Bayesian Nonparametric Clustering in SAS®

Hend Aljobaily; University of Northern Colorado

2019

ABSTRACT

Traditional parametric models use a fixed and finite number of parameters which cannot be used in data mining and machine learning. This is because they may result in the over or under fitting of data due to the complexity of the models used in data mining and machine learning. The Bayesian nonparametric approach is an alternative to the traditional parametric approach. Probabilistic models are appropriate nonparametric models for data mining and machine learning since they are data-driven. One example of a Bayesian nonparametric model is the Dirichlet Process model. The Dirichlet Process model is one of the most popular BNP models. For clustering, the Dirichlet Process in the Gaussian Mixture Model (GMM) is used to find the best number of clusters within the data using the **gmm** action in the **CAS** procedure within SAS®. It is able to add new clusters and remove existing clusters during the clustering process, thus finding the best number of clusters adaptively. In the **gmm** action, the Dirichlet Process serves as the prior for the proportion of the Gaussian mixture. In this study, a real-world example will be used to demonstrate the use of Dirichlet Process Gaussian Mixture Model for nonparametric clustering in SAS® to analyze a large dataset.

INRODUCTION

Generally, statistical problems are described using probability models with random variables $y_1, \dots, y_n$, where y_i could be a vector of random variables corresponding to data collected from some population of interest. One of the common assumptions is that the random variables, y_i, are drawn independently from some underlying probability distribution *G*. However, a statistical problem starts when there is uncertainty about the distribution of *G*. To illustrate, let g denote the probability density function (p.d.f.) of *G*. A statistical model is to be developed when g is known to be a member g_θ of a family $\mathcal{G} = \{\mathrm{g}_\theta : \theta \in \Theta\}$, labeled by a set of parameters θ from a set Θ. Such models are called parametric or finite-dimensional models and can be described as $\mathcal{G} = \{\mathrm{g}_\theta : \theta \in \Theta \subset \mathbb{R}^{\mathrm{p}}\}$. The aim of the analysis is then to use the observed sample to report a possible value for θ or at least to determine a subset of Θ which contains θ.

Nevertheless, in many situations, restricting inference to a specific parametric form may limit the scope and type of inferences that can be drawn from such models. Therefore, relaxing parametric assumptions is an option for obtaining a greater modeling flexibility and robustness against assumptions of a parametric statistical model. In these cases, considering models that require parameters θ in an infinite dimensional space might be a good solution. An example of an infinite-dimensional parameter space is a regression model with an unknown mean function

$m(z)$, where the space of continuous functions defined on the real line, $S = \{m(z): z \in \mathbb{R}, m(.)\}$, is a continous function. These models with infinite-dimensional parameters are referred to as nonparametric models (Ghosh & Ramamoorthi, 2003). Other popular models using nonparametric statistics are Bayesian nonparametric models. Bayesian statistics, or specifically Bayesian nonparametric statistics, remained largely theoretical except for very simple models. However, Bayesian nonparametric models have become popular since the findings of the Markov chain Monte Carlo and other Monte Carlo methods in the 1990s. To proceed with Bayesian inference in a nonparametric model we need to complete the probability model with a prior on the infinite-dimensional parameter. Such priors are known as Bayesian nonparametric (BNP) priors.

DIRICHLET PROCESS (DP)

One of the most popular BNP models is the Dirichlet process (DP) prior. The DP model is a stochastic process and was introduced by Ferguson (1973) as a prior on the space of probability measures. It is essentially a distribution over distributions where each draw from a Dirichlet process is itself a distribution (Teh, 2011).

Dirichlet Process (DP) Models

Density estimation is related to making an inference about an unknown distribution *G* based on an observed independent and identically distributed (i.i.d.) sample,

$$y_i | G \overset{iid}{\sim} G, \qquad i = 1, \dots, n.$$

Assuming a prior model on *G* requires the specification of a BNP prior. A DP, with the parameters M and G_0, is a random probability measure *G* defined on *S* which assigns probability $G(B)$ to every (measurable) set. The DP is usually denoted as $DP(MG_0)$, or $DP(M, G_0)$, where the parameter M is called the precision or total mass, and G_0 is the centering measure, and the product $\alpha \equiv MG_0$ is the base measure of the DP.

An important property of the DP is the discrete nature of *G*. As a discrete random probability measure, G could always be written as a weighted sum of point masses. Another important property of the DP is its large weak support which means that, under mild conditions, any distribution with the same support as G_0 can be approximated weakly by a DP random probability measure. The large support property means that for any finite number of measurable sets $B_1, \dots, B_m$, and $\epsilon > 0$. Some examples of approaches to the Dirichlet process are Chinese Restaurant Process (CRP), Stick-breaking process, and Polya urn scheme.

Dirichlet Process Mixture (DPM) Models

The DP generates distributions that are discrete with a probability of one which raises a problem when dealing with continuous density estimations. This problem can be fixed by using a DP random measure as the mixing measure in a m ixture over some simple parametric forms. Such

an approach was introduced by Ferguson (1983), Lo (1984), Escobar (1988, 1994), and Escobar and West (1995). Let Θ be a typical finite-dimensional parameter space. For each $\theta \in \Theta$, let f_θ be a continuous p.d.f. In many situations a normal kernel $f_\theta(y) = N(y|\mu, \sigma)$ is used with $\theta = (\mu, \sigma)$. Given a probability distribution G defined on Θ a mixture of f_G with respect to G has the p.d.f.

$$f_G(y) = \int f_\theta(y) dG(\theta).$$

The mixture model shown above and a DP prior on the mixing measure G can be written as a hierarchical model. Assume $y_i|G \overset{iid}{\sim} F_G$, then an equivalent hierarchical Dirichlet process model becomes

$$y_i|\theta_i \overset{iid}{\sim} G$$
$$\theta_i|G \overset{iid}{\sim} G$$
$$G \sim DP(MG_0)$$

In this case, the posterior distribution would be written as such:

$$G|n_1, \dots, n_K \sim DP\left(\alpha_0 + N, \frac{\alpha_0}{\alpha_0 + N} + \sum_k \frac{n_k}{\alpha_0 + N}\right)$$

Clustering Using Dirichlet Process Mixture (DPM)

Usually, clustering algorithms require the number of clusters to be determined in advance. This is problematic when the number of clusters exist in the dataset. One solution to this problem is DPM. In Bayesian nonparametric, or specifically DPM, the number of mixture components or clusters is not fixed in advance but is determined by the model and the data. The parameters of each component are generated by a DP which is a distribution over the parameters of other distributions (Vlachos, Ghahramani, & Korhonen, 2008). Thus, each occurrence is generated by the chosen component given the parameters defined as the following:

$$y_i|\theta_i \sim G$$
$$\theta_i|G \sim G$$
$$G \sim DP(MG_0)$$

An important implication of the DPM model is the fact that the DPM model induces a probability model on clusters in the following way. The discrete nature of the DP implies a positive probability for ties among the latent θ_i. Let $\theta_j^*, j = 1, \dots, k$, denote the $k \leq n$ unique values, let $S_j = \{i: \theta_i = \theta_j^*\}$, and let $n_j = |S_j|$ denote the number of θ_i tied with θ_j^*. The multiset $\rho_n = \{S_1, \dots, S_k\}$ forms a partition of the set of experimental units $\{1, \dots, n\}$. Since θ_i are random, the sets S_j are random. In other words, the DPM implies a model on a random partition ρ_n of the experimental units. The probability model on ρ_n might seem like an incidental by-product of the

construction, but the reality is that many applications of the DPM model focus on this partition ρ_n. The posterior model $p(\rho_n|y)$ reports posterior inference on clustering of the data.

SAS® PROCEDURE

To perform the above analyses in SAS®, the procedure PROC CAS in SAS® Viya™ could be used. Let us assume that there is a "big" dataset called DATA that contains multiple variables, but the variables VAR1, VAR2, VAR3, and VAR4 are the variables on interest.

Cloud Analytic Services (CAS)

SAS® Cloud Analytic Services (CAS) is the analytic server in SAS® Viya™ for data management and analytics. It provides data management and an analytics framework that can run on the cloud and provides best-in-class analytics, which SAS is known for. CAS provides user-level sessions which provides security, allows the user to share data between sessions, and provides fault tolerance. The CAS always keeps tables such as data tables and tables obtained from results in memory storage. Sometimes the entire file is kept in memory and other times only pieces of the file are mapped into memory for access to the data. This allows CAS tables to be loaded that are larger than the memory available across the grid. Below is the code that allows a user to create a connection to the CAS server:

```
options cashost="sasserver.demo.sas.com" casport=5570;
```

The CASHOST and the CASPORT statements used to specify the host on which the CAS connection runs and the port on which it listens for communications. The host information and port information for the server are stored and can be retrieved automatically whenever any specific CAS server is used.

The following code could be used to create a session in the CAS:

```
cas SESSIONNAME sessopts=(caslib=LIBRARYNAME timeout=1800
locale="en_US");
libname LIBRARYNAME cas;
caslib _all_ assign;
```

The CAS statement creates the CAS session named SESSIONNAME, and the CASLIB statement creates a library that can be accessed for data storage. The LIBNAME statement can also be used to create a library in the CAS session.

Data

The following DATA statement is used to load the dataset of interest into your CAS session by calling the library name followed by the dataset. The DATA step divides the dataset of interest into training and testing datasets. The **gmm** action uses the indicated number of observations for the training dataset, DATA_train. Next, it uses the rest of the observations, starting at 124 in this case, in the testing dataset, DATA_test.

```
data LIBRARYNAME.DATA_train;
        title "Clustering";
        set DATA (obs=123);
run;

data LIBRARYNAME.DATA_test;
        title "Clustering";
        set DATA (firstobs=124);
run;
```

Dirichlet Process Clustering

The PROC CAS procedure uses the Gaussian Mixture Model **gmm** action. It is able to add new clusters and remove existing clusters during the clustering process which result in finding the ideal number of clusters. In the **gmm** action, the Dirichlet Process serves as the prior for the proportion of the Gaussian mixture. The following PROC CAS statements use the **gmm** action to perform clustering on the training dataset:

```
proc cas;
  action nonParametricBayes.gmm
    table={name="ELS_train"},
    inputs={"VAR1", "VAR2", "VAR3", "VAR4"},
    seed=1234567890,
    nThreads=32,
    maxClusters=100,
    alpha=1,
    inference={method="VB", maxVbIter=1000,
    covariance="DIAGONAL", threshold=0.001},
    output={casOut={name="score", replace=TRUE},
    copyVars={""VAR1", "VAR2", "VAR3", "VAR4", "VAR5"}},
    clusterSumOut={name="clustersum", replace="TRUE"},
    clusterCovOut={name="clustercov", replace="TRUE"},
    display={names={"NObs", "DescStats", "ModelInfo"}},
    saveState={name="astore", replace="TRUE"};
run;
```

The **table** parameter indicates the data table to be analyzed. The **inputs** parameter specifies the input variables to use in clustering. The **seed** parameter specifies the random seed to use in clustering. The **nThreads** parameter specifies the number of threads to use in the parallel computation. The **maxClusters** parameter specifies the maximum number of possible clusters. The **alpha** parameter indicates the mass parameter of the Dirichlet process. The **inference** parameter specifies the method used for clustering which is set to Variational Bayes (VB),

the **maxVbIter** parameter specifies the maximum number of VB iterations, the **covariance** parameter specifies the type of the covariance matrices for the Gaussian mixture model, the **threshold** parameter specifies the threshold of the convergence of the VB iterations, the **output** parameter stores the clustering scores to the LIBRARYNAME.score data table, the **clusterSumOut** parameter stores the clusters summary to the LIBRARYNAME.clustersum data table, and the **clusterCovOut** parameter stores the clusters covariance to the mycas.clustercov data table.

Also, the **gmm** uses the **saveState** parameter to save the trained GMM to the LIBRARYNAME.astore data table to be used for the testing dataset. This can be achieved using the code shown below for the **score** action in the **aStore** action set:

```
proc cas;
  action aStore.score
    table={name='DATA_test'},
    out={name='newscore'},
    rstore={name='astore'};
run;
```

Plotting Clusters' Distributions

Histograms can be used to visualize the distribution of each cluster. When looking at multiple histograms, or distributions in this case, it is best that all histograms have the same bin width and anchor locations for easier comparison. A comparative histogram enables users to compare two or more distributions. The PROC SGPLOT procedure enables you to use more than one HISTOGRAM statement by overlaying the histograms or distributions of different clusters. The data table saved from the previous step as "newscore" is used to plot the distributions of all clusters in the dataset of interest.

```
title "Clusters";
proc sgplot data=LIBRARYNAME.newscore;
 histogram _ClUSTER_1_ / binwidth=2 transparency=0.5
          name="Cluster l" legendlabel="Cluster 1";
 histogram _ClUSTER_2_ / binwidth=2 transparency=0.5
          name="Cluster 2" legendlabel="Cluster 2";
 histogram _ClUSTER_3_ / binwidth=2 transparency=0.5
          name="Cluster 3" legendlabel="Cluster 3";
 histogram _ClUSTER_4_ / binwidth=2 transparency=0.5
          name="Cluster 4" legendlabel="Cluster 4";
 density _ClUSTER_1_ / type=kernel ;
 density _ClUSTER_2_ / type=kernel ;
 density _ClUSTER_3_ / type=kernel ;
 density _ClUSTER_4_ / type=kernel ;
 xaxis label="CLUSTERS" min=0;
```

```
  keylegend "Cluster 1" "Cluster 2" "Cluster 3" "Cluster 4" / across=2 position=TopRight
location=Inside;
run;
```

CONCLUSION

Effective clustering is one of the important tasks in statistics generally and machine learning specifically as well as in other fields of study. Cluster analysis, also known as clustering, is defined as the grouping of a set of objects in such a way that objects in the same cluster are more similar to each other than other objects in different clusters. Traditional and parametric models usually have insufficient results for "big data" and machine learning due to the complexity of data resulting from the wrong fitting. They also have difficulties determining the appropriate number of components or clusters in a mixture model. Thus, the Bayesian nonparametric models are useful alternatives to the traditional parametric models. Bayesian nonparametric methods are data-driven methods which use infinite parameterization to determine an appropriate model complexity in a fully Bayesian manner. These infinite-dimensional nonparametric representations can be used to learn about the structure of data including the appropriate number of clusters within a dataset. An example of a Bayesian nonparametric model is the Dirichlet Process model. The Dirichlet Process model is one of the most popular BNP models. For clustering, the Dirichlet Process in the Gaussian Mixture Model (GMM) is used to find the best number of clusters within the data. It can add new clusters and remove existing clusters during the clustering process which allows for the ideal number of clusters to be found adaptively. In SAS® Viya, the CAS procedure uses the `gmm` action to analyze data using Dirichlet Process prior to the Gaussian mixture model.

REFERENCES

Dubey, A. (2015). *Bayesian Nonparametrics: Dirichlet Processes*. Lecture. Retrieved July 01, 2019, from https://pdfs.semanticscholar.org/4181/e193087e7bb1cd0c267438b14162aeeccee1.pdf

Ferguson, T. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics, 1*(2), 209-230. Retrieved from http://www.jstor.org/stable/2958008

Ghosh, J. K., & Ramamoorthi, R. V. (2013). *Bayesian nonparametrics*. New York: Springer.

Müller, P., Quintana, F. A., Hanson, T., & Jara, A. (2015). *Bayesian nonparametric data analysis*. Cham: Springer.

Pendergrass, J. (2017). *The Architecture of the SAS® Cloud Analytic Services in SAS® Viya™*. Retrieved June 18, 2019, from https://support.sas.com/resources/papers/proceedings17/SAS0309-2017.pdf

SAS Visual Data Mining and Machine Learning 8.3: Procedures. (2018, June 07). Retrieved July 18, 2019, from https://documentation.sas.com/?docsetId=casactml&docsetTarget=casactml_nonparametricbayes_details12.htm&docsetVersion=8.3&locale=en

Teh Y. (2011) Dirichlet Process. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.

Vlachos, A., Ghahramani, Z., & Korhonen, A. (2008). ICML Workshop on Prior Knowledge for Text and Language Processing. In *Dirichlet Process Mixture Models for Verb Clustering*. Helsinki, Finland. Retrieved June 18, 2019, from http://mlg.eng.cam.ac.uk/pub/pdf/VlaGhaKor08.pdf

Wicklin, R. (2016, March 09). Comparative histograms: Panel and overlay histograms in SAS. Retrieved June 26, 2019, from https://blogs.sas.com/content/iml/2016/03/09/comparative-panel-overlay-histograms-sas.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Hend Aljobaily
University of Northern Colorado
hend.aljobaily@unco.edu
hend.aljobaily@hotmail.com