Bayesian Non-Parametrics and Dirichlet Process Clustering Techniques

Radhika Anand, Dipesh Gautam, Princeton Lee
Department of Statistical Sciences
Duke University

1 Introduction

Standard clustering algorithms like K-Means and Gaussian Mixture Models require us to know the total number of clusters a-priori. This has two problems:

- We do not always know the number of clusters a-priori
- Most real-world problems don't have a fixed number of clusters

This brings us to a class of algorithms called Non-Parametric Bayes, which assumes that true dimensionality is unbounded and allows parameters to grow with growing data. This use of countably infinite mixtures (which can easily be treated in a Bayesian framework) bypasses the need to determine the exact number of components/clusters a-priori. Also, a countably infinite mixture is more realistic than a mixture with small number of components.

Commonly studied non-parametric Bayes techniques include:

1. Chinese Restaurant Process (CRP)
2. Polya Urn Model
3. Dirichlet Process
4. Indian Buffet Process (IBP) [for feature allocation]

We worked on Indian Buffet Process (IBP) as part of our class, Statistical Computation, last semester where we used IBP to detect base images used to construct several images. Thus, we studied the application of non-parametric Bayes in feature allocation. We however could not dive into its applications in clustering, which motivated us to study them as part of our Machine Learning project this semester.

In this report, we begin with a discussion of CRP and Polya Urn Models and proceed to the study of Dirichlet Process Mixture Models [1] and DP-Means Algorithm [2] for clustering. We then test the later two on McDonald's dataset of item nutrition profiles to obtain food clusters.

2 Non-Parametric Clustering Models

This section describes the working of Chinese Restaurant Process and Polya Urn Models in detail.

2.1 Chinese Restaurant Process

Consider data points as customers who visit a restaurant and clusters as tables in that restaurant. CRP is based on the idea that there are infinite number of tables in the restaurant and infinite number of seats at each table. The first customer sits at a table. The (n+1)-st customer then sits at a new table with probability $\frac{\alpha}{(n+\alpha)}$ and on an occupied table with probability $\frac{n_k}{(n+\alpha)}$, where n_k is the number of customers occupying table k.

Important points:

1. The more the customers at a table, the more probable it is that the next customer will join that table. This is the rich-get-richer property.
2. Probability of a new table (cluster) is proportional to α. Thus, we consider α as a dispersion parameter which determines the total number of clusters. The higher the α, the higher is the number of clusters in a given set of data.

Finally, CRP specifies a distribution over partitions/table assignments but does not assign parameters to tables.

Fig. 1 shows how the total number of detected clusters increases with increasing α value.

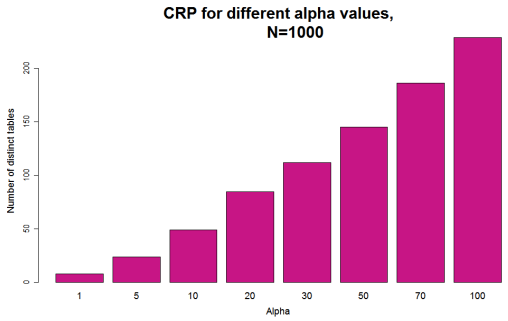


Figure 1: Number of distinct tables for different α values

2.2 Polya Urn Model

Consider data points as balls in an urn and their colors as the clusters they belong to. We now pick a ball from the urn, note its color and drop that ball and another ball of the same color back into the urn.

Thus, it is similar to CRP in that it follows the rich-get-richer property and has α as the dispersion parameter.

Further, it not only specifies a distribution over partitions but also assigns parameters to them. It is like a distribution over distributions as can be visualized from the various runs of the same Polya Urn Model in the plots below. This brings us to the Dirichlet Process.

Fig. 2 shows the density of colors in the urn for several runs. We use standard normal as the base color distribution. Thus, as α increases i.e. we sample more and more new colors from the base distribution, we see that the plots tend towards standard normal i.e. they approach the base distribution.

3 Methodology

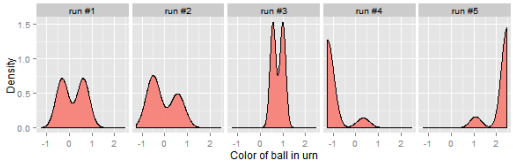
We use the above models to develop clustering techniques like Dirichlet Process Mixture Models and DP-Means and apply them to a real dataset.

3.1 Dirichlet Process Mixture Models (DPMM)

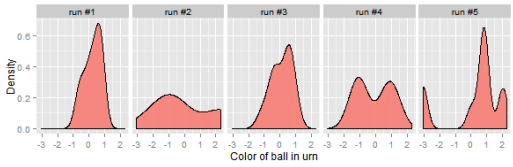
Suppose we have multivariate, real-valued data $y_1, ..., y_n$, which is exchangeable, a DPMM is defined as follows:

$$\begin{aligned} y_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G &\sim D(G_0, \alpha) \end{aligned}$$

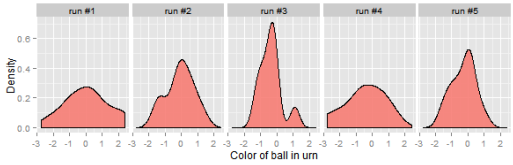
Figure 2: Density of colors for different α values with standard normal as base distribution



(a) $\alpha = 1$



(b) $\alpha = 5$



(c) $\alpha = 50$

i.e., we model the distribution from which y_i's are drawn as a mixture of distributions of the form $F(\theta)$, which is normal for a Gaussian DPMM, with G as the mixing distribution over parameters θ. Further, the prior over G is a Dirichlet Process with base distribution G_0 and dispersion parameter α.

3.2 DP-means

DP-Means is a hard clustering algorithm, obtained by the asymptotic behavior of a DPMM, which behaves similar to K-Means with the exception that a new cluster is created whenever a point is more than λ away from every existing cluster centroid. The algorithm is as follows:

Input: $x_1, ..., x_n$: input data; λ: cluster penalty parameter
Output: Clustering $\ell_1, ..., \ell_k$ and number of clusters k

1. Initialize $k = 1$, $\ell_1 = \{x_1, ..., x_n\}$ and μ_1 the global mean.
2. Initialize cluster indicators $z_i = 1$ for all $i = 1, ..., n$.
3. Repeat until convergence
   - For each point x_i
     - Compute $d_{ic} = \|x_i - \mu_c\|^2$ for $c = 1, ..., k$
     - If $min_c d_{ic} > \lambda$, set $k = k + 1$, $z_i = k$ and $\mu_k = x_i$
     - Otherwise, set $z_i =$argmin$_c d_{ic}$.
   - Generate clusters $\ell_1, ..., \ell_k$ based on $z_1, ..., z_k : \ell_j = \{x_i | z_i = j\}$
   - For each cluster ℓ_j, compute $\mu_j = \frac{1}{|\ell_j|} \sum_{x \epsilon \ell_j} x$

4 Data

We test the above techniques on McDonald's item-nutrition dataset to cluster food items into groups. The dataset consists of 325 rows where each row is a McDonald's food item and 14 columns which

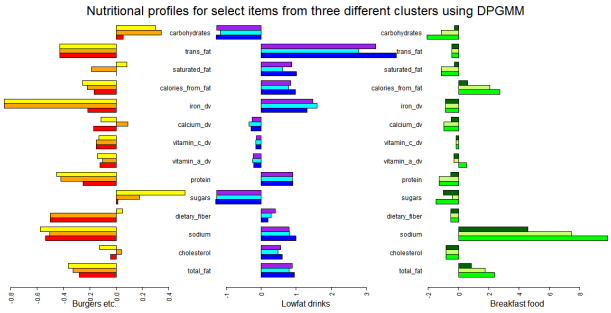


Figure 3: DP-GMM

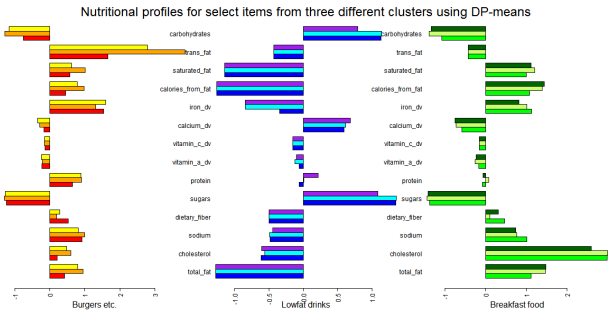


Figure 4: DP-Means

make up their nutritional profile, namely, total fat, cholesterol, sodium, dietary fiber, sugars, protein, vitamin A, vitamin C, calcium, iron, calories from fat, saturated fat, trans fat and carbohydrates.

5 Results

We applied DPGMM and DP-Means to McDonalds item-nutrition dataset to cluster food items into groups based on the values of their nutritional parameters. We obtained around 11 clusters with separate clusters for burgers, lowfat drinks, breakfast items, desserts, salads, desserty drinks etc. Fig 3 and Fig. 4 show the nutrition profiles for select items from three different clusters for both DPGMM and DP-Means respectively.

6 Conclusions

We studied Non-Parametric Bayes techniques for clustering to solve the problem of specifying the number of clusters a-priori. We studied CRP, Polya Urn Models and Dirichlet Process and used them to derive DPGMMs and DP-Means. We tested them on McDonalds dataset to obtain clusters as above and observed that cluster assignments using DPGMM were more accurate than those obtained using DP-Means, which is as we would expect since DP-Means is just a hard clustering algorithm derived using asymptotics on DPGMM.

As mentioned in the introduction, we also explored the field of feature allocation using the Indian Buffet Process (IBP) last semester. We used IBP to detect latent features in an image dataset. The details of this can be found at `https://github.com/ra125/STA663-FinalProject`.

References

[1] Neal, R. Markov Chain Sampling methods for Dirichlet Process Mixture Models, 1998

[2] Kulis B. & Jordan M. Revisiting k-means: New Algorithms via Bayesian Nonparametrics, 2012