Bayesian consensus clustering

Stephen Coleman

22/10/2019

Intro

Dirichlet-Multinomial Allocation mixture model

This model (Lock and Dunson 2013) is based upon a finite approximation of the Dirichlet process known as the Dirichlet-Multinomial Allocation mixture model (Green and Richardson 2001), as is Clusternomics (Gabasova, Reid, and Wernisch 2017) and Multiple Dataset Integration (Kirk et al. 2012).

Consider a dataset of N samples, $X = (x_1, \ldots, x_N)$ where each sample is itself a p-vector of measurements fro some $p \in \mathbb{N}$. We are interested in uncovering structure in the data. To do this we associate each sample, X_i, with a *cluster*. In this way we move from a p-dimensional space to a 1-dimensional space. We want to understand if there are sub-populations in our sample that are responsible for heterogeneity. By partitioning the data into clusters we hope to have some insight into this underlying structure and improve our understanding of, and, as a result of this, our ability to interpret the data. There are a myriad of ways to uncover such partitions. Some methods have more obvious disadvantages than others, a common problem being that the number of clusters allowed in the follow result is somewhat arbitrary. Another common problem is that some of the methods are not model based which gives a somewhat ad-hoc nature. A method that does not suffer from these problems is the Dirichlet process. A tractable approximation of this is the Dirichlet-Multinomial Allocation mixture model which is a common choice in the Bayesian-model based clustering. In this case we model each sub-population by an individual distribution. We also allow the number of clusters present to be inferred from the data, thus avoiding a heuristic or arbitrary means of selecting the number of clusters allowed.

In this model we use a mixture of K components to model the sub-populations. The components are all modelled by a distribution with a density function $f(\cdot)$. The kth component has associated parameters, θ_k, based on the samples allocated to this component and is thus defined by density $f(\theta_k)$. The proportion of samples within the components define the associated mixture weights, $\pi = (\pi_1, \ldots, \pi_K)$. Each sample is allocated to a specific component; this component memberhsip is represented by the latent variable $c = (c_1, \ldots, c_n)$, where $c_i \in \{1, \ldots, K\} \forall i \in \{1, \ldots, N\}$. Therefore we have allocation probability:

$$p(x_i|c_i = k) = \pi_k f(x_i \theta_k),$$

and the full model density, which is a weighted sum of the mixture densities:

$$p(x_i) = \sum_{k=1}^{K} \pi_k f(x_1|\theta_k)$$

Note that while there are component-specific parameters, the density function used, $f(\cdot)$, is common across all components.

In the Dirichlet-Multinomial Allocation mixture model, the mixture weights are given a Dirichlet prior, normally with a symmetric concentration parameter. The allocation variable is then sampled from a Cateogrical distrbution defined by the component weights. The component parameters are then updated based upon the allocation of samples and the full model is then the weighted sum described above. The full

hierarhcical model is described below:

$$\begin{aligned} \pi &\sim \text{Dirichlet}\left(\frac{\alpha_0}{K}, \ldots, \frac{\alpha_0}{K}\right), \\ c_i &\sim \text{Categorical}(\pi), \\ \theta_k &\sim h(\cdot), \\ x_i | c_i = k &\sim f(x_i | \theta_k). \end{aligned}$$

Here $h(\cdot)$ is some distribution, often with additional hyperparameters.

References

Gabasova, Evelina, John Reid, and Lorenz Wernisch. 2017. "Clusternomics: Integrative Context-Dependent Clustering for Heterogeneous Datasets." *PLoS Computational Biology* 13 (10). Public Library of Science: e1005781.

Green, Peter J, and Sylvia Richardson. 2001. "Modelling Heterogeneity with and Without the Dirichlet Process." *Scandinavian Journal of Statistics* 28 (2). Wiley Online Library: 355–75.

Kirk, Paul, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. 2012. "Bayesian Correlated Clustering to Integrate Multiple Datasets." *Bioinformatics* 28 (24). Oxford University Press: 3290–7.

Lock, Eric F, and David B Dunson. 2013. "Bayesian Consensus Clustering." *Bioinformatics* 29 (20). Oxford University Press: 2610–6.