

Random Partition Distributions Indexed by Pairwise Information

David Dahl, Brigham Young University

We present a suite of new random partition distributions indexed by pairwise distances in which items that are "close" to each other cluster with higher probability than distant ones. We develop two novel statistical methods using these distributions. First, the distributions form the basis for a cluster analysis technique which uses the same input data as does hierarchical clustering, yet provides a means to assess clustering uncertainty. Second, the distributions yield a new class of Bayesian models in which pairwise distances can influence the clustering a priori. As with many Bayesian nonparametric models, our approach is based on random partitioning of items into clusters where items in a cluster share a common value for model parameters. Whereas the commonly-used Dirichlet process prior induces a clustering distribution which is uniform across all pairs, our random partition distributions allows pairwise information to enter the prior distribution. We explore the properties of our distributions and show how to make posterior inference. Several applications are presented to demonstrate the advantage of the proposed methodology.