CSC 2541: Bayesian Methods for Machine Learning

Radford M. Neal, University of Toronto, 2011

Lecture 9

Models with Multiple Latent Variables

We previously looked at mixture models, in which the distribution of the observable variables was determined by a single discrete latent class variable.

But often a model with several latent variables makes more sense.

Example: Symptoms of a patient are determined by which diseases they have. This can be modeled easily as a mixture only if we unrealistically assume that patients can have only one disease at a time.

If we have 100 possible diseases, and a patient can have up to 10 diseases, we'd need "100 choose 10" $\approx 10^{13}$ mixture components to model the possible combinations of diseases.

Models with Binary Latent Features

For simplicity, I'll look only at models with binary (0/1) latent features.

Denote the observed data for item i by a vector y_i of dimension p, with y_{ij} being the value of the j'th variable.

Denote the latent features for item i by the binary vector z_i of dimension K, with z_{ik} being the value of the k'th feature.

The model will define some distribution for the latent features in an item, and some distribution for the observed variables given the latent features. We'll assume that items are independent, given the model parameters.

For example, if the observed data is real, we might have

$$y_{ij} \mid z_i,\, \omega_j,\, \sigma_j \quad \sim \quad N(z_i^T \omega_j,\, \sigma_j^2)$$

where σ_j^2 and the vector ω_j are model parameters. For binary data, we might have

$$y_{ij} \mid z_i,\, \omega_j \quad \sim \quad \text{Bernoulli}\,(1\,/\,(1+\exp(-z_i^T \omega_j)))$$

In both cases, we might assume that the features of an item are independent given z_i.

A Bayesian Model with a Finite Number of Binary Features

If there are K binary features associated with each item, we could model them as independent, with π_k being the probability that feature k is 1.

Using the conjugate Beta prior for the π_k, and a general form for the distribution of y_i given z_i, we have the following model:

$$\begin{aligned}
\pi_k &\sim \text{Beta}(\alpha/K, \beta) \\
z_{ik} \mid \pi_k &\sim \text{Bernoulli}\,(\pi_k) \\
\phi_k &\sim \ldots \\
\theta &\sim \ldots \\
y_i \mid z_i,\, \phi,\, \theta &\sim F(z_i, \phi, \theta)
\end{aligned}$$

where $F(z_i, \phi, \theta)$ is some distribution that depends on the parameters, ϕ_k, associated with features for which z_{ik} is 1, as well as on common parameters, θ.

MCMC for the Model with a Finite Number of Features

Due to conjugacy, we can integrate away the π_k parameters. We might sometimes be able to integrate away θ and ϕ too, but I'll assume here that we can't.

We can repeatedly perform the following MCMC updates:

1) For $i = 1, \ldots, n$ and $k = 1, \ldots, K$, do a Gibbs Sampling update for z_{ik}. The conditional probabilities for z_{ik} needed for sampling are given by

$$P(z_{ik} \,|\, z_{-ik}, z_i, y_i, \theta, \phi) \quad \propto \quad \begin{cases} \dfrac{n_{-ik} + \alpha/K}{n - 1 + \beta + \alpha/K}\, F(y_i;\, z_i, \phi, \theta) & \text{if } z_{ik} = 1 \\[2ex] \dfrac{n - 1 - n_{-ik} + \beta}{n - 1 + \beta + \alpha/K}\, F(y_i;\, z_i, \phi, \theta) & \text{if } z_{ik} = 0 \end{cases}$$

where $n_{-ik} = \sum\limits_{i' \neq i} z_{i'k}$. (Note that z_{ik} is part of the z_i argument of F above.)

2) For $k = 1, \ldots, K$, update ϕ_k by Gibbs Sampling if possible, or otherwise by a Metropolis or slice sampling update.

3) Update θ by Gibbs Sampling if possible, or otherwise by a Metropolis or slice sampling update.

Here, $F(y_i;\, z_i, \phi, \theta)$ is the density for y_i according to the distribution $F(z_i, \phi, \theta)$.

Letting the Number of Features go to Infinity

What happens if we let $K \rightarrow \infty$ with this model?

First, note that the expected number of z_{ik} that are 1 in any item, i, is

$$E\Big[\sum_{k=1}^{K} z_{ik}\Big] \;=\; \sum_{k=1}^{K} E[z_{ik}] \;=\; \sum_{k=1}^{K} \frac{\alpha/K}{\beta+\alpha/K} \;=\; \frac{\alpha}{\beta+\alpha/K}$$

which goes to α/β as $K \rightarrow \infty$.

So we don't end up with an infinite number of features with value 1 for a single item. Also, for a finite training set, only a finite number of features will have the value 1 in *any* training item.

So it looks like the infinite model might be sensible.

The Indian Buffet Process

To get more insight into the infinite model, consider the prior distribution for z_n given $z_1, \ldots, z_{n-1}$.

Let n_k be $\sum\limits_{i=1}^{n-1} z_{ik}$, and let A be the set of k for which $n_k > 0$.

Given $z_1, \ldots, z_{n-1}$, the probability that z_{nk} is 1 is $\dfrac{n_k + \alpha/K}{n - 1 + \beta + \alpha/K}$.

So when $K \to \infty$, for the finite number of k in A, the probability that z_{nk} is 1 is $n_k \,/\, (n - 1 + \beta)$.

For any of the infinite number of k not in A, the probability that $z_{nk} = 1$ is zero, but the *total* number of such k for which $z_{nk} = 1$ will have a Poisson distribution with mean $\alpha \,/\, (n - 1 + \beta)$.

This process has been pictured as an "Indian Buffet", in which the n'th dinner samples each dish that a previous dinner sampled with probability $n_k \,/\, (n - 1 + \beta)$, and also samples Poisson$\,(\alpha \,/\, (n - 1 + \beta))$ new dishes. (This process is usually presented with β fixed at 1, apparently due to the lure of a spurious simplicity.)

MCMC for the Infinite Feature Model

As $K \to \infty$, the MCMC updates can be rephrased as follows:

1) For $i = 1, \ldots, n$,

   a) Let A be the set of k for which $n_{-ik} > 0$. For all k in A, in random order, do a Gibbs Sampling update for z_{ik}, using these conditional probabilities:

$$P(z_{ik} \,|\, z_{-ik}, z_i, y_i, \theta, \phi) \quad \propto \quad \begin{cases} \dfrac{n_{-ik}}{n-1+\beta}\, F(y_i;\; z_i, \phi, \theta) & \text{if } z_{ik} = 1 \\[2ex] \dfrac{n-1-n_{-ik}+\beta}{n-1+\beta}\, F(y_i;\; z_i, \phi, \theta) & \text{if } z_{ik} = 0 \end{cases}$$

   b) Let B be the set of k for which $n_{-ik} = 0$ but $z_{ik} = 1$. Proposal to replace the set B with a new set of size Poisson$\,(\alpha \,/\, (n-1+\beta))$, with new values of ϕ_k for k in the new set drawn from their prior. The prior cancels in the Metropolis-Hastings acceptance probability, leaving only the likelihoods.

2) For all k for which $z_{ik} = 1$ for some i, in random order, update ϕ_k by Gibbs Sampling if possible, or otherwise by a Metropolis or slice sampling update.

3) Update θ by Gibbs Sampling if possible, or otherwise by a Metropolis or slice sampling update.