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 Discussion of various clustering models

This section describes the working of CRP 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.

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.

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

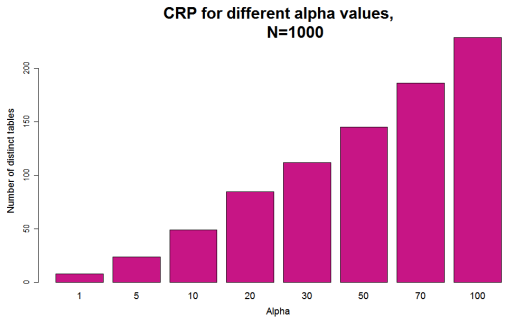


Figure 1: Number of distinct tables for different α values

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

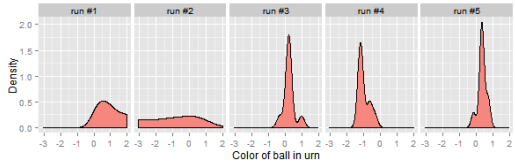
We use the above models to develop clustering techniques like Dirichlet Process Mixture Models and DP-Means which are apply them to a real dataset.

3.1 Dirichlet Process Mixture Models (DPMM)

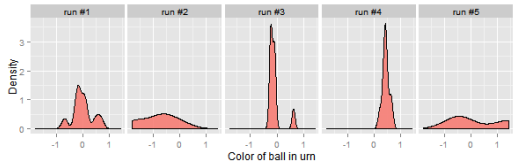
Suppose we have multivariate, real-valued data $y_1, ..., y_n$, which is exchangeable, a DPMM is defined as follows:

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

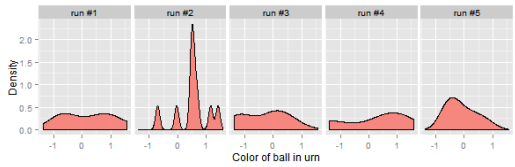
Figure 2: Density of colors for different α values with standard normal as base distribution



(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 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. Init. $k = 1$, $\ell_1 = \{x_1, ..., x_n\}$ and μ_1 the global mean.
2. Init. 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 apply the above techniques to 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

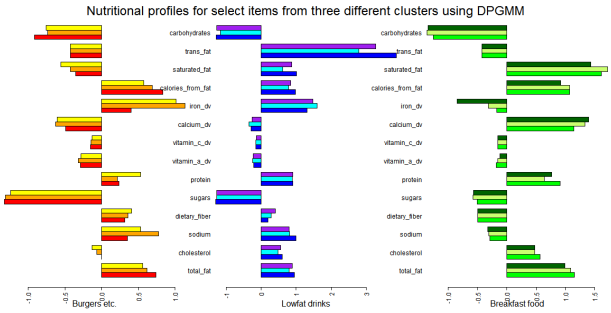


Figure 3: DP-GMM

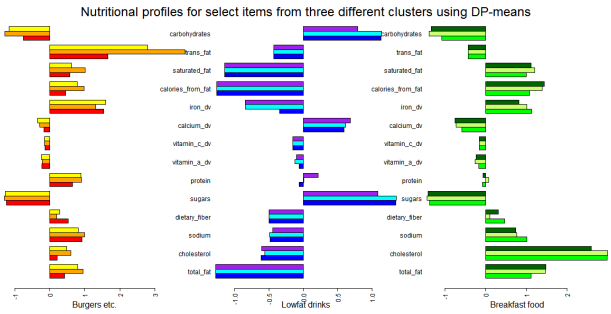


Figure 4: DP-Means

which 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 got around 11 clusters with separate clusters for burgers, lowfat drinks, breakfast items, desserts 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

6.1 Style

Papers describing final project should be up to four pages long (not including bibliography). They should contain an abstract, introduction, related work, methods, results, discussion, and conclusion section, and figures and tables where necessary, and a bibliography. These papers are due on April 16th in class.

6.2 Retrieval of style files

The formatting instructions contained in these style files are summarized in sections 7, 8, and 9 below.

7 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing of 11 points. Times

New Roman is the preferred typeface throughout. Paragraphs are separated by 1/2 line space, with no indentation.

Paper title is 17 point, initial caps/lower case, bold, centered between 2 horizontal rules. Top rule is 4 points thick and bottom rule is 1 point thick. Allow 1/4 inch space above and below title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

8 Headings: first level

First level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 12. One line space before the first level heading and 1/2 line space after the first level heading.

8.1 Headings: second level

Second level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

8.1.1 Headings: third level

Third level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the third level heading and 1/2 line space after the third level heading.

9 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

9.1 Citations within the text

Citations within the text should be numbered consecutively. The corresponding number is to appear enclosed in square brackets, such as [1] or [2]-[5]. The corresponding references are to be listed in the same order at the end of the paper, in the **References** section. (Note: the standard BIBTEX style `unsrt` produces this.) As to the format of the references themselves, any style is acceptable as long as it is used consistently.

9.2 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

9.3 Tables

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

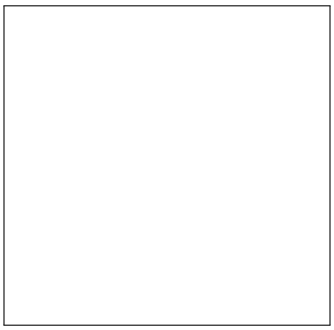


Figure 5: Sample figure caption.

Table 1: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

10 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes. Please note that pages should be numbered.

11 Preparing PostScript or PDF files

Please prepare PostScript or PDF files with paper size "US Letter", and not, for example, "A4". The -t letter option on dvips will produce US Letter files.

- Consider directly generating PDF files using `pdflatex`
- Otherwise, please generate your PostScript and PDF files with the following commands:

```
dvips mypaper.dvi -t letter -Ppdf -G0 -o mypaper.ps
ps2pdf mypaper.ps mypaper.pdf
```

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper.

References

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. CITE A LOT. Any unreferenced methods, prior work, or biological phenomenon, unless it is textbook-common, will be penalized.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.