Mixure model estimation process

Evan Miller

Thursday, June 16, 2016

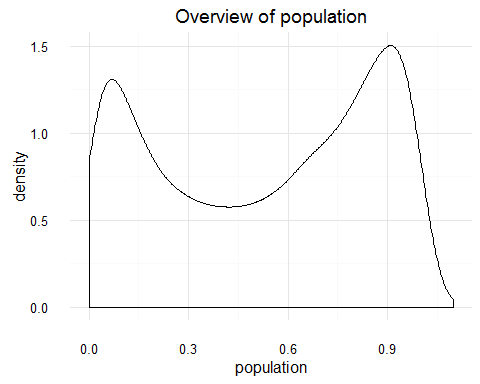
This R markdown document will highlight the process used to estimate a mixture model from the underlying population data. The data has:

* 8000 members;
* data generated from 2 separate beta distributions

I need to:

* Find the maximum likelihood estimate (MLE) for the parameters for each beta distribution ( & )
* Find the percentage of the data coming from each component beta distribution ().

**The generated data looks as follows**:



After looking at the graph there are two peaks. One at 0.1 and one at 0.95. However this doesn't give me much help as the beta distribtuion is characterised by it's alpha and beta parameters, not the mean and standard deviation.

**The approach I have taken is as follows:**

1. Given initial values for the beta distributions calculate the probability that a given observation would arise from either of the two beta distributions
2. Use these probabilities to calculate the class responsibilities
3. Draw randomly from a bernoulli distribution a class that is based off the observation's class responsibilities
4. Use the fitdistr function to fit a beta distribution for each class
5. Repeat this for a given number of iterations (minimum 10) and take the median of these parameters
6. Calculate the mixing proportion pi
7. Check if convergence has been achieved by differencing the initial and calculated parameters
8. If convergence hasn't been achieved repeat the process with the calculated parameters as inputs

**Issues I'm coming up against:**

* I cannot be sure whether I'm seeing a local or global maximum when running this model. To get around this multiple runs will be needed of this recursive approach. I am also experimenting with using k-means to generate initial estimates for the function.
* The code itself takes a long time to run. I am trying to speed up the code by using vectorised functions instead of loops but the improvements have only been marginal.
* I am new at working with mixture models. Because of this I am cross validating my approach against a small mixed distribution derived from two normal subpopulations. Once I am fine with the results of this I will derive my own beta distributions to check against (on a smaller scale than the sample data while I work on speeding up the code).