10-801: Advanced Topics in Graphcal Models 10-801, Spring 2007

Monte Carlo Methods

Lecturer: Eric P. Xing *Scribe: Wenjie Fu*

Monte Carlo EM

$$D = \{X\}, \quad X \sim P(\ \mid \theta)$$

In M-step,

$$\begin{aligned} \theta &= \arg\max P(D \mid \theta) \\ &= \arg\max P(X_H, X_V \mid \theta) \end{aligned}$$

X_H, X_V denotes hidden and visible data respectively.

$$\begin{aligned} L(\theta) &= \langle\ \log P(X_H, X_V \mid \theta)\ \rangle_{P(X_H \mid X_V, \theta)} \\ X_H^{(n)} &\sim P(X_H \mid X_V, \theta) \\ L(\theta) &= \int \log P(X_H, X_V \mid \theta)\ P(X_H \mid X_V, \theta) \mathrm{d}X_H \\ &\approx \frac{1}{M} \sum_{m=1}^{M} \log P(X_H^{(m)}, X_V \mid \theta) \\ \theta^{\mathrm{ML}} &= \arg\max_\theta \sum_{m=1}^{M} \log P(X_H^{(m)}, X_V \mid \theta) \end{aligned}$$

When $M = 1$, it becomes stochastic EM.

Data Augmentation

For Bayesian Inference, we want to get

$$\begin{aligned} &P(\theta \mid X_V) \\ &\theta \sim P(\theta \mid X_V) \end{aligned}$$

But it is hard to margin out X_H. Suppose

$$\theta \sim P(\theta \mid X_H, X_V)$$

is easy.

$$P(\theta \mid X_V) = \int P(\theta \mid X_H, X_V) P(X_H \mid X_V) \mathrm{d}X_H$$
$$P(X_H \mid X_V) = \int P(X_H \mid \theta, X_V) P(\theta \mid X_V) \mathrm{d}\theta$$

Initially, suppose we have samples $\theta^{(1)}$, $\theta^{(2)}$, ..., $\theta^{(m)}$, which could be interpreted as an estimation of distribution of θ given X_V.

In I-step (I for Imputation), draw

$$X_H^{(m)} \sim P(X_H \mid \theta^{(m)}, X_V)$$

In P-step (P for Posterior), draw

$$\theta^{(m)} \sim P(\theta \mid X_H^{(m)}, X_V)$$

Repeat the I-step and P-step iteratively.
Finally,

$$P(\theta \mid X_V) \approx \frac{1}{M} \sum P(\theta \mid X_H^{(m)}, X_V)$$

Invariant Distribution

From this point, we will not distinguish X_H and θ, viewing them both as hidden variables.

How to draw sample?

$$X \sim P(X)$$
$$X^{(1)}, X^{(2)}, \ldots, X^{(m)} \text{ are a sequence of samples.}$$

Assume X follows a Markov Chain, so that

$$X^{(t)} \sim P(X \mid X^{(t-1)}, \ldots, X^{(1)}) = P(X \mid X^{(t-1)})$$
$$X^{(1)} \sim P_0(X)$$

Let T_m denotes the transition probability of the Markov Chain, that is

$$T_m(X^{(m)}, X^{(m+1)}) = P(X^{(m+1)} \mid X^{(m)})$$

Homogeneous Markov Chain: $T_m = T$

Definition

Define **Invariant Distribution** (ID):
P is an ID with respect to T if

$$P(X) = \sum_{X'} T(X', X) P(X')$$

Detail Balance

If $P(X)T(X,X') = P(X')T(X',X)$, then P is an ID with respect to T.

Proof.

$$\begin{aligned} P(X) &= P(X)\sum_{X'} T(X,X') \\ &= \sum_{X'} P(X)T(X,X') \\ &= \sum_{X'} P(X')T(X',X) \end{aligned}$$

According to the definition of I.D., it is proved. □

Ergodicity

Ergodicity: Sampling using T, and starting from P_0. If $P_m \to P$, when $m \to \infty$.

In order to have Ergodicity, P must be I.D. with respect to T. (Necessary condition)

In addition, if $T(X,X') > 0$, then Ergodicity must be held. (Sufficient condition)

Theorem If P^* is I.D. with respect to T,

$$\begin{aligned} P_{m+1}(X) &= \sum_{X'} T(X',X)P_m(X') \\ \gamma &= \min_{X'} \min_{X:P^*(X)>0} \frac{T(X',X)}{P^*(X)} > 0 \\ \text{Then, } &\lim_{m\to\infty} P_m(X) = P^*(X) \end{aligned}$$

Proof. Prove by induction. Suppose

$$P_m(X) = [1-(1-\gamma)^m]P^*(X) + (1-\gamma)^m r_m(X)$$

r_m is an arbitary distribution. Since $\gamma < 1$, it is a convex combination.

1) $m = 0$, let $r_0 = p_0$

2)

$$
\begin{aligned}
P_{m+1}(X) &= \sum_{X'} T(X', X) P_m(X') \\
&= [1-(1-\gamma)^m] \sum_{X'} T(X', X) P^*(X') + (1-\gamma)^m \sum_{X'} T(X', X) r_m(X') \\
&= [1-(1-\gamma)^m] P^*(X) + (1-\gamma)^m \sum_{X'} r_m(X') [T(X', X) + \gamma P^*(X) - \gamma P^*(X)] \\
&= [1-(1-\gamma)^{m+1}] P^*(X) + (1-\gamma)^m \sum_{X'} r_m(X') \frac{T(X', X) - \gamma P^*(X)}{1-\gamma} \\
r_{m+1} &= \sum_{X'} r_m \frac{T(X', X) - \gamma P^*(X)}{1-\gamma}
\end{aligned}
$$

By definition of γ, r_{m+1} is a distribution. (Sum to 1 and be all non-negative.) □

Metropolis Hasting

$$
\begin{aligned}
X' &\sim q(X, X') = P(X' \mid X) \\
A(X, X') &= \min(1, \frac{P(X')}{P(X)} \frac{q(X', X)}{q(X, X')})
\end{aligned}
$$

Accept X' with probability $A(X, X')$. That is, if yes, $X^{t+1} = X'$, otherwise $X^{t+1} = X$ (X^t)

$$
T(X, X') = q(X, X') A(X, X')
$$

We can show $T(X, X')$ satisfied Detail Balance.

Proof.

$$
\begin{aligned}
P(X) T(X, X') &= P(X) q(X, X') A(X, X') \\
&= P(X) q(X, X') \min(1, \frac{P(X')}{P(X)} \frac{q(X', X)}{q(X, X')}) \\
&= \min(P(X) q(X, X'), P(X') q(X', X))
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
P(X') T(X', X) &= P(X') q(X', X) A(X', X) \\
&= \min(P(X') q(X', X), P(X) q(X, X')) \\
&= P(X) T(X, X')
\end{aligned}
$$

Proof done. □

Therefore, $P(X)$ is I.D. with respect to T. We can evaluate $P(X)$ on $\tilde{P}(X)$.