Lecture 9. Dynamic Programming: Sampling & Expectation

Note Title 4/17/2011

$$\Pi(\underline{x}) = \frac{1}{Z} e^{-E[\underline{x}]}, \quad E[\underline{x}] = \sum_{i=1}^{N} \varphi_i(x_{i-1}, x_i)$$

$x_0 \quad x_1 \quad \ldots \quad x_{N-1} \quad x_N$

Previous lecture showed how to use DP to estimate $\hat{\underline{x}} = \arg\min_{\underline{x}} E[\underline{x}]$, and compute $E[\hat{\underline{x}}]$

Forward Pass computes $E[\hat{\underline{x}}]$
Backward Pass gives $\hat{\underline{x}}$.

This lecture will show how to use DP to obtain samples from $\Pi(\underline{x})$, to compute Z, and to $\sum_{\underline{x}} h(\underline{x})\Pi(\underline{x})$ (sometimes by sampling, sometimes exactly).

Note: this is a different algorithm that the one used to compute $\hat{\underline{x}} = \arg\min_{\underline{x}} E[\underline{x}]$ — but it is closed related and is based on the same principles.

x

(1) Sampling from $\Pi(\underline{x})$

Claim — we can use DP to convert $\Pi(\underline{x}) = \frac{1}{Z} e^{-E(\underline{x})}$ into an alternative form $\Pi_0(x_0)\Pi_1(x_1|x_0)\ldots\Pi_N(x_N|x_{N-1})$

or, equivalently, $\Pi_0(x_0|x_1)\Pi_1(x_1|x_2)\ldots\Pi_{N-1}(x_{N-1}|x_N)\Pi_N(x_N)$

(not the same Π's in these two cases).

If we can express $\Pi(\underline{x})$ in one of these two forms then we can obtain samples $\underline{x}^1, \underline{x}^2, \ldots \underline{x}^M$ as follows

x_0^1 — sample from $\Pi_0(x_0)$	x_0^2 — from $\Pi_0(x_0)$
x_1^1 ~ sample from $\Pi_1(x_1\|x_0^1)$	x_1^2 — from $\Pi_1(x_1\|x_0^2)$
x_2^1 — sample from $\Pi_2(x_2\|x_1^1)$	x_2^2 — from $\Pi_2(x_2\|x_1^2)$
⋮	⋮
x_N^1 ~ sample from $\Pi_N(x_N\|x_{N-1}^1)$	x_N^2 — from $\Pi_N(x_N\|x_{N-1}^2)$

Or x_N^1 ~ from $\Pi_N(x_N)$, x_{N-1}^1 ~ from $\Pi_{N-1}(x_{N-1}|x_N^1)$, etc

So only need to sample from $\Pi(x_i|x_{i+1})$ or $\Pi(x_{i+1}|x_i)$
~ use techniques from earlier lectures.

Page 2 How to convert from $\pi(\underline{x}) = e^{-E(\underline{x})}/Z$

Note Title 4/2/2006

to these forms - e.g. $\pi_N(x_N)\pi_{N-1}(x_{N-1}|x_N) \ldots \pi_0(x_0|x_1)$?

Special case: if $N=1$ $\pi(\underline{x}) = \pi(x_0, x_1)$

then it is easy, $\pi(x_0,x_1) = \pi(x_0|x_1)\pi(x_1)$

$= \pi(x_1|x_0)\pi(x_0)$.

Note: Markov Condition $\Rightarrow$ the graph structure of this model (due to potentials $\varphi_i(x_{i-1},x_i)$ means that the model has a local Markov structure. $\Rightarrow$ e.g. $\pi(x_i|x_{i-1},x_{i-2} \ldots x_0) = \pi(x_i|x_{i-1})$

If we know x_{i-1}, then knowing $x_{i-2}, \ldots x_0$ also gives no more information about x_i.

DP Algorithm: To compute $\pi_0(x_0|x_1) \ldots \pi_{N-1}(x_{N-1}|x_N)\pi_N(x_N)$ from $\pi(\underline{x})$

- Define $V_1(x_1) = \sum_{s \in S} e^{-\varphi_1(s,x_1)}$
- Recursively compute for $i = 2, \ldots, N$

$$V_i(x_i) = \sum_{y \in S} V_{i-1}(y)\, e^{-\varphi_i(y,x_i)}$$

Then we compute efficiently (in $O(k^2N)$):

(A) The normalization constant: $Z = \sum_{x_N \in S} V_N(x_N)$

(B) The marginal $\pi_N(x_N) = V_N(x_N)/Z$

(C) The conditionals $\pi_i(x_i|x_{i+1}) = \dfrac{V_i(x_i)\, e^{-\varphi_{i+1}(x_i,x_{i+1})}}{\sum_y V_i(y)\, e^{-\varphi_{i+1}(y,x_{i+1})}}$

(Note: easy to modify algorithm to compute $\pi_0(x_0)\pi_1(x_1|x_0) \ldots \pi_N(x_N|x_{N-1})$.

To estimate $I = \sum_{\underline{x}} h(\underline{x}) \pi(\underline{x})$,

where $h(\underline{x})$ is a quantity you want to estimate, like previous lectures.

Obtain samples $\underline{x}^1, \underline{x}^2, \dots \underline{x}^m$ from $\pi(\underline{x})$

Estimator : $I_m = \frac{1}{m} \sum_{i=1}^{m} h(\underline{x}^i)$. (as before)

Unbiased $\langle I_m \rangle = E_\pi I_m = I$.

efficiency $\sim \frac{1}{m} Var_\pi h(\underline{x})$.

But, if $h(\underline{x})$ takes certain special forms, then we can use DP to compute $\sum_{\underline{x}} h(\underline{x}) \pi(\underline{x})$ exactly.

Observe that: $V_1(x_1) = \sum_{x_0} e^{-\varphi_1(x_0, x_1)}$, $V_2(x_2) = \sum_{x_0, x_1} e^{-\varphi_1(x_0,x_1) - \varphi_2(x_1, x_2)}$

$V_N(x_N) = \sum_{x_0, x_1, \dots x_{N-1}} e^{-\varphi_1(x_0,x_1) - \dots - \varphi_N(x_{N-1}, x_N)}$ → computed by DP (on Page 2)

Suppose $h(\underline{x}) = h_0(x_0) + \dots + h_N(x_N)$

Then to compute $\sum_{\underline{x}} h(\underline{x}) \pi(\underline{x})$ requires:

(i) computing Z, done by DP on page 2

(ii) for each i, compute $\sum_{x_0, x_1 \dots, x_N} h_i(x_i) e^{-\varphi_1(x_0,x_1) - \dots - \varphi_N(x_{N-1}, x_N)}$

→ done by modifying DP on page 2.

e.g. $\tilde{V}_i(x_i) = \sum_{y \in S} \tilde{V}_{i-1}(y) e^{-\varphi_i(y, x_i)} \underbrace{h_i(x_i)}_{\text{new}}$

other updates as before.

This requires more computation $- O(k^2 N^2)$

$O(k^2 N)$ for each $h_i(.)$

We can extend this in the obvious way if

$h(\underline{x}) = h_1(x_0, x_1) + \dots + h_N(x_{N-1}, x_N)$

e.g. $\tilde{V}_i(x_i) = \sum_{y \in S} \tilde{V}_{i-1}(y) e^{-\varphi_i(y, x_i)} h_i(y, x_i)$

for computing $\sum_{x_0, x_1 \dots, x_N} h_i(x_{i-1}, x_i) e^{-\varphi_1(x_0,x_1) - \dots - \varphi_N(x_{N-1}, x_N)}$. //

Note: this can be extended to other cases — e.g. term like $h_i(x_{i-1}, x_{i-2}, x_i)$, but gets more complicated.

page 4

Special Case:

Note $(-1)^2 = 1 = 1^2$
so most probable states are $\underline{x} = (1, 1, \ldots 1)$ and $\underline{x} = (-1, -1, \ldots -1)$

Ising Spin Model

$$\pi(\underline{x}) = \frac{1}{Z} e^{\beta(x_0 x_1 + \ldots + x_{N-1} x_N)}.$$

$x_i \in \{-1, +1\}$

$$V_1(x_1) = e^{\beta x_1} + e^{-\beta x_1} = e^{\beta} + e^{-\beta}$$

(Note: this is a very special case, usually $V_1(x_1)$ is a function of x_1)

$$V_2(x_2) = \sum_{y \in S} V_1(y)\, e^{\beta y x_2}$$

$$= \{e^{\beta} + e^{-\beta}\} \sum_{y \in S} e^{\beta y x_2} = \{e^{\beta} + e^{-\beta}\}^2.$$

In general, $V_t(x_t) = \{e^{\beta} + e^{-\beta}\}^t$ // $V_d(x_d) = \{e^{\beta} + e^{-\beta}\}^d$

Hence $Z = \sum_{x_N \in S} V_N(x_N) = 2\{e^{\beta} + e^{-\beta}\}^N$

$S = \{-1, +1\}$

Marginal Density $\pi(x_N) = \frac{V_N(x_N)}{Z} = \frac{1}{2}$

Conditional Distribution:

$$\pi_t(x_t \mid x_{t+1}) = \frac{\{e^{\beta} + e^{-\beta}\}^t e^{\beta x_t x_{t+1}}}{\sum_{y \in S} \{e^{\beta} + e^{-\beta}\}^t e^{\beta y x_{t+1}}}$$

$$\pi_t(x_t \mid x_{t+1}) = \frac{e^{\beta x_t x_{t+1}}}{e^{\beta} + e^{-\beta}} \quad //$$

The Ising Model was invented by Physicists to study phase transitions → e.g. how does ice become water at a criterical temperature $T = 0°$ centigrade. $\beta = 1/T$.

For small β (high temperature T), the Ising model is disordered, samples from Ising will be like

$1\ -1\ -1\ 1\ 1\ 1\ 1\ -1\ 1\ -1\ -1\ 1$.

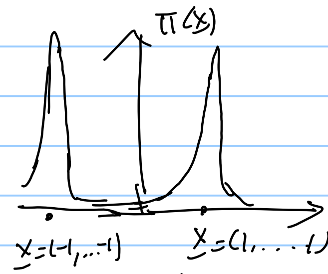
For large β (small temperature T), the Ising model is ordered, samples will tend to be either

$1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1$ i.e. $x_0 = x_1 = \ldots = x_n$

or $-1\ -1\ -1 \qquad -1$

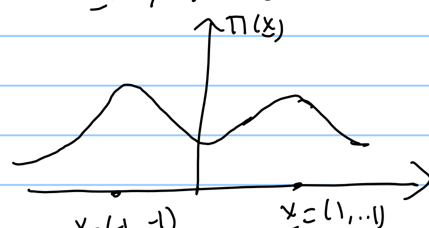
At the critical temperature $T_c = 1/\beta_c$, the model will change from order to disorder (e.g. from Ice to Water)

At low T.



sharp peaks in $\pi(\underline{x})$ at $\underline{x} = (1, \ldots 1)$ and $\underline{x} = (-1, \ldots -1)$

At high T.



peaks in $\pi(\underline{x})$ at the same places $\underline{x} = (1, \ldots, 1)$ & $\underline{x} = (-1, \ldots -1)$.

but $\pi(\underline{x})$ is also quick large at other places.

Samples from $\pi(\underline{x})$ at low temperature will be almost all at $(1, 1, \ldots, 1)$ or $(-1, \ldots, -1)$ the ordered states

Samples at high temperature will also occur at the disordered states – e.g. $(1, -1, 1, 1, -1, -1 \ldots)$.

They will still be most probable at $(1, 1, \ldots 1)$ or $(-1, \ldots -1)$ but there are many more disordered states (exponentially more).

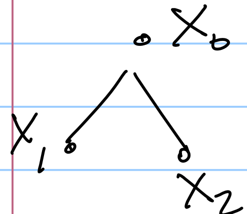
So we expect that a sample will probably be a disordered state.

page 6

DP can be extended to apply to any graph without closed loops.

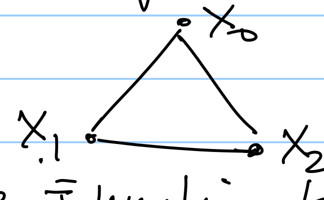
(Both the m computations in previous lecture and the V computations in this lecture.)

E.G.



yes for DP.

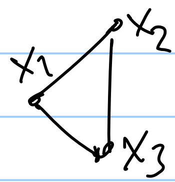
But not if there are closed loops.



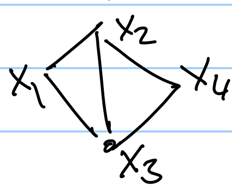
No for DP

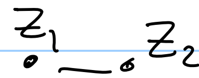
But there are ways — the Junction tree algorithm — which allows us to convert a prob. model on a graph with closed loops into a new model without closed loops — by augmenting the variables

E.G.



Define new variable $Z_1 = (X_1, X_2, X_3)$ $\quad$ Z_1





$Z_1 = (X_1, X_2, X_3)$

$Z_2 = X_4$

etc.

But augmenting variables may make DP impractical. It may require adding extra nodes (i.e. making N very large). or make the number of state values k too large.

E.G. If X_i has k possible values $(S_1, S_2 \ldots S_k)$ then $Z_1 = (X_1, X_2, X_3)$ has k^3 possible values. Beyond Scope of course.

Also, the m-update and V-update algorithms can be applied to graphs with closed loops as approximation. This gives belief propagation (sum-product & sum-max)