Exercises 6: Latent-variable models

Latent class models

Multinomial–Dirichlet conjugacy

The multinomial distribution is the generalization of the binomial to multiple categories. A vector of counts $(n_1, \ldots, n_K)$ has a multinomial distribution with parameter $w = (w_1, \ldots, w_K)$ if the w_k's sum to 1 and if

$$p(n_1, \ldots, n_K \mid w) = \frac{N!}{n_1! \cdots n_K!} \prod_{k=1}^{K} w_k^{n_k},$$

where $N = \sum_{k=1}^{K} n_k$. The normalizing constant can be derived from the same kind of combinatorial argument that applies to the binomial distribution.

1. Show that the Dirichlet distribution, written $w \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ and with p.d.f.

   $$p(w \mid \alpha) = \frac{\Gamma(\sum \alpha_k)}{\sum \Gamma(\alpha_k)} \prod_{k=1}^{K} w_k^{\alpha_k - 1},$$

   is the conjugate prior for the multinomial distribution. (The support of the Dirichlet distribution is the K-dimensional simplex.)

2. Let $x_k \sim \mathrm{Ga}(\alpha_k, 1)$ be independent Gamma draws for $k = 1, \ldots, K$. Show that the random vector with elements $w_k = x_k/X$, where $X = \sum x_k$, has a Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_K)$. (Recall that you showed a very similar thing for the beta distribution; the Dirichlet is easily understood as the multiple-category generalization of the beta.)

A simple mixture model

Imagine a high school where every sophomore takes a standardized math test. This school has three kinds of sophomore math classes:

remedial classes, with a cumulative enrollment of about 100 students. Among students in the remedial classes, the test scores are normally distributed with mean 55 and standard deviation 15.

average classes, cumulative enrollment 400 students. Among these students, the test scores are normally distributed with mean 70 and standard deviation 10.

honors classes, cumulative enrollment 150. Among these students, the
test scores are normally distributed with mean 85 and standard
deviation 5.

The following problems all concern this stylized scenario.

1. Simulate a few different data sets according to this scenario. How
   would you describe the school-wide distribution of test scores?
   (Skewed, multi-modal, etc.)

2. Let γ_i denote a latent indicator variable for student i, that takes
   the value k if student i falls in class k. Let $w = (w_1, w_2, w_3)$ denote
   the prior probabilities of the three classes. Write functions to draw
   from the conditional posterior distributions for: (1) γ_i, given w and
   the information about the distribution of students within classes;
   and (2) w, given all the γ_i's.

3. Suppose you allocate every student into one of three classes by
   drawing from the conditional distribution of γ_i. Now pretend
   you didn't know the true mean or variance of the scores within
   each class, and that you merely had (sensibly chosen) priors for
   them. What would the posterior distribution be for the mean and
   variance of students within bucket k?[1]

Propose a general Gibbs sampler for the following problem: the marginal
density of some outcome Y is a K-component mixture of normals, but
you don't know the probabilities, means, or variances of each of the
components. Try running your Gibbs sampler on one of the simulated
test-score data sets, and see how well it does at recovering the true den-
sity. (Notice how you get error bars for free from your Gibbs sampler.)

Then, use your Gibbs sampler to fit a mixture of normals to the data
in "galaxies.txt," which shows the recessional velocities for 82 different
galaxies (i.e. how fast those galaxies are moving away from us) from
6 different well-separated conic sections of space. Our current under-
standing of galaxy formation would predict just such a mixture distri-
bution, with each cluster corresponding to a "clump" of galaxies, and
each galaxy within a cluster moving at a similar speed.

Latent feature models

[1] If you want to superimpose a plot
of the three densities on top of the
histogram of test scores, use the
"prob=TRUE" flag to the histogram
function. Then you can add lines to the
plot.