Sakshi Jain(B20ME065)

Spectral Relaxation for K-means Clustering

By - Hongyuan Zha & Xiaofeng He,
Chris Ding & Horst Simon, Ming Gu

In Kmeans, clusters are represented by the center of mass of their members. It uses an algorithm in which each data point is assigned to the nearest cluster center and then computing the centroid of its member data vectors and assigning that point as the cluster center. Another approach is to minimize the sum-of-squares cost function using the coordinate descent method. Minimizing the cost function is equivalent to trace maximization formulation and assigning the cluster membership using pivoted QR decomposition, by taking into account the special structure of the partial eigenvector matrix. Then taking the performance of the clustering algorithms using document clustering as an example.

The methods for clustering, used in this algorithm are Kmeans clustering, sum-of-square cost function, QR decomposition, Coordinate descent method, and all these methods have their advantages and disadvantages, that is, pros and cons in this approach of clustering.

The advantage of doing clustering using Kmeans is that it is relatively simple to implement and it scales to large data sets. It has the guarantee to converge to a point and in this, we have the power to start the positions of centroid and then can be generalized to clusters of different shapes, sizes, such as elliptical clusters. It applies to unlabeled data sets and for linearly separable data and this algorithm works quickly, that is, the runtime of the algorithm is low because it has linear time complexity and it can be used with large datasets conveniently. K-Means produces tighter clusters than hierarchical clustering, especially if the clusters are globular. K-Means returns clusters that can be easily interpreted and even visualized, which is a big advantage of Kmeans clustering. K-Means inertia sum of squared means for each point to their respective cluster center (centroid). Higher inertia values can be helpful to question cluster numbers or algorithm's inner workings such as initialization or maximum iteration.

The advantage of this approach is that it uses a sum-of-square cost function which is easy to understand and calculate and the calculations are very fast and these are primarily used for simple, pure data. The method gives optimal estimates of the unknown parameters, it is very sensitive to the presence of unusual data points in the data used to fit a model. The expression for the gradient becomes prettier with the 1/2 because the 2 from the square term cancels out.

The advantage of using QR decomposition is the ease of implementation, which makes it a useful algorithm to use for prototyping if a pre-built linear algebra library is unavailable, which is rarely the case. One of the important advantages of Coordinate Descent is that it is well suited for parallel computation, in simple words, one can perform a full cycle of coordinate descent iterations in p parallel steps (as opposed to n), assuming the availability of a sufficient number of parallel processors. A second advantage of the coordinate descent method lies in the fact that it can be very useful in cases where the actual gradient of the function is not known. The coordinate descent method generally has similar convergence properties to the steepest descent.

The disadvantage of doing the clustering using Kmeans is that in this one needs to choose the value of K manually and it is being dependent on the initial value, which is being decided by a person by a random variable, that is the results will differ based on random centroid initialization. The clustering data varies sizes and density and outliers and one needs to scale it depending on the number of dimensions. With a global cluster, it didn't work well. The final result of clustering depends on the order of the data, initially given. The clustering is sensitive to scale, that is, the result will change if the data is normalized or standardized, which is not a bad property but one needs to spend some extra attention to scale the data, which might result in bad accuracy. It may work poorly with clusters with different densities but spherical shapes.

The disadvantage of this approach is that it uses a sum-of-square cost function which requires isolated spectral bands that are solely related to the constituent of interest and it cannot be used for complex mixture samples in which the individual constituents have overlapping spectral bands. In this approach, band selection can be difficult or impossible if the spectrum of property of interest is not known explicitly and large prediction errors will result from constituents with bands in the same region of the spectrum as the calibration band. This method is not suitable for business, and economic data that conform to the growth curves like Gompertz's curve, Logistic Pearl-Read curve, etc. It can be quite sensitive to the choice of starting values. It tends to overfit data.

Despite the popularity of Kmeans clustering, one of its major drawbacks is that the coordinate descent search method is prone to local minima. The coordinate descent method may not reach the local minimum even for a convex function. The algorithm may get stuck at a non-stationary point if the level curves of a function are not smooth. Coordinate descent performs redundant computation for the same training set for large datasets. If we take the entire dataset for computation, then we can update the weights of the model for the new data. As weights are updated frequently, the cost function fluctuates heavily and the loss is computed for each mini-batch and hence total loss needs to be accumulated across all mini-batches.