Gibbs Sampling

In this section, we describe Gibbs sampling, a general method for probabilistic inference. Gibbs sampling is well suited to coping with incomplete information and is often suggested for such applications. However, generality comes at some computational cost, and for many applications including those involving missing information there are often alternative methods that have been shown to be more efficient in practice. Nevertheless, understanding Gibbs sampling provides some valuable insights into the problems of statistical inference. The description given below proceeds in three parts.

Part I

We begin with some basic ideas about stochastic processes. Consider a Markov process with states Q = {1,2,...,N} and transition probabilities {p_{ij}}. In our examples, the state space is finite but this requirement is not necessary for Gibbs sampling to be applicable. Assume that the process is ergodic, meaning that any state is reachable from any other state with probability greater than zero. Let q_t indicate the state of the process after t steps and q_0 indicate the initial state.

Let {pi_i} be the long-run frequencies for the process. That is pi_i = Pr(q_t=i) for large t. Note that the most likely state after t steps for large t is arg max_i pi_i. One simple way to estimate the most likely state is to run the process for a large number of steps starting from any initial state, keep track of the number of times the process is in each state, and use the corresponding frequencies to estimate the {pi_i}. Ergodicity is what allows us to start the process in an arbitrary initial state.

Part II

Now suppose X = {X_1,X_2,...,X_M} is a collection of random variables and Pr(X) is the joint distribution on X with no probabilities equal to zero or one. In the following, we will treat these variables as state variables for a particular Markov process. We define the Markov process as follows. Let

```
S = Product_{i=1}^M Omega_{X_i}
```

where Omega_{X_i} is the set of values for X_i. Let X_{i,t} represent a random variable corresponding to the ith state variable at time t. Define the state at time t as follows.

```
q_t = (X_{1,t}=x_{1,t},X_{2,t}=x_{2,t},...,X_{M,t}=x_{m,t}).
```

Define the transition probabilities

```
Pr(X_{i,t+1}|q_t) = Pr(X_i|{X_j : j != i})
```

where Pr(X_i|{X_j : j != i}) is obtained from the joint distribution Pr(X). To find the maximum a posteriori (MAP) value of X_i according to Pr(X) run the process a large number of steps starting in some arbitrary initial state and estimate as described in Part I.

That is basically all there is to Gibbs sampling. In Part III, we extend the idea to Bayesian networks and learning problems, but the extension is pretty simple. Geman and Geman [1984] place the idea of Gibbs sampling in a general setting in which the collection of variables is structured in a graphical model and each variable has a neighborhood corresponding to a local region of the graphical structure. Geman and Geman use the Gibbs distribution to define the joint distribution on this structured set of variables. In the case of Bayesian networks, the neighborhoods correspond to the Markov blanket of a variable [Pearl, 1988] and the joint distribution is defined by the factorization of the network.

Part III

Now suppose that Pr(X) is defined by a Bayesian network.

```
Pr(X) = Product_{i=1}^M Pr(X_i|Parents(X_i)
```

Then Pr(X_{i,t+1}|{X_{j,t} : j != i}) is determined as follows.

```
Pr(X_i|{X_j : j != i}) =
    Pr(X_i|Parents(X_i))
        Product_{X_j : X_i in Parents(X_j)}
            Pr(X_j|Parents(X_j))
```

Notice that computing Pr(X_i|{X_j : j != i}) can be done using only local information. Assuming that max_i |Parents(X_i)| is bounded by a constant, then sampling from Pr(q_{t+1}|q_t) can be performed in time linear in M, the number of variables in the Bayesian network. The extension to using learning Bayesian networks is treated in the exercises.

Exercises

1. Gibbs sampling is a general inference algorithm.

Consider the following network.

```
X_1              X_2              X_3
 o--------------->o--------------->o
```

along with the distributions Pr(X_2|X_1), Pr(X_3|X_2), and Pr(X_1). An assignment to the X_i might correspond to a patient history such that X_1=1 indicates that the patient has regularly eaten a diet with a high level of fat, X_2=1 indicates that the amount of plaque on the lining of the patient's arteries exceeds some threshold, and X_3=1 indicates that the patient died from a massive stroke.

Suppose that we know a particular patient has eaten a high-fat diet and died of a massive stroke. How might we estimate the posterior probability that the patient has a significant amount of arterial plaque using Gibbs sampling?

2. Gibbs sampling can be used to learn Bayesian networks with missing data. The first step is to represent the learning problem itself as a Bayesian network.

Continuing with the above example, suppose that we wish to compute the quantity Pr(h|d) where h is a hypothesis in the form of the above Bayesian network structure and d is set of assignments of the n variables in the network where each assignment is of the form (X_1=x_1,X_2=x_2,X_3=x_3).

Suppose further that some of the assignments have missing information. This missing information may, for example, correspond to hidden variables. Note that if the assignments were complete and the variables discrete it is straightforward to compute Pr(h|d) using the methods of Cooper and Herskovits, Heckerman, Lauritzen and others. Suppose d is of the form.

```
X_1    X_2    X_3
 1      _      0
 0      0      0
 1      -      0
 1      1      1
```

where the underscores indicate missing information. If determining the amount of arterial plaque requires a detailed autopsy that is rarely done, it may be that much of the patient data will be missing this information. Now express the problem of learning the parameters of the above network as a Bayesian network and then use Gibbs sampling to estimate the parameters.

Back to Tutorial