CSC 446 Notes: Lecture 14

1 Gibbs sampling.

Last time we had the following algorithm for Gibbs sampling:

repeat
  for $k \leftarrow 1 \dots K$
    $x_k \sim P(\mathbf{x}_k | \mathbf{x}_{\neg k}) = \frac{1}{Z} \prod_{m \in M(k)} f_m(\mathbf{x}_m)$

What about the continuous case particularly where sampling is hard? Here we can have a second sampling step:

repeat
  for $k \leftarrow 1 \dots K$
    $\mathbf{x}_k \sim P(\mathbf{x}_k | \mathbf{x}_{\neg k}) = \begin{cases} \frac{1}{Z} \prod_{m \in M(k)} f_m(\mathbf{x}_m) \\ \text{Metropolis–Hastings} \end{cases}$

2 EM with Gibbs Sampling.

E–step Sample each variable.

M–step Use hard (fixed value of the sample) assignments from sampling in:

$$\lambda_k = \frac{\sum_n \mathbb{I}(z^n = k)}{N} = \frac{N_k}{N} \tag{1}$$

In this $\mathbb{I}$ is counting how many points are assigned k and we denote this sum as N_k.

$$\mu_k = \frac{\sum_n \mathbb{I}(z^n = k)\mathbf{x}^n}{\sum_n \mathbb{I}(z^n = k)} = \frac{\sum_n \mathbb{I}(z^n = k)\mathbf{x}^n}{N_k} \tag{2}$$

The computation of Σ is similar.

2.1 Some problems.

What we have above is an approximation to what EM is doing. If we put the computations in expectation it is the same as EM. One problem with EM in general is that if a probability of a cluster hits zero it never comes back. In sampling we can get unlucky and get all zeros and never get that cluster back.

2.2 Some advantages.

With sampling we can apply EM to complicated models. For example, a factor graph with cycles or highly connected components. Sampling can be improved by sampling L times in the E–step.

In practice a short cut is taken and we combine the E and the M steps and take advantage of samples immediately:

for $n \leftarrow 1 \dots N$
sample z^n

$\lambda_k \leftarrow \lambda_k + \frac{1}{N}\mathbb{I}(z^n = k) - \frac{1}{N}\mathbb{I}(z^n_{\text{old}} = k)$
or
$\lambda_k \leftarrow \frac{N_k}{N}$

$\hat{\mu} \leftarrow \hat{\mu} + \mathbb{I}(z^n = k)\mathbf{x}^n - \mathbb{I}(z^n_{\text{old}} = k)\mathbf{x}^n$
$\mu_k \leftarrow \frac{\hat{\mu}}{\lambda_k N}$

We are keeping a running count and updating it as we sample. New observations move data points from their current cluster to a new cluster so we subtract the old observation and add the new one. This technique is widely used and easy to implement.

If we run this long enough it should correspond to the real distribution of hidden variables $P(z^1 \dots z^N|\mathbf{x}^1 \dots \mathbf{x}^N)$ — but what does that mean here? Although we have defined $P(z^n|\mathbf{x}^n; \lambda, \mu, \Sigma)$, the parameters λ, μ, and Σ are changing as sampling progresses. If we run long enough we know that, because the problem is symmetric, $P(z^n = k) = \frac{1}{K}$.

To make sense of this we will make λ, Σ, and μ variables in our model with some distribution $P(\lambda)$. These variables have no parents so we can pick this distribution. We have seen this before and choose to use the Dirichlet distribution so we let $P(\lambda) = \text{Dir}(\alpha)$. Any point is just as likely to be μ so we let $P(\mu) = 1$ and similarly $P(\Sigma) = 1$.

We can take $P(\lambda)$ into account when we sample:

$$z^n \sim \frac{\lambda_k \mathcal{N}(\mathbf{x}; \mu_k \Sigma_k)}{Z}$$

When we have seen data the probability of the next is:

$$P(z^{N+1}|z^1 \dots z^N) = \frac{c(k) + \alpha}{N + K\alpha}$$

Applying this we have:

$$z^n \sim \frac{\frac{N_k+\alpha}{N+K\alpha}\mathcal{N}(\mathbf{x}; \mu_k \Sigma_k)}{Z}$$
$$\lambda_k = \frac{N_k + \alpha}{N + K\alpha}$$

Now λ_k can never go all the way to zero. Now if we run long enough we will converge to the real distribution. We have a legitimate Gibbs sampler (states with just relabeling have equal probability). We are sampling with

$$\lambda_k = \frac{1}{Z}\int P(z^1 \dots z^N|\lambda)P(\lambda)\mathrm{d}\lambda$$

The λ is integrated out, giving what is called a collapsed Gibbs sampler.

A remaining question is: when should we stop? If we plot $P(x)$ as we iterate we should see a general trend upward with some small dips and then the curve levels off. But, it could be that in just a few steps a better state could continue the upward trend.

The real reason for doing this Gibbs sampling is to handle a complicated model. One example of that could be a factor graph of diseases and symptoms due to its high tree width. If we try to use EM directly

we have exponential computation for the expected values of the hidden variables. With Gibbs we avoid that problem.

Ryan Yates 3/12; DG 3/13