Bayesian Statistics Project: Partition Estimation With Model Misspecification

De Diego,F., Ischia,M., Jimenez,J.

Politecnico di Milano

fernandoantonio.dediego@mail.polimi.it
martino.ischia@mail.polimi.it
juliandavid.jimenez@mail.polimi.it

November 22, 2019

Overview

The Clustering Problem

- Cluster analysis is the process of grouping or segmenting a collection of objects into subsets ("clusters"), such that those within each cluster are more closely related to one another than objects assigned to different clusters [Hastie, Tibshirani, Friedman, 2009].
- Classical Algortithms:
  - Agglomerative Hierarchical Clustering
  - k-means Clustering
- Model-based clustering methods using finite mixture models. Where each mixture component corresponds to a cluster.

Bayesian Nonparametric Clustering

- The data is assumed conditionally i.i.d. with density

$$f(y|P) = \int K(y|P)dP(\theta)$$

  Where $K(y|P)$ is a specified parametric density on the sample space with mixing parameter $\theta \in \Theta$ and P is a probability measure on Θ.
- As a prior on the mixing measure we select a Dirichlet process.
- The Dirichlet process has a stick-breaking representation: $v_1, v_2, ... \overset{i.i.d.}{\sim} B(1, \alpha), w_j = v_j \prod_{i=1}^{j-1}(1 - v_i)$ for $j = 1, 2, ...$ and $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, ... \overset{i.i.d.}{\sim} P_0$, then the random discrete measure

$$P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$$

  is distributed as a $DP(\alpha, P_0)$ [Sethuraman, 1994].

Bayesian Nonparametric Clustering

- The density is modeled with a countably infinite mixture model
- Since P is discrete a.s. this model induces a latent partinioning c of the data.
- The partition can be represented by $\mathbf{c} = (C_1, ..., C_{k_N})$, where C_j contains the indices of data points in the j^{th} cluster and k_N is the number of clusters in the sample of size N.

Bayesian Nonparametric Clustering

- The number of mixture components is infinite.
- There are

$$S_{N,k} = \frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(k-j)^N$$

  a Stirling number of the second kind ways to partition the N data points into k groups and

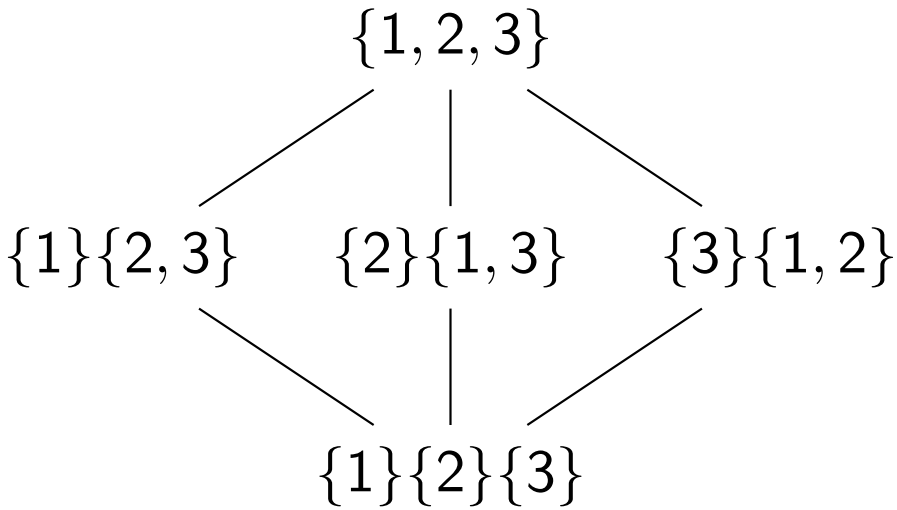
$$B_N = \sum_{k=1}^{N} S_{N,k}$$

  a Bell number possible partitions of the N data points (Intractable).
- Thus MCMC techniques are employed.

Latent Partitioning

- The partition space $\mathbf{C}$ equipped with $\leq$ is a partially ordered set.
- For $\mathbf{c}, \hat{\mathbf{c}} \in \mathbf{C}, \mathbf{c} \leq \hat{\mathbf{c}}$ if for all $i = 1, ..., k_N$, $C_i \subseteq \hat{C}_j$ for some $j \in \{1, ...\hat{k_N}\}$.
- For any $\mathbf{c}, \hat{\mathbf{c}} \in \mathbf{C}, \mathbf{c}$ is covered by $\hat{\mathbf{c}}$, denoted by $\mathbf{c} \prec \hat{\mathbf{c}}$ if $\mathbf{c} < \hat{\mathbf{c}}$ and there is no $\hat{\hat{\mathbf{c}}} \in \mathbf{C}$ such that $\mathbf{c} < \hat{\hat{\mathbf{c}}} < \hat{\mathbf{c}}$.
- The partition space forms a lattice as every pair of partitions has a *greatest lower bound* **(g.l.b.)** and a *lowest upper bound* **(l.u.b.)**.
- Meet operation: $\mathbf{c} \wedge \hat{\mathbf{c}} = g.l.b.(\mathbf{c}, \hat{\mathbf{c}})$

Latent Partitioning



Hasse Diagram for the lattice of partitions with sample size $N = 3$. A line is drawn from $\mathbf{c}$ up to $\hat{\mathbf{c}}$ when $\mathbf{c}$ is covered by $\hat{\mathbf{c}}$.

Point Estimation for Clustering

What is an appropriate point estimate of the partition, based on the posterior?

Some simple approaches are:

- The posterior mode as point estimate.
- Use the posterior similarity matrix to get a point estimate.

A more rigorous approach is to define a loss function over partitions to obtain a point estimate. Two specific loss functions will be analyzed:

- Binder's Loss.
- Variation of Information (VI).

Binder's Loss

- The Binder's Loss is a quadratic function that penalizes locating two observations in the same cluster when they should be in different clusters, and locating the observations in different clusters when they should be in the same one.

- If $\mathbf{c}$ is the true clustering and $\hat{\mathbf{c}}$ is its estimation, then the Binder's Loss is

$$B(\mathbf{c}, \hat{\mathbf{c}}) = \sum_{n<n'} \mathbb{I}(c_n = c_{n'})\, \mathbb{I}(\hat{c_n} \neq \hat{c_{n'}}) + \mathbb{I}(c_n \neq c_{n'})\, \mathbb{I}(\hat{c_n} = \hat{c_{n'}})$$

where $\mathbb{I}$ is the indicator function, and the partitions of size N are represented as $\mathbf{c} = (c_1, ..., c_N)$, where $c_n = j$ if the n^{th} data point is in the j^{th} cluster.

Variation of Information

- The Variation of Information (VI) compares the information contained in two clusterings with the information shared between the two clusterings. Its formula is the following:

$$VI(\mathbf{c}, \hat{\mathbf{c}}) = H(\mathbf{c}) + H(\hat{\mathbf{c}}) - 2I(\mathbf{c}, \hat{\mathbf{c}})$$

  Here, the first two terms represent the entropy of the two clusterings, and the last term is the mutual information between them.
- Since $I(\mathbf{c}, \hat{\mathbf{c}}) = H(\mathbf{c}) + H(\hat{\mathbf{c}}) - H(\mathbf{c}, \hat{\mathbf{c}})$, the VI can be rewritten as

$$VI(\mathbf{c}, \hat{\mathbf{c}}) = -H(\mathbf{c}) - H(\hat{\mathbf{c}}) + 2H(\mathbf{c}, \hat{\mathbf{c}})$$

Comparison between Binder's Loss and VI

The ($N-$invariant) formulas of both loss functions are:

$$\tilde{B}(\mathbf{c}, \hat{\mathbf{c}}) = \sum_{i=1}^{k_N} (\frac{n_{i+}}{N})^2 + \sum_{j=1}^{\hat{k_N}} (\frac{n_{+j}}{N})^2 - 2 \sum_{i=1}^{k_N} \sum_{j=1}^{\hat{k_N}} (\frac{n_{ij}}{N})^2$$

$$VI(\mathbf{c}, \hat{\mathbf{c}}) = \sum_{i=1}^{k_N} \frac{n_{i+}}{N} log_2(\frac{n_{i+}}{N}) + \sum_{j=1}^{\hat{k_N}} \frac{n_{+j}}{N} log_2(\frac{n_{+j}}{N}) - 2 \sum_{i=1}^{k_N} \sum_{j=1}^{\hat{k_N}} \frac{n_{ij}}{N} log_2(\frac{n_{ij}}{N})$$

where $n_{ij} = |C_i \cap \hat{C}_j|$, $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$.

Comparison between Binder's Loss and VI

Binder's Loss and VI share multiple properties, and also have important differences.

Among the *shared properties* we find:

- Both loss functions are metrics on the space of partitions.
- If $\mathbf{c} \geq \hat{\mathbf{c}} \geq \hat{\hat{\mathbf{c}}}$, then:
  - $d(\mathbf{c}, \hat{\hat{\mathbf{c}}}) = d(\mathbf{c}, \hat{\mathbf{c}}) + d(\hat{\mathbf{c}}, \hat{\hat{\mathbf{c}}})$.
  - $d(\mathbf{c}, \hat{\mathbf{c}}) = d(\mathbf{c}, \hat{\mathbf{c}} \wedge \mathbf{c}) + d(\hat{\mathbf{c}}, \hat{\mathbf{c}} \wedge \mathbf{c})$.

Some of their *differences* are the following:

- Binder's Loss tends to estimate a partition with more clusters than VI.
- $VI(\mathbf{c}, \hat{\mathbf{c}}) \leq log_2(N)$, while $\tilde{B}(\mathbf{c}, \hat{\mathbf{c}}) \leq 1 - \frac{1}{N}$.

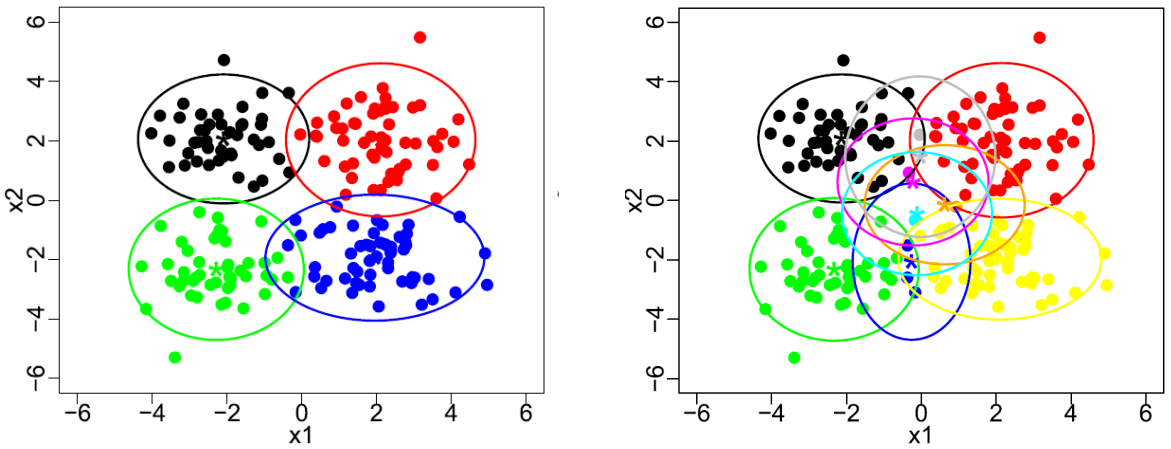
Actual implementation

$$\mathbf{c}^* = \arg\min_{\hat{\mathbf{c}}} \mathbb{E}\left[L(\mathbf{c}, \hat{\mathbf{c}}) \mid data\right]$$

- The posterior distribution of the partition is estimated from the output of a Markov Chain
- In the case Variation Information, for a given $\hat{\mathbf{c}}$ the Monte Carlo approximation of $\mathbb{E}[L(\mathbf{c}, \hat{\mathbf{c}}) \mid data]$ is of order MN^2 where M is the numbers of elements of the Markov chain
- For any practical problem, exploring every partition is impossible: $B_{20} = 51.7 \cdot 10^{12}$
- A possibility is to restrict the set of partitions for the optimization problem
- Another possibility: greedy search algorithm

Building up a simulation study

- Validation of the results found by [Wade, Ghahramani, 2018]: the VI criterion gives as result a partition that is in line with our intuitive idea of clustering .



(c) Ex 1 VI: 4 clusters

(a) Ex 1 Binder's: 9 clusters

Figure: [Wade, Ghahramani, 2018]

Building up a simulation study

- Simulate data from known distributions.
- Running simulations multiple times due to the intrinsic randomness.
- Computational issues: we will have to write the code in C++.

The project

Study how this two different strategies behave in the case the **model is misspecified**. 3 possible ways to misspecify the model:

- Through the mean parameter, setting it far away from the true value.
- Through the variance parameter, setting a flat distribution or a spiky one.
- Through the α parameter of the Dirichlet prior, related to the weights of the mixture and consequently the number of clusters.

Do the two methods behave in the same way?

Possible Extension

- Consider a different formulation of the Binder's loss, that takes into account triplets instead of pairs.

References

- Wade, S., Ghahramani, Z. (2018) *Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion).* Bayesian Anal. 13 (2018), no. 2, 559–626. doi:10.1214/17-BA1073. https://projecteuclid.org/euclid.ba/1508378464
- Sethuraman, Jayaram. (1994). A Constructive Definition of the Dirichlet Prior. Statistica Sinica. 4. 639-650.
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Springer-Verlag.
- Miller, J. (2018). *An elementary derivation of the Chinese restaurant process from Sethuraman's stick-breaking process*. arXiv e-prints, arXiv:1801.00513. process from Sethuraman's stick-breaking process