K-means vs GMM Cost Function

Thomas Athey

April 4, 2019

1 Problem Definition

1.1 K-means

Say we have the dataset

$$X = \{x_1, x_2, \ldots, x_n\} \epsilon (R^p)^n$$

that we seek to partition into k clusters. The k-means algorithm seeks to find the partition $\mathscr{P}$ that minimizes the cost function:

$$J_{km,p}(\mathscr{P}; X) = \sum_{i=1}^{k} \sum_{x \epsilon \mathscr{P}_i} ||x - \mu(\mathscr{P}_i)||_2^2 \tag{1}$$

where $\mu(\mathscr{P}_i)$ is the mean of the points in $\mathscr{P}_i$. Since the mean is the least squares estimator of a set of data, we know that J_{km} is a lower bound to the more general cost function that is also a function of a set of means:

$$J_{km,g}(\mathscr{P}, \Theta; X) = \sum_{i=1}^{k} \sum_{x \epsilon \mathscr{P}_i} ||x - \mu_i||_2^2 \tag{2}$$

where $\Theta = \{\mu_1 ... \mu_k\}$. Given a particular Θ, it is easy to see that the partition $\mathscr{P}$ that minimizes (2) is that which associates each datapoint to the closest mean. Let's denote the group assignment as c_i and thus, the closest mean as μ_{c_i}. Now, we can write an equivalent k-means cost function that depends on Θ, and not $\mathscr{P}$:

$$J_{km}(\Theta; X) = \sum_{i=1}^{n} ||x - \mu_{c_i}||_2^2 \tag{3}$$

Note that:

$$\min_{\mathscr{P}} J_{km,p}(\mathscr{P}; X) = \min_{\Theta} J_{km}(\Theta; X)$$

1.2 Maximum Likelihood Estimation for Gaussian Mixture

Say we observe data, X, and we have a probability model for this data, $P(X;\Theta)$, but Θ is unknown. The maximum likelihood method estimates Θ as:

$$\Theta_{MLE} = \underset{\Theta}{\operatorname{argmax}}\, P(X;\Theta) = \underset{\Theta}{\operatorname{argmax}}\, logP(X;\Theta)$$

If the observations are independent then we can factor $P(X;\Theta)$ into a product to obtain:

$$\Theta_{MLE} = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} logP(x_i;\Theta)$$

In Gaussian mixture models (GMMs), we assume that the probability density of x is a weighted sum of Gaussian densities, and the parameters we are trying to estimate are the weights, means, and covariances (i.e. $\Theta = \{(\tau_1,\mu_1,\Sigma_1),...(\tau_k,\mu_k,\Sigma_k)\}$). An intuitive interpretation of this model is that each datapoint has an associated hidden variable z_i which denotes the particular Gaussian from which x_i is generated:

$$P(x_i,z_i;\Theta) = P(z_i|\Theta)P(x_i|z_i;\Theta) = \tau_{z_i}N(x_i;\mu_{z_i},\Sigma_{z_i})$$

$$P(x_i;\Theta) = \sum_{z=1}^{k} P(z;\Theta)P(x_i|z;\Theta) = \sum_{z=1}^{k} \tau_z N(x_i;\mu_z,\Sigma_z)$$

So, the function that we are trying to maximize is:

$$L_{gmm}(\Theta;X) = \sum_{i=1}^{n} log \sum_{z=1}^{k} \tau_z N(x_i;\mu_z,\Sigma_z) \tag{4}$$

An important results of maximum likelihood estimators is that in the case of independent observations, as the number of observations increases, the estimator will converge to the true parameter values.

Once the parameters are obtained, datapoints can be assigned to a particular mixture according to the Bayes decision rule:

$$c_i = \underset{j}{\operatorname{argmax}}\, \tau_j N(x_i;\mu_j,\Sigma_j) \tag{5}$$

1.3 Problem

Minimizing (3), and maximizing (4) followed by (5) both result in a partition of the observations X into k groups. Our goal is to investigate when and how these clustering procedures are related.

2 Connecting GMM and K-Means with a Simple Case

2.1 GMM

Say we make the following assumptions in GMM estimation:

- $k = 2$
- $\Sigma_i = I$
- $\tau_i = \frac{1}{2}$

Therefore, we only need to estimate the two mixture means, and our cost function becomes (after removing constants):

$$L_{gmm,s}(\{\mu_1, \mu_2\}; X) = \sum_{i=1}^{n} log(e^{-\frac{1}{2}||x_i - \mu_1||_2^2} + e^{-\frac{1}{2}||x_i - \mu_2||_2^2}) \tag{6}$$

2.2 Approximation of $J_{gmm,s}$

In (6), consider the approximation:

$$e^{-\frac{1}{2}||x_i - \mu_1||_2^2} + e^{-\frac{1}{2}||x_i - \mu_2||_2^2} \approx \max_j e^{-\frac{1}{2}||x_i - \mu_j||_2^2}$$

If $|||x_i - \mu_1||_2^2 - ||x_i - \mu_2||_2^2| \geq 3$ then the approximation has less that 5% error.

With this approximation, we have:

$$L_{gmm,s}(\{\mu_1, \mu_2\}; X) \approx -\frac{1}{2} \sum_{i=1}^{n} ||x_i - \mu_{c_i}||_2^2$$

where c_i chooses the mean closest to x_i. Notice that this is the negation of the cost function in (3). Thus, even though the problems of k-means and GMM estimation are fundamentally different, they can be linked through this approximation.

3 Approximations of Other Cases, k=2

In all of these cases, c_i chooses the cluster membership that minimizes the cost function. Also, constants are removed from the sum within the log.

3.1 Unconstrained τ_i

$$L_{gmm,s}(\{\mu_1, \mu_2\}; X) = \sum_{i=1}^{n} log(\tau_1 e^{-\frac{1}{2}||x_i - \mu_1||_2^2} + \tau_2 e^{-\frac{1}{2}||x_i - \mu_2||_2^2})$$

$$L_{gmm,s}(\{\mu_1, \mu_2\}; X) \approx \sum_{i=1}^{n} log(\tau_{c_i}) - \frac{1}{2}||x_i - \mu_{c_i}||_2^2$$

3.2 $\Sigma_i = v_i I$

$$L_{gmm,s}(\{\mu_1, \mu_2\}; X) = \sum_{i=1}^{n} log(\frac{1}{v_1^{p/2}} e^{-\frac{1}{2v_1}||x_i - \mu_1||_2^2} + \frac{1}{v_2^{p/2}} e^{-\frac{1}{2v_2}||x_i - \mu_2||_2^2})$$

$$L_{gmm,s}(\{\mu_1, \mu_2\}; X) \approx -\sum_{i=1}^{n} \frac{p}{2} log(v_{c_i}) + \frac{1}{2v_{c_i}}||x_i - \mu_{c_i}||_2^2$$

3.3 Unconstrained τ_i, Σ_i

$$L_{gmm,s}(\{\mu_1, \mu_2\}; X) = \sum_{i=1}^{n} log(\frac{\tau_1}{\sqrt{|\Sigma_1|}} e^{-\frac{1}{2}(x_i - \mu_1)^T \Sigma_1^{-1}(x_i - \mu_1)} + \frac{\tau_1}{\sqrt{|\Sigma_2|}} e^{-\frac{1}{2}(x_i - \mu_2)^T \Sigma_2^{-1}(x_i - \mu_2)})$$

$$L_{gmm,s}(\{\mu_1, \mu_2\}; X) \approx \sum_{i=1}^{n} log(\tau_{c_i}) - \frac{1}{2} log(|\Sigma_{c_i}|) - \frac{1}{2}(x_i - \mu_{c_i})^T \Sigma_{c_i}^{-1}(x_i - \mu_{c_i})$$

4 Algorithms and Performance on Simple Synthetic Data

The most popular algorithm for solving the problem presented in 2 is Lloyd's algorithm. Lloyd's algorithm consists of repeating two simple steps:

1. Partition datapoints according to the closest cluster centers
2. Recalculate cluster centers as the mean of the corresponding datapoints

This algorithm is guaranteed to converge, and each iteration decreases the k means cost function (1). However it is not guaranteed to converge to a global minima of the k means problem.

Lloyd's algorithm can be extended to the cases presented in 3.

1. Distance to mean is no longer the only component in the approximate log likelihood, so points should be assigned to the cluster that minimizes the appropriate log likelihood term
2. Besides just keeping track of cluster means, the algorithm may need to compute the cluster proportions, and/or variance terms at each iteration.

Below we create simple synthetic datasets that correspond to each of the cases in 3 and compare the Lloyd's algorithm for k-means to the appropriately extended version of the algorithm.

4.1 Unconstrained τ_i

- $\tau_1 = 0.01$, $\tau_2 = 0.99$
- $\mu_1 = [0, 0]$, $\mu_2 = [10, 0]$
- $\Sigma_1 = \Sigma_2 = I$

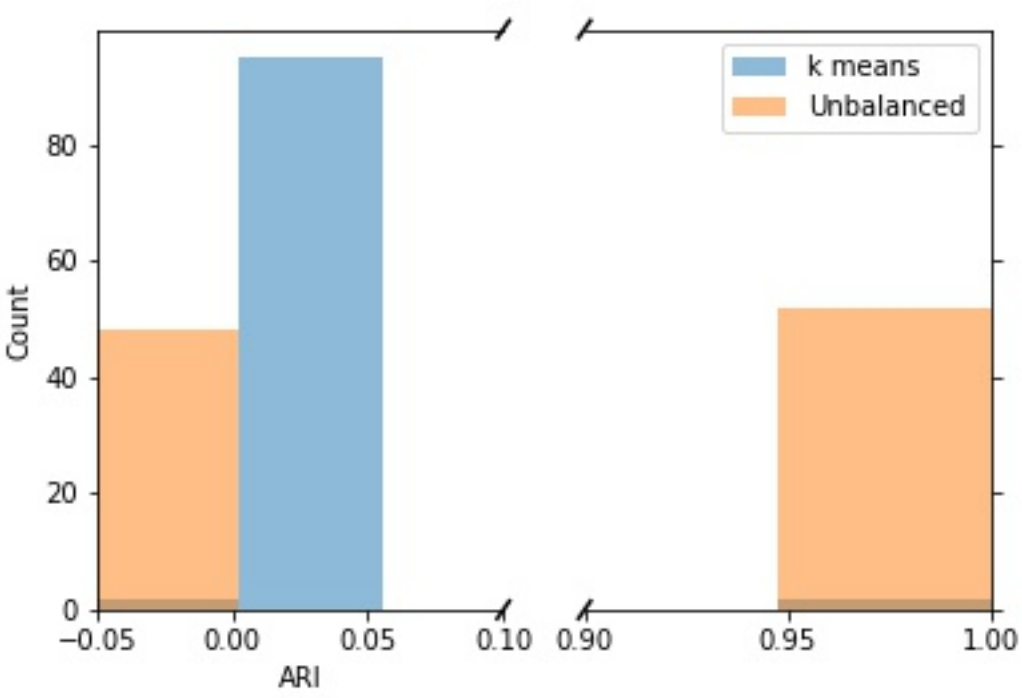

Figure 1: ARI histograms of k-means and unbalanced algorithms