Comparison of Some Clustering Methods

Eric Chi

5/21/2018

Please write a short report (10 pages maximum, 11 point font, 1 inch margins all around) comparing the following four clustering methods: k-means++ [1], k-harmonic means [2], mean shift [3], and Gaussian mixture modelling estimated using the EM algorithm, in terms of (a) algorithmic implementation; (b) selection of tuning parameters if applicable; (c) practical performance in a simulation study.

[1] Arthur and Vassilvitskii (2007), "k-means++: the advantages of careful seeding," Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

[2] Zhang, Hsu, Meichun, and Umeshwar (1999), "K-Harmonic Means - A Data Clustering Algorithm," HP Labs Techreport.

[3] Fukunaga and Hostetler (1975), "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," IEEE Transactions on Information Theory, 21, 32-40.

You should think of your report as a well crafted tutorial that provides guidance to the reader on the pros and cons of the different methods. When describing the computations involved, you must include a flop count per iteration. At least five simulations must be included. Four of these simulation scenarios must highlight when each of the four methods is the method of choice. A fifth should be a comparison of wall clock run times as the number of data points and the number of features accompanying each observation increases.

To evaluate clustering quality, please use all the measures included in the clues package, the variation of information (available in the mcclust package), as well as the normalized mutual information (available in the NMI package).

Please include a brief discussion on how to interpret the values of these measures, and the pros and cons of each measure.

Please also include a brief discussion on how to choose the number of clusters and compare two methods for choosing the number of clusters in the literature (your choice) in the simulations section.

You may use R packages if available; however, you must review the code to ensure that the algorithmic steps explained in your report match those in the R code provided. These methods are not new and some comparisons, as required in this exam, have already appeared in the literature and in online sources. Your report will be evaluated in terms of the value added by your analysis. You should include a section in your report entitled "Recommendations" which summarizes the strengths and weaknesses of each method and advise the reader on which method is best suited for which scenario. Be sure to cite your sources generously. References do not count against the 10 pages maximum.