Infinite Multiway Mixture with Factorized Latent Parameters

Işık Barış Fidaner
Computer Engineering
Boğaziçi University
Bebek, İstanbul
fidaner@gmail.com

Ali Taylan Cemgil
Computer Engineering
Boğaziçi University
Bebek, İstanbul
taylan.cemgil@boun.edu.tr

Abstract

In this paper, we develop an infinite multiway mixture model, whose parameters are represented as a tensor factorization. We define a D-way Poisson mixture, where a large observed tensor X is generated by the mixture proportions π_d and a smaller latent tensor Θ, which is represented as a factorization of M latent factors Θ_m of varying dimensionalities. We first derive an EM algorithm for the finite mixture. Then, we formulate an infinite multiway mixture, and propose an MCMC method to sample the assignments.

1 Introduction

Clustering has been a primary problem in Bayesian nonparametrics that led to the development of a literature on DP mixtures. Multiway clustering is a problem that has recently gained attention. In [1], the multiway clustering problem is formulated with reference to earlier work on using hypergraphs to approach VLSI and PCB clustering placement problem. As elaborated in [2], a hypergraph is a general representation that contains hyperedges of any size that relate any combination of entities. In [3], a general multiway framework is presented to handle various kinds of hyperedges. Our model assigns D-tuples of objects to D-tuples of clusters, thus only involves D-way hyperedges that relate D objects of different types.

We use D-way tensors to represent the variables. In [4], a probabilistic tensor factorization framework is presented for multiway analysis. We use a similar framework to represent the latent D-way tensor of component parameters in the multiway mixture. In [5], an MCMC method was proposed for nonparametric biclustering problem. Biclustering is a special case of D-way clustering for $D = 2$, thus its solution can be applied to the general multiway problem.

In the following sections, we first formulate the multiway clustering problem as a finite mixture, and present a variational inference method. Then, we present an MCMC inference method to use with the infinite mixture model.

2 The D-way Poisson mixture model

In our problem, we have D types of objects, and N_d objects from each object type $d \in \{1, \ldots, D\}$. Each of the observations is given for a D-tuple of these objects, together making up the D-way tensor X with sizes $N_1, \ldots, N_D$ in its D dimensions. We model X as a Poisson mixture of latent parameters in a smaller D-way tensor Θ with sizes $K_1, \ldots, K_D$. Here, K_d is the number of clusters for a given dimension, and in general it is significantly smaller than N_d.

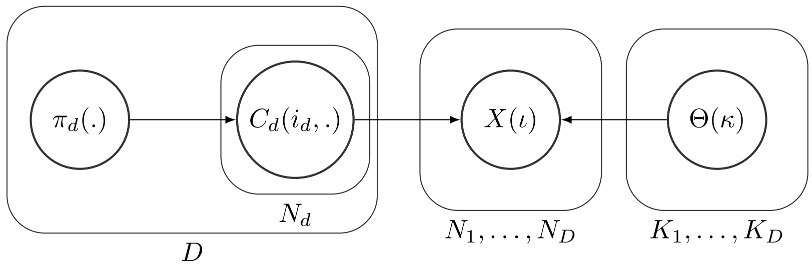

Figure 1: The D-way mixture model

A D-way tensor is indexed by an 'index set' of D indices. Each observation in the tensor X is denoted $X(\iota)$ where ι is the index set $\{i_1,\ldots,i_D\}$ and $i_d \in \{1,\ldots,N_d\}$. Similarly, the tensor Θ is indexed by κ that is the index set $\{k_1,\ldots,k_D\}$ where $k_d \in \{1,\ldots,K_d\}$.

2.1 Layer assignments

In D-way clustering, for each object type d, N_d objects of this type are to be assigned to the corresponding K_d clusters. For $D=1$, single observations in X are assigned to single parameters in Θ. For $D=2$, rows of observations in X are assigned to rows of parameters in Θ, and columns to columns. When $D=3$, matrices in three different orientations from X are assigned to matrices of corresponding orientations in Θ. For the general case, (D-1)-way tensors in X are assigned to (D-1)-way tensors in Θ. We call such a (D-1)-way tensor a 'layer', and indicate it by a dimension d, and a value for its index j_d. The layer's orientation is given by d, and its placement inside the tensor by j_d.

We call C_d an indicator variable. For each orientation d, that $C_d(i_d,k_d)=1$ indicates that the layer at i_d of X is assigned to the layer at k_d of Θ. A single observation $X(\iota)$ is thus assigned by the indicators $C_1,\ldots,C_D$ to the layers at the indices $k_1,\ldots,k_D$ of Θ. These indices form the set κ, and as a result, the observation is assigned to the latent parameter at $\Theta(\kappa)$.

In the mixture model, we assume that each observation $X(\iota)$ is Poisson distributed with the intensity as the latent parameter $\Theta(\kappa)$ to which it is assigned.

$$X(\iota) \mid \Theta, C \;\sim\; \prod_{\kappa} \mathcal{PO}(\Theta(\kappa))^{\prod_{d=1}^{D} C_d(i_d,k_d)}$$

We model each of the vectors $C_d(i_d,.)$ by a discrete distribution $\pi_d(.)$ of size K_d. This vector is in turn modeled by a symmetric Dirichlet prior with concentration α_d. The full model is shown in Figure 1.

$$\begin{aligned} C_d(i_d,.) \mid \pi_d \;&\sim Discrete(\pi_d(.)) \\ \pi_d(.) \;&\sim Dir(\frac{\alpha_d}{K_d},\ldots,\frac{\alpha_d}{K_d}) \end{aligned}$$

2.2 Representing the latent tensor

Up to this point, we have described a general D-way mixture model. What makes our model specific is the representation of the latent tensor Θ. We assume that Θ is a function of other M latent factors Θ_m of different dimensionalities.

$$\Theta(\kappa) \;=\; \sum_{\beta} \prod_{m=1}^{M} \Theta_m(\gamma_m)$$

The factorization is summed over a set of indices β to get the latent tensor Θ indexed by κ. Here, β denotes an 'additional' set of indices $\{k_{D+1},\ldots,k_{D+\Delta}\}$ that extends the 'original' set denoted by κ. The union of these two gives the full set of indices $\kappa \dot{\cup} \beta = \{k_1,\ldots,k_{D+\Delta}\}$. Each of the M factors is indexed by γ_m, such that $\cup_{m=1}^{M}\gamma_m = \kappa\dot{\cup}\beta$.

We consider the factors Θ_m as the actual hidden parameters, and put on them a Gamma prior, which is conjugate with the Poisson distribution.

$$\Theta_m(\gamma_m) \sim \mathcal{G}(A, \frac{B}{A})$$

2.3 Partitioning a factor's indices

The index set γ_m of a factor is partitioned into three disjoint sets in two consecutive steps as follows:

$$\gamma_m = \eta_m \dot{\cup} \beta_m = \eta_m \dot{\cup} \sigma_m \dot{\cup} \lambda_m$$

In the first step, it is partitioned into its 'original' indices $\eta_m = \gamma_m \cap \kappa$ and 'additional' indices $\beta_m = \gamma_m \cap \beta$. We then introduce $\beta_{-m} = \cup_{m' \neq m} \beta_m$ to denote the additional indices that belong to any factor other than m. In the second step, β_m is further partitioned into the indices shared with other factors $\sigma_m = \beta_m \cap \beta_{-m}$ and the indices that are not shared $\lambda_m = \beta_m \backslash \beta_{-m}$.

2.4 Determining the parameters of the model

For a given D, a finite multiway mixture model is selected by determining the following:

1. The sizes $\{K_1, \ldots, K_D\}$ of the latent tensor Θ (the number of clusters for all object types).
2. The concentration parameters $\alpha_1, \ldots, \alpha_D$ for each of the object types.
3. The number of factors M, and the index sets $\gamma_1, \ldots, \gamma_M$ for each of these factors. When these are given, we can also determine the set of additional indices $\beta = \cup_{m=1}^{M} \gamma_m \backslash \kappa$.
4. The prior parameters A and B for the factors Θ_m.

Various factorizations of Θ lead to different models. To mention two basic examples: When $M = 1$ and $\gamma_1 = \kappa$, the latent tensor $\Theta(\kappa)$ is modelled directly. When $M = D$ and $\gamma_d = \{k_d\}$, the tensor Θ is the product of D vectors $\Theta_d(k_d)$.

3 Variational inference for the finite mixture

We develop an Expectation-Maximization algorithm that involves the following steps

1. Calculate the expectation of $p(C \mid X, \Theta, \pi)$.
2. For each $d \in \{1, \ldots, D\}$, calculate the π_d that maximizes $p(X, C, \Theta, \pi)$.
3. For each $m \in \{1, \ldots, M\}$, calculate the Θ_m that maximizes $p(X, C, \Theta, \pi)$.

In the expectation step (1), we use the posterior of the layer assignments.

$$p(C \mid X, \Theta, \pi) \propto \{\prod_{\iota} \prod_{\kappa} \mathcal{PO}(X(\iota) \mid \sum_{\beta} \prod_{m=1}^{M} \Theta_m(\gamma_m))^{\prod_d C_d(i_d, k_d)}\} \{\prod_{d} \prod_{i_d} \prod_{k_d} \pi_d(k_d)^{C_d(i_d, k_d)}\}$$

In the maximization step (2), we update $\pi_d(k_d)$ by the equation:

$$\pi_d^*(k_d) = \frac{\frac{\alpha_d}{K_d} - 1 + \sum_{i_d} \mathbb{E}[C_d(i_d, k_d)]}{\alpha_d - K_d + N_d}$$

In the next step (3), we update $\Theta_m(\gamma_m)$ by the following formula:

$$\Theta_m^*(\gamma_m) = \frac{A - 1 + \sum_{\iota} \{X(\iota) \frac{\prod_{d:k_d \in \lambda_m} K_d}{\sum_{\lambda'_m \neq \lambda_m} \Theta_m(\eta_m \cup \sigma_m \cup \lambda'_m)}\} \mathbb{E}[\prod_d C_d(i_d, k_d)]}{\frac{B}{A} + \sum_{\iota} \{\sum_{\beta} \prod_{m' \neq m} \Theta_{m'}(\gamma_{m'})\} \mathbb{E}[\prod_d C_d(i_d, k_d)]}$$

For any factor $\Theta_m(\gamma_m)$ with no additional indices (for each m where $\beta_m = \emptyset$), the fraction in the numerator of the formula reduces to 1. When there is no factorization ($M = 1$), the coefficient in the summation in the denominator also reduces to 1.

4 The infinite mixture and MCMC inference

By taking a finite D-way mixture, and bringing $K_d \to \infty$ for some or all of the object types, we can obtain an infinite multiway mixture. We are developing an MCMC method for inferring the layer assignments in such a nonparametric multiway mixture model.

A variety of MCMC methods for DPM are presented in [6] including Gibbs sampling, Metropolis-Hastings updates and auxiliary parameters to handle both conjugate and non-conjugate priors. In [5], such a method is developed for a nonparametric biclustering model, which can be obtained from our D-way model for $D = 2$. The method proposed is based on a property which we will express in our terms as follows.

When C_d are given, for each κ, there is a set or a 'block' of observations that are assigned to it:

$$\{X(\iota) : \kappa\} \;=\; \{X(\iota) : \prod_d C_d(i_d, k_d) = 1\}$$

Conditional to Θ, these blocks of observations are independent and thereby their likelihoods:

$$p(X \mid \Theta, C) \;\sim\; \prod_\kappa p(\{X(\iota) : \kappa\} \mid \Theta(\kappa))$$

In case of a conjugate prior, we can also integrate out Θ to get the following:

$$p(X \mid C) \;\sim\; \prod_\kappa \int p(\{X(\iota) : \kappa\} \mid \Theta(\kappa))\, p(\Theta(\kappa))\, d\Theta(\kappa) \;=\; \prod_\kappa p(\{X(\iota) : \kappa\})$$

Using this property, an MCMC algorithm similar to [5] can be developed for the infinite D-way mixture.

References

[1] Shashua, A., Zass, R. & Hazan, T. (2006) Multi-way clustering using super-symmetric non-negative tensor factorization. In Proc. of the European Conference on Computer Vision (ECCV)

[2] Zhou, D., Huang, J. & Schölkopf, B. (2007) Learning with Hypergraphs: Clustering, and Classification, Embedding. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems 19*, pp. 1601-1608. MIT Press.

[3] Banerjee, A., Basu, S. & Merugu, S. (2007) Multi-Way Clustering on Relation Graphs, In Proc. SIAM Conf. Data Mining.

[4] Yilmaz, K. & Cemgil, A. T. (2010) Probabilistic Latent Tensor Factorisation, In Proc. of International Conference on Latent Variable analysis and Signal Separation, 6365, 346-353

[5] Meeds, E. & Roweis S. (2007) Nonparametric Bayesian Biclustering. UTML-TR-2007-001, Technical Report, University of Toronto.

[6] Neal, R. M. (2000) Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9, 249–265