AISTATS 2010
Thirteenth International Conference on Artificial Intelligence and Statistics
May 13-15, 2010
Sardinia
Italy

Reviews For Paper

Track	Peer-reviewed Papers
Paper ID	65
Title	An Alternative Prior Process for Nonparametric Bayesian Clustering

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Overall Rating	Borderline Accept--Reasonably good paper but lacking in some respects
Confidence	Very Confident--I am an expert and have fully checked the paper
Summary	The paper discusses the probability law on a random partition of (1,...,n) implied by the Dirichlet and the Pitman-Yor processes, and in particular its "rich-get-richer" characteristic. The authors argue that, in some applications, this property is not desirable, and discuss an alternative predictive rule, called "uniform process". This process has been suggested in the literature, and here the authors further explore its implications on the clustering structure, providing some theoretical results and studying its performance in some applications. The authors underline that the uniform process does not lead to an exchangeable partition probability law, but they argue that this may be natural in several applied contexts. The paper is well presented, and the discussed issues are of interest; however, there are some aspects that I would ask the authors to clarify.
Detailed Comments	As reviewed in Section 2.1 of the paper, the (infinite colors) Polya urn scheme defines the probability law of a sequence of random variables; Blackwell and Mac Queen showed that this sequence is exchangeable, with de Finetti measure given by a Dirichlet process. In this scheme, the predictive probability of ϕ_{n+1} given $\phi_1, \ldots, \phi_n$ only depends on the number of distinct values of the sample, on on their frequencies; it is this characteristic that implies the "rich-get-richer" property of the clustering. So, some care is needed in modifying the latter property, since it implies to remove a quite natural assumptions on the role of the empirical frequencies in the updating scheme. A more general framework is given by species sampling models, and the Pitman-Yor process is an interesting example, that allows more flexibility by introducing an additional parameter. These results are nicely reviewed in the paper; a further useful reference would be Pitman (1996). In particular, Theorem 14 in Pitman (1996) clarifies the relationship between exchangeability of the sequence of draws $(\phi_i$) and exchangeability of the partition probability function (PPF). The uniform process seems to be a special case of a species sampling model, where the predictive rule only depends on the number of distinct values. If this is correct, and, as the authors underline, the uniform process does not lead to an

	exchangeable PPF, then by the above result the sequence (ϕ_i) is not exchangeable. This is not necessarily a drawback; but I think that the authors should explore what kind of dependence is implied for the sequence of parameters (ϕ_i). In some applications in mixture models and clustering, exchangeability of the individual parameters (ϕ_i) may not be a natural assumption; but, with a uniform process, what are we assuming on the probability law of the sequence (ϕ_i)? At the beginning of Section 5, the authors argue that, in some applications, there is a natural temporal ordering; would it be possible to give further hints about that? Does the uniform process succeed in modeling a temporal dependence? The predictive rule of the uniform process is somehow weaker than the DP, since it only depends on the number of distinct values; intuitively, I would expect that this implies a quite weak dependence among the ϕ_i's. ADDITIONAL REFERENCES Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. In "Statistics, Probability and Game Theory; Papers in honor of David Blackwell", 245--267.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Overall Rating	Accept--Good solid paper, top 50%
Confidence	Very Confident--I am an expert and have fully checked the paper
Summary	The abstract proposes to study the relative performance of random partitions implied by the Dirichlet process (DP), the Pitman-Yor (PY) process and another, less commonly used one, the uniform process. In particular, the authors focus on the stochastic ordering of cluster sizes implied by the DP and PY priors.
Detailed Comments	The abstract addresses the important problem of choice of probability models for random partitions. The authors ask the right question by focusing on the implied distribution of cluster sizes, which is severely biased under the ubiquitous DP model. They instead propose to use uniform predictive probabilities to define an alternative random partition. The authors correctly recognize a major drawback of the uniform model -- it is not exchangeable. I disagree with the authors that exchangeability is just another model assumption. Without exchangeability the model is not on the partitions of a set of experimental units, but rather on a numbered sequence of units. Order suddenly matters, in inappropriate ways.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Overall Rating	Accept--Good solid paper, top 50%
Confidence	Very Confident--I am an expert and have fully checked the paper

Summary	The authors describe a particular process, the uniform process, which can be used for clustering. The process provides an interesting alternative to the methods which currently dominate nonparametric Bayesian approaches to clustering. The key distinction that the authors draw is between models that describe the data as conditionally independent and identically distributed draws from some distribution (exchangeable in the language of the paper) and the uniform process (which does not lead to an exchangeable distribution on the observables). Properties of the process are developed and contrasted to those of the Dirichlet process and Pitman-Yor process. Comparisons are made on a document clustering example. In the comparisons, the uniform process outperforms the Dirichlet process for out-of-sample prediction.
Detailed Comments	This is an engaging and well-written paper, and the need for the development of novel prior distributions for nonparametric Bayesian clustering is clear. A few comments. 1. Technical bits. Page 2, section 2.1. Better to refer to G_0 as the base distribution or base cdf. For the Dirichlet process, the base measure generally refers to "\theta G_0". (A similar comment applies in 2.2). The description "are said to belong to the same cluster iff \psi_n = \psi_m" relies on a continuous G_0. If not, two distinct draws from G_0 (hence, two distinct clusters) can have the same value of \psi. The Polya urn scheme (see Blackwell and MacQueen, Annals of Mathematical Statistics) is an earlier description of the predictive probability phenomenon described at the end of page 2. To me, it is more descriptive than is the Chinese restaurant. End of section 2.1. Earlier breaks of the stick "tend to" lead to larger random weights ... Page 3, top of right-hand column. I'm comfortable with the description of a model as exchangeable (or not). The description of a partition as exchangeable is, I think, less standard. You might consider alternative wording here. 2. The authors advocate that one discard the assumption of "exchangeability for all N"--equivalent to a model of conditionally i.i.d. data--as a means of changing the large-sample behavior of the number of clusters and cluster sizes. This is an interesting approach, as it implicitly discards the tradition in nonparamtric Bayesian analysis of placing a distribution over distributions and then modelling the data as draws from the realized distribution. In contrast, my reaction would be to attempt to model the latent distributions and the relationship between different observations. No need for a change here, just a different perspective. One exchangeable (for finite N) model would first write the uniform process and then follow it with a uniform distribution over all permutations of the data (including the hold-out sample, if desired). I suspect that one could add a permutation step to the MCMC algorithm. Is this feasible? Does it change

performance?

3. The treatment of \beta is mysterious. At one point, it is a parameter that governs a prior distribution. Later, it appears to be fixed. A word or two to clarify would help!

4. The comparisons show that the uniform process outperforms the Dirichlet process. How does a Pitman-Yor process (with parameters that make it different from the Dirichlet process) perform? If possible, add an example or two to Figure 4.

5. Additional connections to the literature.

Peter Green, possibly with Sylvia Richardson, had strong commentary on the "rich get richer" property of the Dirichlet process circa 1995. Though I don't have a reference handy, it would be worth trying to track down their work.

The uniform process may be connected to dependent Dirichlet processes or, more generally, to dependent nonparametric processes (e.g., MacEachern, 1999, 2000). I haven't taken the time to work through the math to see whether the uniform process fits into this framework, but a surprising number of models do. Since the hierarchical Dirichlet process does, this is a direction worth exploring.