Literature Report on DP-means

Aaron Maurer

STAT 302, Spring Quarter 2015

1 Overview

For this report, I chose to read "Revisiting k-means: New Algorithms via Bayesian Nonparametrics"[1]. This article introduces a alternative to the k-means clustering algorithm called DP-means, where instead of specifying a number of clusters at the start, then working to optimize their fit, new clusters are introduced when a point in the data set is too far from one of the existing clusters. This in itself is a 'hard' clustering algorithm, which outputs an assignment to clusters, rather than a posterior distribution one might normally expect from a Bayesian method. However, this paper demonstrates that this algorithm is the limit of a Bayesian model where the total distribution is a mixture of Gaussian distributions, with the number of components coming from a Dirichlet process. This is the source of the name of the algorithm: Dirichlet process means or DP-means.

In addition to a description and derivation of this algorithm, the paper offers a few extensions and simulations. The most interesting of theses is a hierarchical clustering method where, over multiple data sets of the same variables, local clusters are simultaneously fit on each data set so as to match a set of global clusters across all of the data sets. This is again a hard clustering algorithm, but it can be derived as well in a Bayesian fashion by taking the limit when the parameters of the Dirichlet process which generates the components in each data set arises from a prior global Dirichlet process. The next extension shows how DP-means extends to spectral clustering; where one would spectrally cluster using k-means by performing k-means on the first k eigenvectors of a similarity matrix for the data, instead with DP-means one takes all eigenvectors where the eigenvalues are above a given threshold, then cluster via DP-means using that same threshold to determine when to add additional components. Similarly, the author also shows that DP-means can be extended to graph cut problems. The paper concludes with a few simulations demonstrating the effectiveness of DP-means and its multiple data set extension.

[1] Brian Kulis and Michael I. Jordan, "Revisiting k-means: New Algorithms via Bayesian Nonparametrics", *CoRR* (2011): http://arxiv.org/abs/1111.0352

This article offers obvious extensions of how the classical Bayesian statistics we learned can be extended to a machine learning algorithm. The prior Dirichlet process is used to guide the posterior number of components. The method for fitting the model is the limit of a Gibbs sampling algorithm. Finally, the multiple data set version of DP-means is a hierarchical model, based on the exchangeability of the data sets. All together, this forms an interesting competitor to k-means, being similarly easy to compute, but built on classic Bayesian principles.

2 Paper Contents

2.1 DP-Means Algorithm

The idea of clustering arises, in statistical terms, from the idea that a set of random variables are drawn from a mixture distribution. In other words, a random variable X_i is drawn in a two stage process. First, a multinomial variable z_i with k possible outcomes is drawn. Then, $X_i \mid z_i \sim F_{z_i}$, where $\left\{F_j \mid j \in \{0, ..., k\}\right\}$ is some set of probability distributions. The goal of a clustering method is then to impute the z_i based on the X_i, revealing important underlying structure in the data. Bayesian models provide a natural way to fit probability distributions to z_i, in particular without choosing k a priori, but there issue is that they, classically, are complicated to compute on large data sets and don't necessarily scale well.

Thus, the k-means algorithm, where k must be specified beforehand, remains the most commonly used algorithm. This method can be thought of as designed for a model where z_i is multinomial k, for fixed k, and then $X \mid z_i \sim N_p(\mu_j, \sigma I_p)$ for some set of mean vectors $\left\{\mu_j \mid j \in \{0, ..., k\}\right\}$. To find $\{\mu_i\}$ using k-means, one picks initial guesses for the $\hat{\mu}_j$, then alternates assigning $\hat{z}_i = \arg\min_j \|X_i - \hat{\mu}_j\|$ and

$$\hat{\mu}_j = \frac{1}{\left|\{i \mid z_i = j\}\right|} \sum_{i \mid z_i = j} X_i$$

until the μ_j and z_i have converged. This method allows little flexibility, but is relatively easy to compute, even for large data sets. The authors of the paper notes that the fitting process is the limit of what the EM algorithm would do if σ goes to 0.

In designing the DP-means algorithm, the authors "attempt to achieve the best of both worlds by designing scalable hard clustering algorithms from a Bayesian nonparametric viewpoint." To this end, they start with a purely Bayesian model for how the data would arise which puts a prior on k, and then similarly derives a simple algorithm for hard clustering that represents the limit of fitting this Bayesian algorithm as a variance parameter is sent to 0.