Clustering bayesiano no paramétrico para grandes volúmenes de datos

Eduardo Vásquez
Departamento de Estadística, Pontificia Universidad Católica de Valparaíso

Resumen

Dado su alto nivel de flexibilidad, los modelos Bayesianos no paramétricos son cada vez más utilizados en el contexto de aprendizaje estadístico. En esta presentación, se introducirán los Procesos de Dirichlet (DP), centrándonos principalmente en la característica discreta de sus realizaciones, la cual es ampliamente explotada en el contexto de estimaciones de densidades y clustering mediante los *Dirichlet Process Mixtures*. Para realizar la inferencia a posteriori, se han desarrollado diferentes algoritmos a partir del muestreo de Gibbs, partiendo por los trabajos de Michael Escobar y Mike West, los cuales solo eran aplicables en un número reducido de casos (Escobar y West, 1995), culminando con el trabajo de Neal que funciona para el caso general (Neal, 2000). Pese a que este algoritmo es de aplicación general, aún está el inconveniente del alto costo computacional de estos métodos para tamaños muestrales grandes. Así, se presentará un nuevo enfoque al problema de grandes volúmenes de datos, que considera la aplicación de este algoritmo en diferentes etapas. Finalmente, se presentan algunas aplicaciones de este nuevo algoritmo, junto con comparaciones a otros algoritmos de clustering.

Referencias Bibliográficas

1. Escobar, M.D., West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577-588.
2. Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249-265.
3. Ni, Y., Müller, P., Diesendruck, M., Williamson, S., Zhu, Y., Ji, Y. (2020). Scalable bayesian nonparametric clustering and classification. *Journal of Computational and Graphical Statistics* **29**, 53-65.
4. Ross, G.J., Markwick, D., Mulder, K., Sighinolfi, G. (2020). dirichletprocess: Build Dirichlet process objects for bayesian modelling. URL: `https://CRAN.R-project.org/package=dirichletprocess`.