CS281: Advanced ML **November 27, 2017**

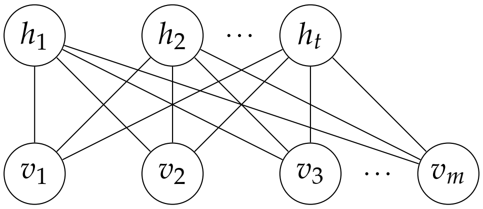
Lecture 22: MCMC and Boltzmann Machines

Lecturer: Sasha Rush *Scribes: Vasileios Nakos and Sebastien Lemieux-Codere*

22.1 The Restricted Boltzman Machine(RBM)

Restricted Boltzman Machine were used in the area of collaborative filtering. Many people thought they would be great for Deep Learning applications, but tend not to believe in it anymore. Restricted Boltzman Machine:

- Are undirected graphical models.
- Have the structure of a bipartite graph.
- Have binary random variables.



The join distribution of v and h is:

$$\mathbb{P}[v,h|\theta] = \frac{1}{Z(\theta)} \exp\left\{ \sum_{i:h_i,j:v_j} \theta_{ij}(h_i v_j) \right\}.$$

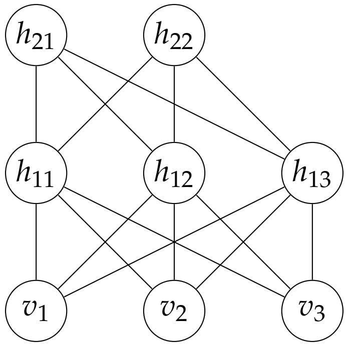
RBMs have been proven useful for the Netflix Grand prize; the interested reader can have a look at this paper from Andreas Toscher and Michael Jahrer for more details. A natural question for this graphical model is whether we can do inference or not. There are no big cliques, but it's relatively densely connected. However, inference is hard! But... **Inference is easier under conditioning!** For example, if we have observed all the v nodes, we might be able to deduce the distribution on the h nodes, by performing (roughly) logistic regression. For example, in the case of binary variables we have that:

$$\mathbb{P}[h|v,\theta] = \Pi_k \text{Ber}(h_k|\text{sigm}(w_{:,k}^T v),$$

and

$$\mathbb{P}[v|h,\theta] = \Pi_r \text{Ber}(v_r|\text{sigm}(w_{r,:}, h),$$

where $w \in \mathbb{R}^{R\times K}$. We can see that the above setup is very similar to a hidden layer in a neural network. This is the first time we see a connection between graphical models and a hidden layer. Upon conditioning on the v's, h node probabilities are a neural network layer. One can have Deep Boltzmann Machines, in which higher layers capture complicated, higher-order relationships between the hidden features in the previous layer:



To perform the inference, we could use Lossy Belief Propagation (LBP), Mean Field (MF) or Expectation Mazimixation (EM). In practice, it turns out that Gibbs sampling works well for Restricted Boltzman Machines when we observe $\mathbb{P}[h|v]$ or $\mathbb{P}[h|v]$. If we observe $p(h|v)$, we can compute $p(v|h)$ in the following way

```
$i = 1$
Set initial $v^{(1)}$
while condition holds do
    Compute $p(h|v^{(i)})$ (Neural Network Layer).
    Sample each node to get $h^{(i)}$.
    Compute $p(v|h^{(i)})$.
    Sample each node to get $v^{(i+1)}$
    $i = i + 1$
```

Eventually, this process will reach an equilibrium and we can interpret new samples as coming from the joint distribution of v and h. This is procedure is called Gibbs sampling, which is a type of Markov Chain Monte Carlo algorithm.

22.2 Markov Chain Monte Carlo (MCMC)

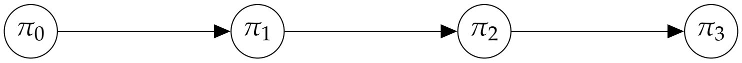
We are interested in computing $\mathbb{E}_{x \sim p} f(x)$, where p is a distribution and f is a function. Depending on what p is, we might be able to do this with sampling methods already covered:

1. Just sample from p
2. Use rejection Sampling. This would involve proposing $q(x)$ and a bound M on the likelihood ratio $f(x)/q(x)$.
3. Use importance sampling
4. Use particle filtering

Another option is to use Markov Chain Monte Carlo (MCMC) methods, a class of algorithms for sampling from a probability distribution where we construct a Markov chain whose stationary distribution is the target density p(x). Gibbs sampling is a MCMC method applicable when the joint distribution is not known explicitly, but the conditional distribution of each variable is known. The Gibbs sampling algorithm is used to generate an instance from the distribution of each variable in turn, conditional on the current values of the other variables. Often it can be shown that the sequence of samples comprises a Markov chain, and the stationary distribution of that Markov chain is the sought-after joint distribution.
Example: Gibbs Sampling on RBM. **Trade-off**: Samples are now correlated, and hence we lose independence. On the other hand, we do not waste many samples in low probability events.

22.2.1 Markov Chains

Markov Chains are a sequence of a random variables $\{\pi\}_{i\in\mathbb{N}}$, which satisfy $\pi_{i+1} = T\pi_i$ and π_0 is given. T is called the Transition Matrix and the i, j entry $T(i|j)$ is the probability to transition to state i from state j. One can think of the Markov chain as starting from an initial distribution π_0 and applying the transition matrix t times to arrive at distribution $\pi_t = T^t\pi_0$.



Key Definitions:

- **Irreducibility**: It's possible to go from any state to any state. In other words, the probability of reaching any state starting from any other state is greater than 0.
- **Aperiodicity**: We have aperiodicity if no state is periodic. A state is periodic with period $k > 1$ if any return to that state from that state must occur in multiples of k time steps. More formally, a state is aperiodic if $\gcd\{n \in \mathbb{N} : \mathbb{P}(\pi_n = i|\pi_0 = i) > 0\} = 1$.
- **Positive recurrent**: A Markov chain is recurrent if the probability of eventually returning from every state to itself is 1. A Markov chain is positive recurrent if it is recurrent and the expected value of the number of steps before returning from every state to itself for the first time is finite.
- **Ergodicity**: A Markov chain is ergodic if it is positive recurrent and aperiodic.
- **Stationary Distribution / Invariance**: A distribution π^* is stationary (or invariant) if $\pi^* = T\pi^*$. In the continuous case if $\pi^*(x') = \int T(x'|x)\pi^*(x)dx$. By definition, π^* is an eigenvector with corresponding eigenvalue $\lambda = 1$.

Key Theorems:

- **Uniqueness of Stationary Distribution**: If a finite state Markov chain is irreducible, aperiodic and has a stationary distribution, then (1) this stationary distribution is unique and (2) the Markov chain will asymptotically converge to that unique stationary distribution as $t \to \infty$.
- **Detailed Balance Condition**: $\pi^*(x)T(x'|x) = \pi^*(x')T(x|x')$ implies that π^* is a stationary distribution. It is a sufficient but not necessary condition for π^* to be a stationary distribution.
- **Convergence to a Unique Stationary Distribution**: If a Markov chain is irreducible and ergodic, then (1) it has a limiting distribution π^* such that the Markov chain will converge to π^* as $t \to \infty$ regardless of the starting distribution, and (2) this limiting distribution is the Markov chain's unique stationary distribution. By definition, π^* is an eigenvector with corresponding eigenvalue $\lambda = 1$. The rate of convergence is determined by the second largest eigenvalue.

If we ensure the detailed balance property, we can approximate the original distribution with a Markov chain. We do this by picking transition probability T and its stationary distribution π^* such that the above Detailed Balance condition holds and that π^* approximates p. Thus, asymptotically applying T to get the next correlated sample yields samples of π^* (assuming the Markov chain is irreducible and ergotic), which can be used as an approximation for p.

Properties:

- *Expected Value*:

$$\mathbb{E}[\frac{1}{N}\sum_{n=1}^{N} f(x_n)] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[f(x_n)] \sim \frac{1}{N}\sum_{n=1}^{N} \int \pi^*(x)f(x)dx = \mathbb{E}_{x\sim\pi^*}f(x).$$

- *Variance*:

$$\mathrm{Var}(1/N\sum f(x_n)) = \frac{1}{N^2}\mathrm{Var}(\sum f(x_n)) = \frac{1}{N^2}[\sum Var(f(x_n)) + 2\sum_n\sum_{n'}\mathrm{cov}(f(x_n), f(x_{n'})].$$

The first term in the variance is small for large N, while the second term is the additional variance from the correlation of sampling.

22.2.2 Metropolis-Hastings

The Metropolis–Hastings algorithm is a Markov chain Monte Carlo (MCMC) algorithm that was originally developed in the Manhattan project: we define a random walk that goes from x to x' with probability $q(x'|x)$. $q(x'|x)$ is called the proposal distribution and it should have a non-zero probability for states that have non-zero probability in $p(x)$. A common proposal distribution is a Gaussian distribution centered at the last sample. The algorithm proceeds in the following way the get the (k+1)th sample:

```
while condition holds do
    Draw $x^*$ from $q(x^*|x^{(k)})$
    Compute $A = min(1, \frac{p(x^*)q(x^{(k)}|x^*)}{p(x^{(k)})q(x^*|x^{(k)})})$
    Set $x^{(k+1)} = x^*$ with probability A. Otherwise, we keep $x^{(k)}$ as the next sample: $x^{(k+1)} = x^{(k)}$.
```

Note that we only need to be able to compute $p(x)$ up to a multiplicative constant.

The algorithm is designed such that $\pi^*(x) = p(x)$. Consider the case where $x \neq x'$. The detailed balance condition holds:

$$T(x'|x) = q(x'|x) \cdot \min\{1, \frac{p(x')q(x|x')}{(p(x)q(x'|x)})$$

and so

$$\pi^*(x)T(x'|x) = p(x)q(x'|x) \cdot \min\{1, \frac{p'(x)q(x|x')}{(p(x)q(x'|x))}\} = \min\{p(x)q(x'|x), p(x')q(x|x')\}.$$

In the above expression we have that the role of x', x is symmetric and hence we get that

$$\pi^*(x')T(x|x') = \min\{p(x')q(x|x'), p(x)q(x'|x)\},$$

Since the detailed balance condition holds with $p(x) = \pi^*(x)$, $p(x)$ is a stationary distribution. Under the assumption that the Markov chain is ergodic and irreducible, $p(x)$ is the unique stationary distribution and the Markov chain will asymptotically converge to it regardless of the initial distribution.

Gibbs sampling is a special case of Metropolis-Hastings. It is equivalent to using Metropolis-Hastings with proposals of the form $q(x'|x) = p(x'_i|x_{i-1})\mathbb{I}[x'_{-i} = x_{-i}]$. The acceptance probability is then 1:

$$\frac{p(x')}{p(x)}\frac{q(x|x')}{q(x'|x)} = \frac{p(x'_i|x'_{-i})p(x'_{-i})p(x_i|x'_{-i})}{p(x_i|x_{-i})p(x_{-i})p(x'_i|x_{-i})} = 1.$$