta S = S^{1,2,...,N_e}(F_k) - S(F_k) = \quad (7)$$
$$= \sum_{F \in F} \frac{H[L(F)]}{H[J]} \log_2 \frac{H[L(F)]}{H[J]},$$

donde, dado que esta variación de entropía inferencial no depende de la función F_k aprendida, hemos eliminado la indicación de esta última.

En (7), como H [L(F)] es menor o igual que H [J], resulta $\Delta S < 0$. En los casos en que $\Delta S < 0$ el aprendizaje produce, en promedio, una reducción de la entropía inferencial $-\Delta S$. Notar que las funciones no representables no contribuyen al valor de ΔS en (7). Esto significa que durante el aprendizaje de funciones irrepresentables no se produce ninguna variación de entropía. De la misma manera, si la red sólo pudiese representar una única función F_k ($H[L_k] = H[J]$, $H[L(F \neq F_k)] = 0$, la variación de entropía durante el aprendizaje en dicha red también sería nula. Por otro lado, es posible demostrar[1] que la reducción de entropía inferencial $-\Delta S$ es máxima para el caso en que todas las funciones son representables por un cantidad igual de matrices sinápticas($H[L(F)] = H[J] / N_{bool}$), donde N_{bool} es la cantidad total de funciones booleanas. Dicho valor máximo de reducción de entropía inferencial es:

$$\Delta S_{max} = \log_2 N_{bool} \quad (8)$$

Vemos entonces que la variación de entropía inferencial dada por (7) permite caracterizar la estructura interna de la red, de acuerdo con el grado de uniformidad con que la red es capaz de aprender todas las funciones booloanas pertenecientes a F. Si $\Delta S = 0$ la red posee una arquitectura tan especializada que sólo es capaz de representar una única función booleana (cualquiera sea ella). Consecuentemente, la capacidad de "aprender" de esta red es nula. Si $\Delta S = -\Delta S_{max}$, se trata de una red versátil, capaz de representar todos los problemas con igual probabilidad. Es decir, ΔS permite clasificar a las redes neuronales en redes de propósito específico ($\Delta S \rightarrow 0$) o redes de propósito general ($\Delta S \rightarrow -\Delta S_{max}$).

En trabajos anteriores[2, 3] hemos descripto un proceso de aprendizaje basado en el método del recocido simulado. El mismo consiste en realizar varios "ciclos" de barridos Monte Carlo a distintas temperaturas efectivas T. Al final de cada ciclo se debe alcanzar el equilibrio termodinámico, es decir, para cada temperatura T la energía promedio $< \varepsilon(J) >$ debe estabilizarse en un valor constante. Este promedio se obtiene dejando que el sistema evolucione temporalmente y tomando en cuenta las sucesivas configuraciones que van apareciendo (el proceso Monte Carlo simula dicha evolución temporal).

También podemos obtener este promedio utilizando las técnicas usuales de mecánica estadística, es decir realizando promedios en el ensamble de redes neuronales correspondiente. La función de partición está dada por:

$$Z_{F_k}^{j_1 j_2 \cdots j_n} = \sum_{J \in J} e^{-\beta \varepsilon_{j_1 j_2 \cdots j_n}(J, F_k)} =$$

$$= \sum_{F \in F} H\left[L(F)\right] e^{-\beta \varepsilon_{j_1 j_2 \cdots j_n}(F, F_k)} \qquad (9)$$

donde J actúa como variable microscópica y F como macroscópica. Asimismo, H [L (F)] puede entenderse como la degeneración de la energía y F_k como un parámetro o "campo" externo. Es importante remarcar que $T = 1/\beta$ coincide con la temperatura efectiva utilizada en el proceso de aprendizaje.

Como, por definición, $e^{j1, j2, \ldots, jn}(F, F_k)$ no puede ser negativa, entonces para T = 0 resulta:

$$Z_{F_k}^{j_1 j_2 \cdots j_n}(T=0) =$$

$$= \sum_{F \in C j_1 j_2 \cdots j_n(F_k)} H\left[L(F)\right] = H\left[M_k^{j_1 j_2 \cdots j_n}\right] \qquad (10)$$

La entropía está dada por:

$$S_{F_k}^{j_1 j_2 \cdots j_n} = -\left\langle \log_2 \frac{H\left[L(F)\right] e^{-\beta \varepsilon_{j_1 j_2 \cdots j_n}(F, F_k)}}{Z_{F_k}^{j_1 j_2 \cdots j_n}} \right\rangle =$$

$$= -\sum_{F \in F} \frac{H\left[L(F)\right] e^{-\beta \varepsilon_{j_1 j_2 \cdots j_n}(F F_k)}}{Z_{F_k}^{j_1 j_2 \cdots j_n}} \cdot$$

$$\cdot \log_2 \frac{H\left[L(F)\right] e^{-\beta \varepsilon_{j_1 j_2 \cdots j_n}(F F_k)}}{Z_{F_k}^{j_1 j_2 \cdots j_n}} \qquad (11)$$

Para T = 0, resulta:

$$S_{F_k}^{j_1 j_2 \cdots j_n}(T=0) = -\sum_{F \in C j_1 j_2 \cdots j_n(F_k)} \frac{H\left[L(F)\right]}{H\left[M_k^{j_1 j_2 \cdots j_n}\right]} \cdot$$

$$\cdot \log_2 \frac{H\left[L(F)\right]}{H\left[M_k^{j_1 j_2 \cdots j_n}\right]} \qquad (12)$$

Si F_k es la función representable o si, siendo irrepresentable, el conjunto de ejemplos utilizados durante el aprendizaje es memorizable, la expresión (12) coincide con la ec. (5) que define la entropía inferencial de la red. Notar que esta correspondencia está ligada a la finalización del proceso de aprendizaje, ya que (12) sólo es válida para $T \doteq 0$.

REFERENCIAS

1. Abramson, Teoría de la información y codificación, Ed. Paraninfo, 1974.
2. Ferrán y Perazzo, Symmetry properties of feed–forward neural networks, CNEA preprint, 1989.
3. Ferrán y Perazzo, Redes neuronales que aprenden, Comunicación a la 73ª Reunión de la AFA, 1988.