k-means and em algorithm

Rebecca C. Steorts

Overview of K-means and EM Algorithm

K-means Clustering

The objective of the K-means algorithm is to minimize the sum of squared distances between data points and their respective cluster centroids. The procedure consists of two main steps:

- **Assignment Step:** Assign each data point to the nearest centroid.
- **Update Step:** Recalculate the centroids as the mean of all points assigned to each centroid.

The K-means algorithm minimizes the following cost function:

$$J(\mu) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

where μ_j is the centroid of cluster C_j.

Expectation-Maximization (EM) Algorithm

The EM algorithm aims to find maximum likelihood estimates of parameters in probabilistic models, especially when they depend on latent variables. The two main steps are:

- **E-step:** Compute the expected value of the log-likelihood function, given the current parameters.
- **M-step:** Maximize the expected log-likelihood from the E-step to update parameters.

For a Gaussian Mixture Model (GMM), the E-step assigns probabilities to latent variables, while the M-step updates the parameters based on these assignments.

Relationship Between K-means and EM

To connect K-means to the EM algorithm, we consider a specific probabilistic model: K-means as a GMM with certain assumptions.

Assume we have a GMM where each component (cluster) is a spherical Gaussian with the same variance σ^2 (which we can let tend towards zero). The Gaussian for each cluster is described as:

$$p(x|\mu_j) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|x - \mu_j\|^2}{2\sigma^2}\right)$$

Each data point x_i is assumed to belong to one of k clusters, with a latent variable z_i indicating the cluster assignment.

Proof Steps

E-step

The E-step computes the responsibilities (probabilities of membership in each cluster):

$$\gamma_{ij} = P(z_i = j | x_i, \mu) = \frac{p(x_i|\mu_j)P(z_i = j)}{\sum_{l=1}^{k} p(x_i|\mu_l)P(z_i = l)}$$

In the K-means context, if we set $P(z_i = j)$ to be equal across clusters (uniform prior), we have:

$$\gamma_{ij} = \begin{cases} 1 & \text{if } j = \text{argmin}_l \|x_i - \mu_l\|^2 \\ 0 & \text{otherwise} \end{cases}$$

Thus, in K-means, we have hard assignments instead of soft assignments.

M-step

In the M-step, K-means updates the centroids as:

$$\mu_j = \frac{\sum_{i:z_i=j} x_i}{N_j}$$

This is equivalent to maximizing the likelihood of the data given the current assignments, similar to how we would update the means of a Gaussian in a GMM.

Limiting Case

As the variance σ^2 tends to zero, the Gaussian distributions become increasingly peaked around the centroids, leading to hard assignments. In the limit, the probabilistic assignment of K-means approaches the hard assignment.

Conclusion

Thus, we conclude that K-means can be seen as a specific case of the EM algorithm applied to a GMM with spherical Gaussian components and uniform priors, particularly when the variance of the Gaussians approaches zero. This demonstrates that K-means clustering is indeed a limiting case of the EM algorithm.