Bayesian Non-Parametrics and Dirichlet Process Clustering Techniques

Radhika Anand, Dipesh Gautam, Princeton Lee

Duke Statistical Science

Motivation

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:
(a) We do not always know the number of clusters a-priori
(b) 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 the parameters to grow with growing data. It doesn't require a prior specification of the number of latent clusters, which it instead infers from the data itself.

Commonly studied non-parametric clustering techniques include:
1) Chinese Restaurant Process
2) Polya Urn Model
3) Dirichlet Process
4) Indian Buffet Process (for feature allocation)

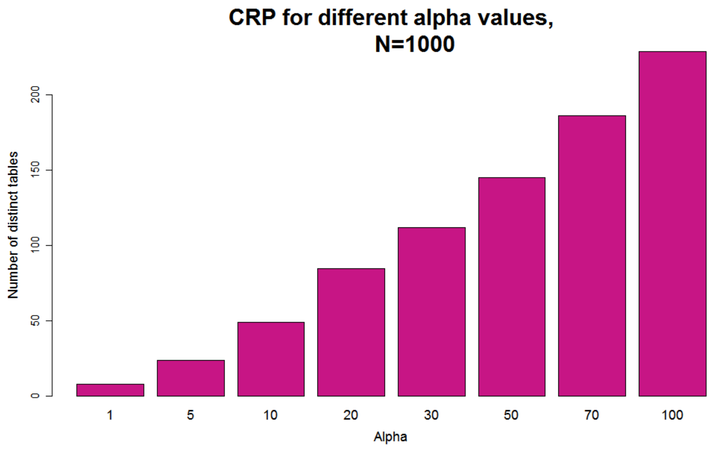
Chinese Restaurant Process (CRP)

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 $\alpha/(n+\alpha)$ and on an occupied table with probability $n_k/(n+\alpha)$, where n_k is the number of customers occupying table k.

Points to note:
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 (which can be done by the Polya Urn Model as in the next section).

The graph below shows how the total number of detected clusters increases with increasing α value.



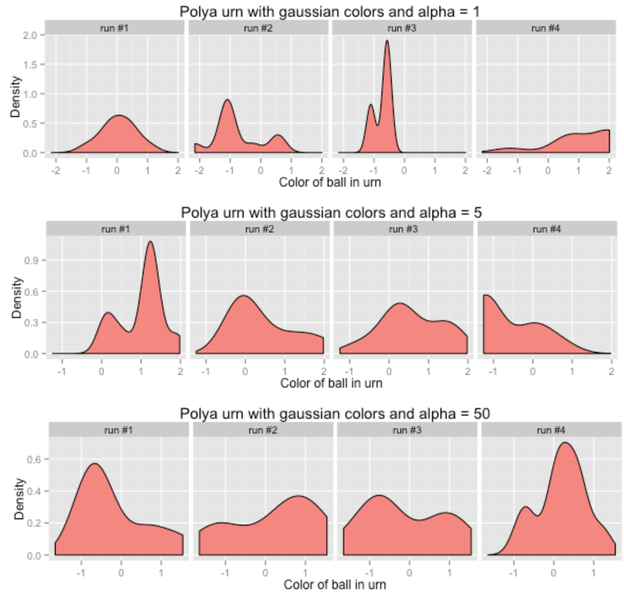
Polya Urn Model

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**.

The plots below show 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.



Dirichlet Process Mixture Model (DPMM)

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

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

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.

References

1) Neal, R. *Markov Chain Sampling methods for Dirichlet Process Mixture Models*, 1998
2) Kulis B. and Jordan M. *Revisiting k-means: New Algorithms via Bayesian Nonparametrics*, 2012

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:

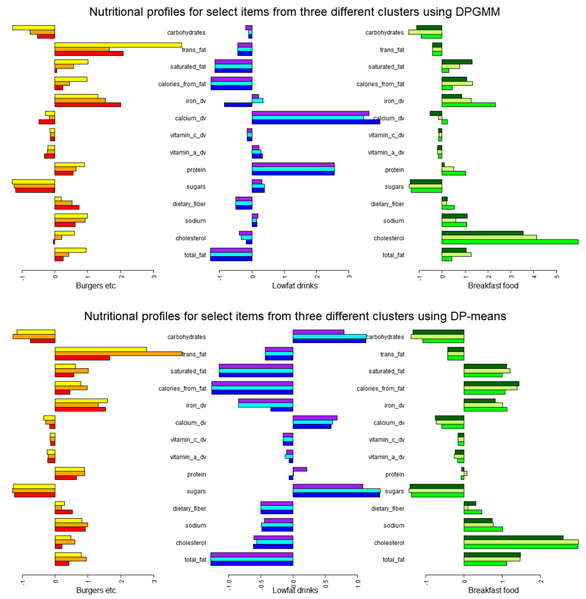
Algorithm 1 DP-means

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

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

Application to McDonald's Dataset

We applied DPGMM and DP-Means to McDonald's item-nutrition dataset to cluster food items into groups. We got around 11 clusters with separate clusters for burgers, lowfat drinks, breakfast items, desserts etc. The plots below show the nutrition profiles for select items from three different clusters for both DPGMM and DP-Means.



Conclusion

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 these on McDonald's dataset to obtain clusters as above and observed that the 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.

Next, we also explored the field of feature allocation using the Indian Buffet Process (IBP). We used IBP to detect latent features in an image dataset (refer to the report for results).