Energy and Probability

Chunpai Wang

October 11, 2018

1 Boltzmann's Law

Assume N particles in a system, and each particle has two states, positive and negative, where the states of the i^{th} particle are labeled by x_i. The overall state of the system is denoted by a vector $\mathbf{x} = x_1, ..., x_N$.Each configuration $\mathbf{x}$ of those N particles associates with an energy $E(\mathbf{x})$. The idea of energy comes from the statistical mechanics, which derives a fundamental result that, in thermal equilibrium, the probability of a state will be given by **Boltzmann's Law** or **Gibbs distribution**:

$$p(\mathbf{x}) = p(x_1, \cdots, x_N) = \frac{1}{Z(T)} \exp\{-E(\mathbf{x}/T)\}$$

where T is the temperature, and $Z(T)$ is the partition function, which is used as normalization:

$$Z(T) = \sum_{\mathbf{x} \in S} \exp\{-E(\mathbf{x})/T\}$$

where S is the space of all possible states $\mathbf{x}$ of the system. As we can see, this partition function usually is intractable. Here we simply view the Boltzmanns's Law as postulate that serves to define the energy for the system, and we view T as the parameter of the function, and usually set $T = 1$ for the sake of simplicity.

2 Potential Function and Energy

In the Markov random fields, based on the *Hammersley-Clifford* we define the joint probability of N variables as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c=1}^{M} \phi_c(\mathbf{x}_c)$$

which is a product of **potential function** $\phi(\mathbf{x}_c)$ over M number of maximal cliques of the graph. If we connect those two joint probability distributions, we have

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c=1}^{M} \phi_c(\mathbf{x}_c) = \frac{1}{Z} \exp\{-E(\mathbf{x})\}$$

and

$$E(\mathbf{x}) = -\sum_{c=1}^{M} \ln \phi_c(\mathbf{x_c})$$

Here, we assume the total energy $E(\mathbf{x})$ in the system can be decomposed as the summation of energies of all maximal cliques, which is

$$E(\mathbf{x}) = \sum_{c=1}^{M} E(\mathbf{x}_c)$$

then we have the energy of each of the maximal cliques as

$$E(\mathbf{x}_c) = -\ln \phi_c(\mathbf{x_c})$$
$$\phi_c(\mathbf{x}_c) = \exp\{-E(\mathbf{x}_c)\}$$

We can see that if the probability $\phi_c(\mathbf{x_c})$ is small, then the energy is large, which indicates the incompatibility of the states x_c.

3 Energy Minimization as Maximizing the Joint Probability

If we want to find the most provable states, then we opt to find the highest joint distribution, which can be interpreted as minimizing the energy function as below:

$$\begin{aligned}\arg\max_{\mathbf{x}\in S} p(\mathbf{x}) &= \arg\max_{\mathbf{x}\in S} \frac{1}{Z}\prod_{c=1}^{M}\phi_c(\mathbf{x}_c) \\ &= \arg\max_{\mathbf{x}\in S} \frac{1}{Z}\exp\{-E(\mathbf{x})\} \\ &= \arg\min_{\mathbf{x}\in S} E(\mathbf{x})\end{aligned}$$

Therefore, the structure prediction problem can be viewed as a problem to minimizing the overall energy in the system [].

4 Representing Potential Functions

Now, here we bring back the parameter T in Gibbs distribution, and treat it as the parameter of joint probability

$$p(\mathbf{x};T) = \frac{1}{Z}\prod_{c=1}^{M}\phi_c(\mathbf{x}_c|T_c)$$

As we known before, the potential is associated with the energy

$$E(\mathbf{x}_c|T_c) = -\log\phi_c(\mathbf{x}_c|T_c)$$

which indicates

$$\text{low energy} = \text{high potential} = \text{high compatibility of states}$$

We can view energy of system from the friction between particles, in other words, incompatibility of states, which is also impacted by the temperature parameter T.
A more general approach is to define the log-potential as the linear function

$$\log\phi_c(\mathbf{x}_c|T_c) = -E(\mathbf{x}_c|T_c) = T_c^\top f_c(\mathbf{x}_c)$$

where the $f_c(\mathbf{x}_c)$ is a feature vector derived from the values of the variables $\mathbf{x}_c$. The feature vector can be derived from the feature functions, which usually are hand-crafted by domain expert. Thus, we have the log joint distribution as

$$\begin{aligned}\log p(\mathbf{x}|T) &= \log\frac{1}{Z(T)}\prod_{c=1}^{M}\phi_c(\mathbf{x}_c|T_c) \\ &= \log\frac{1}{Z(T)}\prod_{c=1}^{M}T_c^\top f_c(\mathbf{x}_c) \\ &= \log\frac{1}{Z(T)} + \sum_{c}^{M} T_c^\top f_c(\mathbf{x}_c) \\ &= \sum_{c}^{M} T_c^\top f_c(\mathbf{x}_c) - \log(Z(T))\end{aligned}$$

which is also known as a maximum entropy or a log-linear model.

References

[1] Nowozin, Sebastian, and Christoph H. Lampert. "Structured learning and prediction in computer vision." Foundations and Trends® in Computer Graphics and Vision 6.3–4 (2011): 185-365.

[2] LeCun, Yann, et al. "A tutorial on energy-based learning." Predicting structured data 1.0 (2006).