Advanced MCMC, Computer Classes

Mateusz K. Łącki

9 December 2016

You will now look on the problem of finding the parameters of normal mixtures in more detail.

Task 1 Make a function that will draw the sample

$$X_1, \ldots, X_n \sim f(x|\theta) = \sum_k w_k g_{m_k}(x),$$

where g_{m_k} stands for the density of a normal distribution with mean m_k and variance equal to 1. We will sometimes call these different distributions clusters. Say there are $K = 5$ clusters spaced so that the modes are equally spaced and the distance between two consecutive modes equals $d = 1.5$, somewhere between 0 and 20. Draw the weight heights from Dirichlet distribution and store them. Draw $N = 200$ sample points.

Task 2 Consider a data-augmented model that reveals additional information: the tags of data points to the different clusters. Each tag corresponds to a binary vector $Y_n = (Y_{nk})$, where $Y_{nk} = \mathbb{1}(X_n \text{was drawn from k-th mode})$. With this additional information, the likelihood of the sample $(X_1, Y_1), \ldots, (X_N, Y_N)$ can be written as:

$$L \propto \prod_n \prod_k \Big[w_k g_{m_k}(X_n) \Big]^{Y_{nk}}.$$

Of course, we have no idea of what these clusters assignments are. But we can draw them. To this end, please encode the Metropolis within Gibbs algorithm (again: both random scan and the deterministic full scan - reuse the previous code if possible). The algorithm should implement the drawing of weights and modes given the data and the cluster assignments, $w, m|X, Y$, as well as the drawing of cluster assignments given the data, the weights and the modes, $Y|w, m, X$. Find the proper formulas. If needed, propose a nice proposal kernel for the Metropolis step that might be necessary to draw one of the parameters (which one?).

As a prior for the parameters w consider the product of the Dirichlet distribution with pseudo-counts set initially to 1. As in the previous project, we assume that modes can be enumerated in an increasing order, $m_1 < m_2 < \cdots < m_K$. Moreover, $0 = m_{\text{min}} < m_1$ and $m_K < m_{\text{max}} = 20$. Consider normalized

modes: $n_i = \frac{m_i - m_{\min}}{m_{\max} - m_{\min}}$. They too are to be drawn from the Dirichlet distribution with pseudo-counts set to 1.

The Gibbs sampler should therefore in principle produce tuples (w, m, Y). However, you can skip the collection of assignments Y, as we are mainly interested in the marginal distribution of this extended vector with respect to weights and modes.

Task 3 Use any drawing module to plot the histogram of the data and the density of the mixture model. For example, you might consider trying out module SEABORN.

Please send the answers by the end of this calendar year.

On the next classes we will try to implement a philosophically alternative method to the one above to estimate the true values of the mixture models. This way we will explore the non-bayesian world and pay homage to it.

Best wishes, Matteo