Probabilistic Foundations of Artificial Intelligence

Solutions to Problem Set 4

Nov 11, 2016

1. Bayesian networks and Markov chains

Consider the query $P(R|S = t, W = t)$ in the Bayesian network on Slide 9 of https://las.inf.ethz.ch/courses/pai-f16/slides/pai-06-bayesian-networks-sampling-annotated.pdf and how Gibbs sampling can answer it.

(i) How many states does the Markov chain have?

(ii) Calculate the transition matrix T containing $P(X_{t+1} = y \mid X_t = x)$ for all x, y.

(iii) What does T^2, the square of the transition matrix, represent?

(iv) What about T^n as $n \to \infty$?

(v) Explain how to do probabilistic inference in Bayesian networks, assuming that T^n is available. Is this a practical way to do inference?

Solution

(i) There are two uninstantiated Boolean variables (*Cloudy* and *Rain*) and therefore four possible states.

(ii) First, we compute the sampling distribution for each variable, conditioned on its Markov blanket.

$$
\begin{aligned}
P(C|r, s) &= \frac{1}{Z} P(C)P(s|C)P(r|C) \\
&= \frac{1}{Z} \langle 0.5, 0.5\rangle \langle 0.1, 0.5\rangle \langle 0.8, 0.2\rangle = \frac{1}{Z} \langle 0.04, 0.05\rangle = \langle 4/9, 5/9\rangle \\
P(C|\neg r, s) &= \frac{1}{Z} P(C)P(s|C)P(\neg r|C) \\
&= \frac{1}{Z} \langle 0.5, 0.5\rangle \langle 0.1, 0.5\rangle \langle 0.2, 0.8\rangle = \frac{1}{Z} \langle 0.01, 0.2\rangle = \langle 1/21, 20/21\rangle \\
P(R|c, s, w) &= \frac{1}{Z} P(R|c)P(w|s, R) \\
&= \frac{1}{Z} \langle 0.8, 0.2\rangle \langle 0.99, 0.9\rangle = \frac{1}{Z} \langle 0.792, 0.18\rangle = \langle 22/27, 5/27\rangle \\
P(R|\neg c, s, w) &= \frac{1}{Z} P(R|\neg c)P(w|s, R) \\
&= \frac{1}{Z} \langle 0.2, 0.8\rangle \langle 0.99, 0.9\rangle = \frac{1}{Z} \langle 0.198, 0.72\rangle = \langle 11/51, 40/51\rangle
\end{aligned}
$$

Strictly speaking, the transition matrix is only well-defined for the variant of MCMC in which the variable to be sampled is chosen randomly[1]. (In the variant where the variables are chosen in a fixed order, the transition probabilities depend on where we are in the ordering.) Now consider the transition matrix.

- Entries on the diagonal correspond to self-loops. Such transitions can occur by sampling *either* variable. For example, for the self-loop on (c, r), we obtain:

  $$t((c, r) \to (c, r)) = 0.5P(c|r, s) + 0.5P(r|c, s, w) = 17/27,$$

  where the two factors of 0.5 are corresponding to the probability that the variables to be sampled are C and R, respectively.
- Entries where one variable is changed must sample that variable. For example,

  $$t((c, r) \to (c, \neg r)) = 0.5P(\neg r|c, s, w) = 5/54$$

- Entries where both variables change cannot occur. For example,

  $$t((c, r) \to (\neg c, \neg r)) = 0$$

This gives us the following transition matrix T, where the transition is from the state given by the row label to the state given by the column label:

$$\begin{array}{c c} & \begin{array}{cccc} (c,r) & (c,\neg r) & (\neg c, r) & (\neg c,\neg r) \end{array} \\ \begin{array}{c} (c,r) \\ (c,\neg r) \\ (\neg c, r) \\ (\neg c,\neg r) \end{array} & \begin{pmatrix} 17/27 & 5/54 & 5/18 & 0 \\ 11/27 & 22/189 & 0 & 10/21 \\ 2/9 & 0 & 59/153 & 20/51 \\ 0 & 1/42 & 11/102 & 310/357 \end{pmatrix} \end{array}$$

(iii) T^2 represents the probability of going from each state to each state in two steps.

(iv) T^n (as $n \to \infty$) represents the long-term probability of being in each state starting in each state; for ergodic T these probabilities are independent of the starting state, so every row of T is the same and represents the posterior distribution over states given the evidence.

(v) We can produce very large powers of T with very few matrix multiplications. For example, we can get T^2 with one multiplication, T^4 with two, and T^{2^k} with k. Unfortunately, in a network with n non-event Boolean variables, the matrix is of size $2^n \times 2^n$, so each multiplication takes $O(2^{3n})$ operations.

2. Gibbs sampling

See `.zip` file on course website.

[1] Slide 33 of https://las.inf.ethz.ch/courses/pai-f16/slides/pai-06-bayesian-networks-sampling-annotated.pdf

3. Markov chains and detailed balance

Assume that you are given a Markov chain with state space Ω and transition matrix T, which is defined for all $x, y \in \Omega$ and $t \geq 0$ as $T(x, y) := P(X_{t+1} = y \mid X_t = x)$. Furthermore, let π be the stationary distribution of the chain.

(i) Show that, if for some t the current state X_t is distributed according to the stationary distribution and additionally the chain satisfies the detailed balance equations

$$\pi(x)T(x, y) = \pi(y)T(y, x), \text{ for all } x, y \in \Omega,$$

then the following holds for all $k \geq 0$ and $x_0, \ldots, x_k \in \Omega$:

$$P(X_t = x_0, \ldots, X_{t+k} = x_k) = P(X_t = x_k, \ldots, X_{t+k} = x_0).$$

(This is why a chain that satisfies detailed balance is called *reversible*.)

(ii) Show that, if T is a symmetric matrix, then the chain satisfies detailed balance, and the uniform distribution on Ω is stationary for that chain.

Solution

(i) We use the chain rule, as well as the detailed balance condition:

$$\begin{aligned}
&P(X_t = x_0, \ldots, X_{t+k} = x_k) \\
&= P(X_t = x_0)P(X_{t+1} = x_1 \mid X_t = x_0) \ldots P(X_{t+k} = x_k \mid X_{t+k-1} = x_{k-1}) && \text{ch. rule} \\
&= \pi(x_0)T(x_0, x_1) \ldots T(x_{k-1}, x_k) && X_t \sim \pi \\
&= T(x_1, x_0)\pi(x_1) \ldots T(x_{k-1}, x_k) && \text{detailed balance} \\
&= \ldots && \vdots \\
&= T(x_1, x_0) \ldots T(x_k, x_{k-1})\pi(x_k) && \text{detailed balance} \\
&= \pi(x_k)T(x_k, x_{k-1}) \ldots T(x_1, x_0) \\
&= P(X_t = x_k)P(X_{t+1} = x_{k-1} \mid X_t = x_k) \ldots P(X_{t+k} = x_0 \mid X_{t+k-1} = x_1) && X_t \sim \pi \\
&= P(X_t = x_k, \ldots, X_{t+k} = x_0). && \text{ch. rule}
\end{aligned}$$

(ii) By definition of a symmetric matrix, we have that $\pi(x)T(x, y) = \pi(x)T(y, x)$, for all $x, y \in \Omega$. Therefore, if $\pi(x) = \frac{1}{|\Omega|}$, for all $x \in \Omega$, then $\pi(x)T(x, y) = \pi(y)T(y, x)$, which means that detailed balance holds for the chain and the uniform distribution is stationary.