Contrastive Divergence

27 de junho de 2020

Boltzmann Machine — [Hertz, Hinton, Aggarwal]

Stocastic neural network of random binary variables s_i. Connections between units i and j are symmetrical $\omega_{ij} = \omega_{ji}$ (undirected).

Energy based model. Energy is used to determine the configurations of the network, by calculating the probabilities of occurrence, using the Boltzmann distribution as the probability distribution of the observed network states.

The network consists of N **visible** units and M **hidden** units.

Boltzmann Machine

Energy function is

$$H_{vh} = -\frac{1}{2}\sum_i \sum_j s_i^{(vh)} \omega_{ij} s_j^{(vh)} - \sum_i \phi_i s_i^{(vh)}, \qquad (1.1)$$

where ϕ_i is the bias (!!!!) of unit i. The superscript (vh) identifies the joint state, where visible units have configuration v and hidden units have configuration h. s_i and s_j identify two different units that are connected to each other by a synapse with weight ω_{ij}.

Boltzmann Machine

From the Boltzmann distribution, the joint probability of finding the system with energy H_{vh}, corresponding to a specific configuration (vh) is

$$P_{vh} = \frac{e^{-H_{vh}/T}}{Z}. \tag{1.2}$$

Z is the partition function, a normalization factor that considers all possible configurations of the system, i.e.,

$$Z = \sum_u \sum_k e^{-H_{uk}/T}, \tag{1.3}$$

where T is the *pseudo-temperature*, a dimensionless quantity.

Boltzmann Machine

(!!!) We are interested in modelling the configuration of the visible units in order to use the machine to represent a dataset, i.e., we would like to use the BM to model the marginal probability distribution of the visible units P_v that matches the probability distribution of a given dataset from the real world, (!!!)

rede vai representar alguma propriedade do dado, rede usa a probabilidade para modelar o processo de convergência.

$$P_v = \frac{\sum_h e^{-H_{vh}/T}}{Z} \tag{1.4}$$

Boltzmann Machine

(!!!) Consider the real world dataset has a probability distribution R_v which is an unknown probability distribution of observed events, where v refers to the dimensionaly, i.e., measured variable of each event. We want to use BM to model a probability distribution P_v which *replicates* the distribution of the data. (!!!)

We use Kullback-Leibler divergence to measure the differences between distributions,

$$D_{KL}(R_v||P_v) = \sum_v R_v \ln\left(\frac{R_v}{P_v}\right). \tag{1.5}$$

We want to find the best set of parameters ω_{ij} and ϕ_i that minimazes D_{KL} (turns P_v into the best approximation of R_v.)

Boltzmann Machine

Minimization of the relative entropy, i.e., $D_{KL}(R_v||P_v)$,

$$\frac{\partial D_{KL}}{\partial \omega_{ij}} = 0, \tag{1.6}$$

$$\frac{\partial D_{KL}}{\partial \phi_i} = 0. \tag{1.7}$$

In equation (1.5), only P_v is dependent on ω_{ij} and ϕ_i.

Boltzmann Machine — Learning

Learning in a BM consists in updating the parameters ω_{ij} and ϕ_i through an iterative process,

$$\omega_{ij}^{(updated)} = \omega_{ij}^{(current)} + \Delta\omega_{ij}, \tag{1.8}$$

$$\phi_i^{(updated)} = \phi_i^{(current)} + \Delta\phi_i, \tag{1.9}$$

where

$$\Delta\omega_{ij} = -\eta \frac{\partial D_{KL}}{\partial \omega_{ij}}, \tag{1.10}$$

and analogously for $\Delta\phi_i$.

Boltzmann Machine — Learning

Derivation of $\Delta\omega_{ij}$. Skip that for the moment. 😄

Updates of ω_{ij} and ϕ_i can consume a lot of computational time. The difficulty consists in calculating the partition function every-time the parameters ω_{ij} and ϕ_i are updated, which involves a summation over all possible configurations of the nodes of the network. 😵

In the BM, there can be a connection between each pair of nodes. For each unit, its activation depends on the state of every other unit in the network.

Restricted Boltzmann Machine — [Smolensky, Aggarwal, Hinton and Larochelle]

(!!!) RBM has the same properties of a BM, where a probability distribution of a sample of a real world observation is being modelled. (!!!)

In the RBM there are restrictions in the connections among units. There are no connections among two visible units or two hidden units. Only connections between a visible and a hidden unit are allowed.

Restricted Boltzmann Machine

We can consider that the configuration of a network is given by the joint configuration of a visible state vector $\mathbf{v} = (v_1, v_2, \ldots, v_N)$ and a hidden state vector $\mathbf{h} = (h_1, h_2, \ldots, h_M)$. The energy function is

$$H(\mathbf{v}, \mathbf{h}) = -\sum_i \sum_j v_i \omega_{ij} h_j - \sum_j b_j h_j - \sum_i c_i v_i, \qquad (1.11)$$

where b_j and c_i are the biases related to hidden unit j and visible unit i, respectively. The parameter $\omega_{ij} = \omega_{ji}$ corresponds to the symmetric connection between visible unit v_i and hidden unit h_j.

Restricted Boltzmann Machine

The joint probability $P(\mathbf{v}, \mathbf{h})$ of finding the network at visible state vector $\mathbf{v}$ and hidden state vector $\mathbf{h}$ is

$$P(\mathbf{v}, \mathbf{h}) = \frac{e^{-H(\mathbf{v},\mathbf{h})}}{Z}, \tag{1.12}$$

where Z is the partition function defined as in the case of the BM, by equation (1.3),

$$Z = \sum_{\mathbf{u}} \sum_{\mathbf{k}} e^{-H(\mathbf{u},\mathbf{k})}. \tag{1.13}$$

Restricted Boltzmann Machine

The probability $P(\mathbf{v})$ of finding the visible units in state $\mathbf{v}$, regardless of the hidden state $\mathbf{h}$ is

$$P(\mathbf{v}) = \sum_{\mathbf{h}} P(\mathbf{v}, \mathbf{h}) = \sum_{\mathbf{h}} \frac{e^{-H(\mathbf{v},\mathbf{h})}}{Z}. \tag{1.14}$$

(!!! Colocar lá em cima !!!) In equation (1.14), we see that the numerator increases as the energy increases.

!!! For a giver visible state vector $\mathbf{v}$ and the RBM finds the hidden state vector $\mathbf{h}$ that nicely in order to low the energy. The denominator is searching for global configurations of visible and hidden states that provide low energy which means a large contribution to Z. !!!

Restricted Boltzmann Machine

Both visible and hidden units are random binary variable. For N visible units, $\mathbf{v} \in \{0,1\}^N$, and for M hidden units, $\mathbf{h} \in \{0,1\}^M$. Then we can expand equation (1.14),

$$
\begin{aligned}
P(\mathbf{v}) &= \sum_{\mathbf{h} \in \{0,1\}^M} \frac{e^{-H(\mathbf{v},\mathbf{h})}}{Z} \\
&= \frac{1}{Z} \sum_{\mathbf{h} \in \{0,1\}^M} e^{\left(\sum_i c_i v_i + \sum_j b_j h_j + \sum_i \sum_j v_i \omega_{ij} h_j\right)} \\
&= \frac{1}{Z} \left(e^{\sum_i c_i v_i}\right) \left[\sum_{\mathbf{h} \in \{0,1\}^M} e^{\left[\sum_j \left(b_j + \sum_i v_i \omega_{ij}\right) h_j\right]}\right]
\end{aligned}
$$

Restricted Boltzmann Machine

$$= \frac{1}{Z}\left(e^{\sum_i c_i v_i}\right)\left[\sum_{h_1\in\{0,1\}} e^{\left(b_1+\sum_i v_i\omega_{i1}\right)h_1} \ldots \sum_{h_M\in\{0,1\}} e^{\left(b_M+\sum_i v_i\omega_{iM}\right)h_M}\right]$$

$$= \frac{1}{Z}\left(e^{\sum_i c_i v_i}\right)\prod_{j=1}^{M}\left[\sum_{h_j\in\{0,1\}} e^{\left(b_j+\sum_i v_i\omega_{ij}\right)h_j}\right]$$

$$= \frac{1}{Z}\left(e^{\sum_i c_i v_i}\right)\prod_{j=1}^{M}\left[1+e^{\left(b_j+\sum_i v_i\omega_{ij}\right)}\right]$$

Restricted Boltzmann Machine

$$= \frac{1}{Z}\left(e^{\sum_i c_i v_i}\right) e^{\ln\left[\prod_{j=1}^{M}\left[1+e^{\left(b_j+\sum_i v_i\omega_{ij}\right)}\right]\right]}$$

$$\Rightarrow P(\mathbf{v}) = \frac{1}{Z} e^{\left(\sum_i c_i v_i + \sum_j \ln\left[1+e^{\left(b_j+\sum_i v_i\omega_{ij}\right)}\right]\right)} \tag{1.15}$$

The term in the exponential in equation (1.15) is know as the Free Energy of visible state vector $\mathbf{v}$,

$$F(\mathbf{v}) = -\sum_i c_i v_i - \sum_j \ln\left[1 + e^{\left(b_j+\sum_i v_i\omega_{ij}\right)}\right] \tag{1.16}$$

$$\Rightarrow P(\mathbf{v}) = \frac{e^{-F(\mathbf{v})}}{Z} \tag{1.17}$$

Restricted Boltzmann Machine

Modelling the marginal probability distribution of visible state $P(\mathbf{v})$ means finding a good approximation based on the Boltzmann distribution which models the likelihood of the observed events. To do so, we can maximize the likelihood by,

$$\frac{\partial P(\mathbf{v})}{\partial \omega_{ij}} = 0 \tag{1.18}$$

$$\frac{\partial P(\mathbf{v})}{\partial b_j} = 0 \tag{1.19}$$

$$\frac{\partial P(\mathbf{v})}{\partial c_i} = 0 \tag{1.20}$$

Restricted Boltzmann Machine

One strategy is to maximize the log-likelihood, as the logarithm functions is monotonic and, thus both likelihood and log-likelihood have the same maximum point

$$P(\mathbf{v}) = \frac{\sum_{\mathbf{h}} e^{-H(\mathbf{v},\mathbf{h})}}{Z}$$

$$\ln\left[P(\mathbf{v})\right] = -F(\mathbf{v}) - \ln[Z] \tag{1.21}$$

$$\Rightarrow \frac{\partial \ln[P(\mathbf{v})]}{\partial \omega_{ij}} = 0. \tag{1.22}$$

Restricted Boltzmann Machine

Suppose we have M hidden units, and N visible units, for a given visible state vector $\mathbf{v}$ all hidden units activation can be computed at the same time, as they are independent (no hidden-hidden connection), this means that

$$P(\mathbf{h}|\mathbf{v}) = \prod_{j=1}^{M} P(h_j|\mathbf{v}) \tag{1.23}$$