AI2613 随机过程	2020–2021 春季学期

Lecture 11 – Markov Random Field and Hidden Markov Model

2021 年 5 月 10 日

Lecturer: 张驰豪 *Scribe*: 杨宽

1 Markov Random Field

Today we are going to talk about a generalization of a type of simple Markov chains — discrete-time Markov chains with finite state space.

Consider a Markov chain $X_0 \to X_1 \to X_2 \to \cdots \to X_n \to \cdots$. Each X_t is a random variable and determined by X_{t-1} and the transition probabilities. Furthermore, given X_{t-1}, the value of X_t is independent of $X_0, \dots, X_{t-2}$.

Now we would like to generalize this model. Assume that each X_t depends on several others. We use an undirected graph to describe the dependency among random variables so that the marginal distribution of X_v only depends on the value of its neighbours. Specifically, we have the following definition.

Definition 1 (Markov Random Field). Given a graph $G = (V, E)$ of size $|V| = n$ and the state space $[q]$, we say a set of random variables $X = \{X_v\}_{v \in V}$ is a *Markov random field* with respect to G if their joint distribution satisfies that

$$\forall\, i, (j_w)_{w \in V \setminus \{v\}}, \qquad \mathbf{Pr}\left[X_v = i \,\middle|\, \bigwedge_{w \in V \setminus \{v\}} X_w = j_w \right] = \mathbf{Pr}\left[X_v = i \,\middle|\, \bigwedge_{w \in N(v)} X_w = j_w \right],$$

where $N(v)$ is the set of v's neighbours in graph G. In other words, X_v is independent of all other nonadjacent variables.

Remark. The underlying graph of a Markov random field may be finite or infinite.

A natural question is, given the dependency graph $G = (V, E)$, how to construct the joint distribution of X_vs so that they become a Markov random field. Clearly if each X_v is independent of all other variables then they form a Markov random field. But now we would like to design a nontrivial one.

For convenience we introduce some notations (probably there are some abuses). Given a distribution μ over $x = (x_v)_{v \in V} \in [q]^V$ and $S \subseteq V$, let $x_S = (x_v)_{v \in S}$ and

$$p(x_S) = \mathbf{Pr}_{X \sim \mu}[X_S = x_S] \text{ for all } x_S .$$

We should also mention here that we use $\{\cdot\}$ to denote *unordered tuples* (i.e., sets) and use $(\cdot)$ to denote *ordered tuples* (i.e., vectors).
Moreover, given $x, y \in [q]^V$ and $S, T \subseteq V$, let

$$p(x_S \mid y_T) = \mathbf{Pr}_{X \sim \mu}\left[X_S = x_S \mid X_T = y_T\right] .$$

Suppose that for all $v \in V$ and $y_{N(v)} \in [q]^{N(v)}$, $p(\cdot \mid y_{N(v)})$ is known. Is it possible to recover the joint distribution μ?
Actually, such a joint distribution may not exist, and it is not difficult to design a counterexample. So we are going to talk about under which condition the joint distribution exists and is a Markov random field.

2 Gibbs Distribution and Sampling

We now introduce another description of Markov random fields.

Definition 2 (Gibbs Distribution). Given an underlying graph $G = (V, E)$ and a state space $\Omega = [q]^V$, a distribution μ on Ω is called the *Gibbs distribution*, if there exists a family of functions $V_A(\cdot)\colon \Omega \to \mathbb{R}_{\geq 0}$ such that $V_A(x)$ only depends on $x|_A$ and

$$\forall\, x, \qquad \mu(x) = \prod_{\substack{A \subseteq V \\ A \text{ is complete}}} V_A(x) ,$$

where A is a complete set if all vertices in A form a *clique*, namely, i, j are adjacent for all $i, j \in A$.

Example 3 (Independent set). An *independent set* in a graph is a set of vertices, no two of which are adjacent.
The uniform distribution over all independent sets in a graph is a Gibbs distribution.
To see this, we let $q = 2$ and the state space be $\{0, 1\}$. Then x is an indicator vector to denote which variables are chosen to form a set. If $p(x)$ is a constant for every x that indicates an independent

set, and $p(x) = 0$ otherwise, μ is a Gibbs distribution. Now we define $V_A(\cdot)$ by

$$V_A(x) = \begin{cases} 1 & \text{if } |A| > 2 \\ \mathbb{1}_{[x(i)=0 \vee x(j)=0]} & \text{if } A = \{i,j\} \text{ for some } \{i,j\} \in E \\ 1 & \text{if } |A| = 1 \\ 1/Z & \text{if } A = \emptyset \end{cases}$$

where Z is the number of all independent sets and thus $1/Z$ is the normalizing factor.

Generally, consider the following model. For each $x \in [q]^V$, let

$$w(x) = \prod_{\{i,j\} \in E} w_{\{i,j\}}(x(i), x(j)),$$

and

$$\mu(x) \propto w(x).$$

In fact, let $Z = \sum_x w(x)$. We can define $\mu(x)$ by $\mu(x) = w(x)/Z$. We call Z the *partition function.*

Example 4 (Proper Coloring). A coloring is a configuration $c\colon V \to [q]$. Then the state space is $\Omega = [q]^V$. A coloring is *proper* if no two adjacent vertices have the same color, that is, for all $\{i,j\} \in E$, $c(i) \neq c(j)$.

Let μ be the uniform distribution over all proper colorings. Then μ is a Gibbs distribution.

$$\mu(x) \propto w(x) = \prod_{\{i,j\} \in E} \mathbb{1}_{x(i) \neq x(j)}.$$

Example 5 (Ising Model). The *Ising model* has a parameter $\beta > 0$. The state space is $\Omega = \{0,1\}^V$. The weight of each configuration is given by

$$w(x) = \beta^{\text{\# of monochromatic edges}}.$$

Then the distribution $\mu(x) \propto w(x)$ is a Gibbs distribution.

$$\mu(x) \propto w(x) = \prod_{\{i,j\} \in E} \beta^{\mathbb{1}_{x(i) \neq x(j)}}.$$

Theorem 1 (Hammersley–Clifford Theorem). *Given an underlying graph G and a distribution μ on Ω, if*

$$\forall x, \qquad \mu(x) > 0,$$

then variables $\{X_v\}_{v \in V}$ with distribution $\mu(\cdot)$ is a Markov random field if and only if μ is a Gibbs distribution.

The proof is omitted here. Please see our course reference (Chang's note) for details.

This theorem tells us that Markov random fields could be described by Gibbs distributions.

A natural question is, given $V_A(\cdot)$, how to efficiently sample from the Gibbs distribution, or equivalently how to compute those marginal probabilities induced by the distribution. Unfortunately, this problem is #P-hard. So we turn into considering approximating these values. However, this task is NP-hard in general. (Formally, this problem is in $\mathsf{FP}^{\mathsf{NP}}$, which means we can approximately sample in polynomial-time given an NP oracle.)

Nevertheless, if we do not care about the efficiency of the sampler at the moment, there is a simple sampling algorithm — Metropolis algorithm, which we introduced before.

Consider the following example. We would like to uniformly sample an independent set. Then we construct a Markov chain where the state space is the set of all independent sets. We start from $X_0 = \emptyset$. At each step X_t, we uniformly pick a vertex $v \in V$ and flip a uniform coin. Let $X' = X_t \cup \{v\}$ if Head is showing and let $X' = X_t \setminus \{v\}$ otherwise. Finally, let $X_{t+1} = X'$ if X' is an independent set and $X_{t+1} = X_t$ otherwise. The correctness of this algorithm is left as an exercise.

3 Hidden Markov Model

Now we consider the *hidden Markov model.* Suppose there is a Markov chain $X_0 \to X_1 \to \cdots \to X_n \to \cdots$ with the initial distribution ξ and transition probabilities $\mathbf{A}$. Each X_i is associated with a random variable Y_i where the transition probabilities are $\mathbf{B}$.

$$
\begin{array}{ccccccccccc}
\xi & & & & & & & & & & \\
\wr & & & & & & & & & & \\
X_0 & \xrightarrow{\mathbf{A}} & X_1 & \xrightarrow{\mathbf{A}} & X_2 & \xrightarrow{\mathbf{A}} & \cdots & \xrightarrow{\mathbf{A}} & X_n & \xrightarrow{\mathbf{A}} & \cdots \\
\downarrow \mathbf{B} & & \downarrow \mathbf{B} & & \downarrow \mathbf{B} & \cdots & & & \downarrow \mathbf{B} & & \\
Y_0 & & Y_1 & & Y_2 & & \cdots & & Y_n & & \cdots
\end{array}
$$

We can observe Y_i, but ξ, $\mathbf{A}$, $\mathbf{B}$ are all unknown. For example, X_i is the weather of i-th day (suppose it is unknown!), and Y_i is the behavior, such as staying home or going out. Our goal is to determin ξ, $\mathbf{A}$ and $\mathbf{B}$.

Let $\theta = (\xi, \mathbf{A}, \mathbf{B})$ be our goal, and $y = (y_0, y_1, \ldots, y_n)$ be our observation. Formally, our goal is to find

$$\arg\max_{\theta} P_\theta(y)\,,$$

where $P_\theta(y) = \mathbf{Pr}[y \mid \theta]$, and it is clear to see that $P_\theta(y) = \sum_x P_\theta(x, y)$.
We first consider the problem of computing $P_\theta(y)$ given θ and y. Here we could use *dynamic programming*. So this problem is efficiently solvable. However our goal is much more difficult to solve. Now we introduce an algorithm to compute $\arg\max_\theta P_\theta(y)$

Example 6 (Expectation Maximization Algorithm). Since y is given, $P_\theta(y)$ is a function of θ. Let $L(\theta) = P_\theta(y)$. We further have $\arg\max_\theta L(\theta) = \arg\max_\theta \log L(\theta)$.
We start from an initial θ_0, and then let

$$\theta_{t+1} = \arg\max_\theta \mathbb{E}_{\theta_t}[\log P_\theta(X, y) \mid Y = y],$$

for all $t > 0$.
Note that y is given, and thus $\mathbb{E}_{\theta_t}[\cdot]$ stands for $\mathbb{E}_{X\sim\theta_t}[\cdot]$.

We would like to justify the correctness of EM algorithm.

Lemma 2. $\mathbb{E}_{\theta_0}[P_{\theta_1}(X, y) \mid y] > \mathbb{E}_{\theta_0}[P_{\theta_0}(X, y) \mid y] \implies P_{\theta_1}(y) > P_{\theta_0}(y)$.

To prove this lemma, we should introduce *KL divergence* first.

Definition 7 (KL Divergence). Given two distributions p, q on Ω, *KL divergence* is a measure of the distance between p and q, which is given by

$$D_{\mathrm{KL}}(p, q) \triangleq \sum_i p_i \cdot \log p_i - \sum_i p_i \cdot \log q_i = \sum_i p_i \cdot \log\left(\frac{p_i}{q_i}\right).$$

Proposition 3. $D_{\mathrm{KL}}(p, q) \geq 0$.

Proof. Since $p_i, q_i > 0$, applying the inequality $\log x < x - 1$, we have that

$$-D_{\mathrm{KL}}(p, q) = \sum_i p_i \cdot \log\left(\frac{q_i}{p_i}\right) = \sum_i p_i \cdot \left(\frac{q_i}{p_i} - 1\right) = \sum_i q_i - p_i = 0.$$

□

This proof also shows that $D_{\mathrm{KL}}(p, q) = 0$ iff $p = q$.
Now we are ready to prove Lemma 2.

Proof of Lemma 2. It is equivalent to $\mathbb{E}_{\theta_0}[\log P_{\theta_1}(X, y) \mid y] > \mathbb{E}_{\theta_0}[\log P_{\theta_0}(X, y) \mid y] \implies P_{\theta_1}(y) > P_{\theta_0}(y)$.

By our assumption and Proposition 3, we have that

$$\begin{aligned}
0 &< \mathbb{E}_{\theta_0}\left[\log \frac{P_{\theta_1}(X,y)}{P_{\theta_0}(X,y)} \mid Y = y\right] \\
&= \sum_x P_{\theta_0}(x \mid y) \cdot \log \frac{P_{\theta_1}(x,y)}{P_{\theta_0}(x,y)} \\
&= \sum_x P_{\theta_0}(x \mid y) \cdot \log \frac{P_{\theta_1}(y)}{P_{\theta_0}(y)} - \sum_x P_{\theta_0}(x \mid y) \cdot \log \frac{P_{\theta_0}(x \mid y)}{P_{\theta_1}(x \mid y)} \\
&\le \log \frac{P_{\theta_1}(y)}{P_{\theta_0}(y)}.
\end{aligned}$$

□

Actually, EM algorithm does not use any information on the hidden Markov model. We now consider how to compute

$$\arg\max_{\theta} \mathbb{E}_{\theta_t}\left[\log P_\theta(X,y) \mid Y = y\right]$$

in the hidden Markov model, which is an optimization problem.

Note that

$$P_\theta(x,y) = \xi(x_0) \cdot \prod_{t=0}^{n-1} \mathbf{A}(x_t, x_{t+1}) \cdot \prod_{t=0}^{n} \mathbf{B}(x_t, y_t).$$

It follows that

$$\mathbb{E}_{\theta_0}\left[\log P_\theta(X,y) \mid y\right] = \mathbb{E}_{\theta_0}\left[\log \xi(X_0) \mid y\right] + \sum_{t=0}^{n-1} \mathbb{E}_{\theta_0}\left[\log \mathbf{A}(X_t, X_{t+1}) \mid y\right] + \sum_{t=0}^{n} \mathbb{E}_{\theta_0}\left[\log \mathbf{B}(X_t, y_t) \mid y\right].$$

We optimize the three terms on the right hand side respectively.

First,

$$\mathbb{E}_{\theta_0}\left[\log \xi(x_0) \mid y\right] = \sum_i P_{\theta_0}(X_0 = i \mid y) \cdot \log \xi(i).$$

Since $P_{\theta_0}(\cdot \mid y)$ and $\xi(\cdot)$ are two distributions, by Proposition 3, we have

$$\arg\max_{\xi(i)} \mathbb{E}_{\theta_0}\left[\log \xi(x_0) \mid y\right] = P_{\theta_0}(X_0 = i \mid y).$$

Second,

$$\begin{aligned}
\sum_{t=0}^{n-1} \mathbb{E}_{\theta_0}\left[\log \mathbf{A}(X_t, X_{t+1}) \mid y\right] &= \sum_{t=0}^{n-1} \sum_{i,j} P_{\theta_0}(X_t = i, X_{t+1} = j \mid y) \cdot \log \mathbf{A}(i,j) \\
&= \sum_i \sum_j \log \mathbf{A}(i,j) \cdot \sum_{t=0}^{n-1} P_{\theta_0}(X_t = i, X_{t+1} = j \mid y).
\end{aligned}$$

Again, to optimize the right hand side, we apply Proposition 3 and it yields that

$$\arg\max_{\mathbf{A}(i,j)} \sum_{t=0}^{n-1} \mathbb{E}_{\theta_0}\left[\log \mathbf{A}(X_t, X_{t+1}) \mid y\right] = \frac{\sum_{t=0}^{n-1} P_{\theta_0}(X_t = i, X_{t+1} = j \mid y)}{\sum_{t=0}^{n-1} P_{\theta_0}(X_t = i \mid y)}.$$

Next,

$$
\begin{aligned}
\sum_{t=0}^{n} \mathbb{E}_{\theta_0}\left[\log \mathbf{B}(X_t, y_t) \mid y\right] &= \sum_i \sum_{t=0}^{n} P_{\theta_0}(X_t = i \mid y) \cdot \log \mathbf{B}(i, y_t) \\
&= \sum_i \sum_j \log \mathbf{B}(i, j) \cdot \sum_{t:\ y_t = j} P_{\theta_0}(X_t = i \mid y)\,.
\end{aligned}
$$

Again, applying Proposition 3, it implies that

$$
\arg\max_{\mathbf{B}(i,j)} \sum_{t=0}^{n} \mathbb{E}_{\theta_0}\left[\log \mathbf{B}(X_t, y_t) \mid y\right] = \frac{\sum_{t:\ y_t = j} P_{\theta_0}(X_t = i \mid y)}{\sum_{t=0}^{n} P_{\theta_0}(X_t = i \mid y)}\,.
$$

Finally, combining all of above together we obtain $\theta^* = \arg\max \mathbb{E}_{\theta_0}\left[\log P_\theta(X, y) \mid y\right]$.

Now the remaining problem is to compute $\xi(i)$, $\mathbf{A}(i, j)$ and $\mathbf{B}(i, j)$ given θ_0. Clearly, it is sufficient to compute $P_{\theta_0}(X_t = i, X_{t+1} = j \mid y)$ efficiently.

This problem is also solvable by *dynamic programming*, and the whole algorithm is left as an exercise again.