L06 – Approximate Inference via Sampling

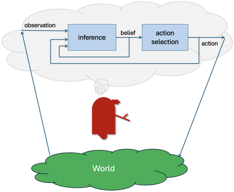
AIMA 13.4

What you should know after this lecture

- Ancestral sampling in directed models
- Gibbs sampling in Bayes nets and factor graphs
- Intro to more general MCMC methods

Probabilistic belief representation

- Belief is a probability distribution : $B \in \mathcal{P}(\mathcal{S})$
  (an element of the set of all distributions over $\mathcal{S}$)
- Important questions:
  - What is $p_B(\text{event})$?
  - What is the most likely state $\text{argmax}_s\, p_B(s)$?



Approximate inference

What if your network is highly connected? Exact inference is expensive.
Approximations exist!

- In a loopy factor graph, perform multiple rounds of message passing
    - Not guaranteed to converge
    - If Gaussian loopy BP converges, the means will be correct; under some conditions, the covariances will be, as well.
    - Otherwise, not too much we can say.
- Sampling methods try to draw samples from $P(X_1, \ldots, X_n)$ and compute answers to queries from them. Generally they are consistent: estimates converge in the limit to the true answers, but can take a long time

Sampling in Bayes nets

Easy to do ancestral sampling to get samples of any unconditional marginal or joint $\bar{v} \sim P_\alpha(\bar{V})$ when α is a BN

1. Sort nodes in α into topological order so that all nodes $\mathrm{pa}(V)$ come before V in the ordering.
2. For $i = 1$ to M `// number of samples
   - For $j = 1$ to N `// number of nodes in network
     $x^i_j = \mathrm{sample}(P_\alpha(V_j \mid \mathrm{pa}(V_j) = x^i_j[\mathrm{pa}(V_j)])))$
3. Use $\{x^i\}_{i=1..M}$ to estimate whatever you want, e.g.

$$\hat{P}_\alpha(V_k = 1, V_j = 0) = \frac{1}{M}\mathbb{I}(x^i_k = 1 \wedge x^i_j = 0)$$

where $\mathbb{I}(a) = 1$ if a else 0

Conditional sampling

What if we want $P_\alpha(V \mid E = e)$? Two general approaches:

- Rejection sampling: do ancestral sampling, but throw away all examples in which $E \neq e$.
  - Can be very slow if $P(E = e)$ is small.
- Importance sampling: sample from an easier distribution Q, but reweight the samples to compute your result
  - Let Q be a distribution over same domain as desired distribution P and $\{x^i\}_{1,\dots,M} \sim Q(x)$
  - Then,

$$E_{x \sim P}[f(x)] \approx \frac{1}{Z} \sum_{x^i \sim Q} \frac{P(x^i)}{Q(x^i)} f(x^i)$$

  - Necessary that $Q(x) > 0$ for any x where $P(x) > 0$.
  - Have to be able to evaluate $P(x)$ and $Q(x)$
  - If P and Q are very different, you will need large M to get a good estimate.

Bayes net importance sampling

In Bayes nets, let $Z = V \setminus E$ (the unobserved nodes)

- Fix all $E = e$ and then use ancestral sampling to get samples from

$$Q(Z) = \prod_i P(Z_i \mid \text{pa}(Z_i))$$

- Importance weights

$$\frac{P(z \mid e)}{Q(z)} \propto \frac{P(z, e)}{Q(z)} = \frac{\prod_i P(z_i \mid \text{pa}(Z_i)) \prod_j P(e_j \mid \text{pa}(E_j))}{\prod_i P(z_i \mid \text{pa}(Z_i)}$$
$$= \prod_j P(e_j \mid \text{pa}(E_j))$$

The name importance sampling also used in a context of sampling from continuous distributions for a different method.

Markov chains

- set $\mathcal{S}$ of states; S_t is a random variable representing the state at time t; $s_i \in \mathcal{S}$ is a possible state
- initial state distribution $p(S_0) = \pi_0$ (row vector)
- transition distribution $P(S_t = s_j \mid S_{t-1} = s_i) = P_{ij}$

We can ask questions like:

- If $S_0 \sim \pi_0$, what is the distribution on S_1?
  Ans: $\pi_0 P$ (check dimensions! be sure it makes sense!)
- What's the probability it will hit s_9 before it hits s_3?
- What is the limiting behavior?
  - Could absorb into a single state and never escape
  - Could enter a deterministic cycle
  - Could have a stationary distribution: $\pi = \lim_{t\to\infty} P^t$ with the property that $\pi P = \pi$ independent of π_0
    Guaranteed if no 0's in P (read about ergodicity)
  - Fun facts: π is an eigenvector of P and the second largest eigenvalue governs the convergence rate

Gibbs sampling

Can be applied in Bayes nets but easiest to think of in factor graphs. Simple type of Markov chain Monte Carlo (MCMC).

- Define a Markov chain where
    - States are assignments of values to all variables
    - The stationary distribution of the chain is the desired distribution $P(V_1, \ldots, V_n)$
    - Samples will be identically distributed but not necessarily independent (because temporally correlated)
- To do estimation:
    - Run the chain for a while and throw those samples away ("burn in" phase) so we are in the stationary distribution
    - Keep (every kth) sample and use them to estimate the quantity of interest

Gibbs sampling in graphical model

1. Initialize values $\bar{v} = (v_1, \ldots, v_N)$ at random
2. Loop
   - Choose i randomly from $1, \ldots, N$
   - Set v_i to a sample from

     $$P(V_i \mid (V \setminus V_i) = (\bar{v} \setminus v_i)) = P(V_i \mid \text{mb}(V_i)) = \bar{v}_{\text{mb}(V_i)})$$

     where $\text{mb}(V_i)$ is the set of variables in the Markov blanket of V_i and $v_{\text{mb}(V_i)}$is their values in the current assignment $\bar{v}$
3. Use the $\bar{v}$ samples to estimate quantity of interest.

Markov blanket: In a factor graph, it's the neighboring nodes

$$P(V_i \mid \text{mb}(V_i) = \text{mb}(v_i)) = \prod_{\phi \in N(V_i)} \phi[\text{mb}(v_i)]$$

where $\phi[\text{mb}(v_i)]$ is the vector of values for variable V_i that remains after selecting the other dimensions of factor ϕ to have their associated values in $\text{mb}(v_i)$.

Example: one step

Assume a factor graph with binary variables A, B, C, D and

- $\phi_{AB} = [[1, 10], [10, 1]]$
- $\phi_{AC} = [[1, 2], [3, 4]]$
- $\phi_{AD} = [[5, 2], [1, 1]]$

Assume

- the current assignment is $(1, 1, 1, 0)$
- we pick variable A to update
- we construct a distribution on A by
  - Finding all the factors mentioning A
  - For each of those factors, use the current assignments $B = 1$, $C = 1$, $D = 0$ to select out each factor's opinion about A
  - This gives us $[10, 1]$, $[2, 4]$, $[5, 1]$
  - Table multiplication gives us $[100, 4]$
- with probability $100/104$, we change A to have value 0

Gibbs sampling properties

If there are no 0 entries in the factors then, the chain is ergodic, which means

- the chain is aperiodic
- every state is reachable with non-zero probability from every other state

It's not too hard to prove that the joint distribution encoded by the network is the stationary distribution of the Markov chain induced by Gibbs sampling.

Fine in discrete / continuous, loopy graphs.

Read about Metropolis/Hastings (a more general class of algorithms) in AIMA4e to learn more.

Next time

- Hidden Markov models
- Start Kalman filters