Learning K in K-Means

Greg Hamerly, Charles Elkan
Department of Computer Science and Engineering
University of California, San Diego

Abstract

The K-means Clustering despite its popularity has two major shortcomings :

- It scales poor computationally
- Parameter k has to be provided which is not obvious and and is a hard algorithmic problem.

This paper deals with the second problem and provide a algorithm to predict k using G-means.

Related Works

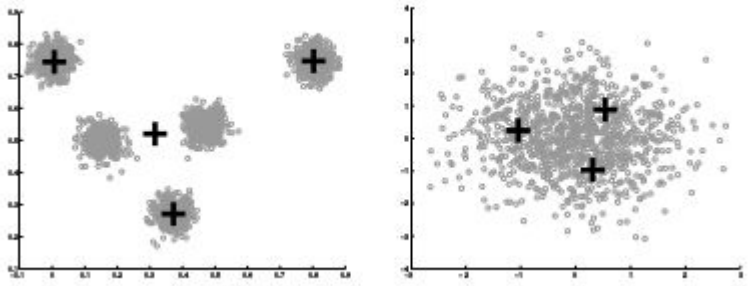
- Clustering algorithms are useful tools for data mining, compression, probability density estimation.
- Most clustering algorithms require the user to specify the number of clusters (called k), and it is not always clear what is the best value for k.
- Choosing k is often an ad hoc decision based on prior knowledge, assumptions, and practical experience.

Assumptions ….

Assumption involved in center-based clustering is :

Center-based clustering algorithms (in particular k-means and Gaussian expectation-maximization) **usually assume that each cluster adheres to a unimodal distribution, such as Gaussian**. With these methods, only one center should be used to model each subset of data that follows a unimodal distribution.

If multiple centers are used to describe data drawn from one mode, the centers are a needlessly complex description of the data, and in fact the multiple centers capture the truth about the subset less well than one center.



Previous Algorithms

Several algorithms have been proposed previously to determine k automatically

- Pelleg and Moore proposed a regularization framework for learning k, which they call X-means. The algorithm searches over many values of k and scores each clustering model using the so-called Bayesian Information Criterion.
- Bischof used a minimum description length (MDL) framework, where the description length is a measure of how well the data are fit by the model.

- One is to build a merging tree of the data based on a cluster distance metric, and search for areas of the tree that are stable with respect to inter- and intra-cluster distances.

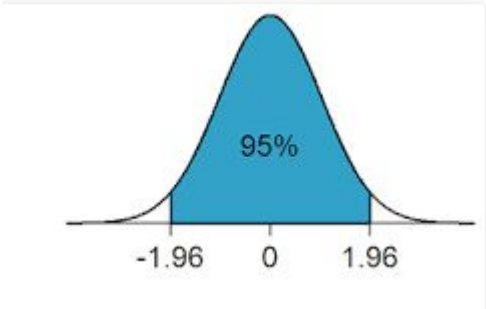
The Gaussian-means (G-means) algorithm

- The G-means algorithm starts with a small number of k-means centers, and grows the number of centers.
- Each iteration of the algorithm splits into two those centers whose data appear not to come from a Gaussian distribution.
- Between each round of splitting, we run k-means on the entire dataset and all the centers to refine the current solution.
- We can initialize with just k = 1, or we can choose some larger value of k if we have some prior knowledge about the range of k.

Statistical Test in G-means

Two key advantages of the hypothesis test is that

- It does not limit the covariance of the data and does not compute a full covariance matrix.
- Additionally, G-means only requires one intuitive parameter, the standard statistical significance level α.



Algorithm

- **1: Let C be the initial set of centers (usually C ← {$\bar{x}$}).**
- **C ← kmeans(C, X).**
- **Let {x(i) |class(x (i)) = j} be the set of data points assigned to center c j .**
- **Use a statistical test to detect if each {x (i) |class(x (i)) = j} follow a Gaussian distribution (at confidence level α).**
- **If the data look Gaussian, keep c (j) . Otherwise replace c (j) with two centers.**
- **Repeat from step 2 until no more centers are added.**

Algorithm Description….

- G-means repeatedly makes decisions based on a statistical test for the data assigned to each center.
- If the data currently assigned to a k-means center appear to be Gaussian, then we want to represent that data with only one center.
- However, if the same data do not appear to be Gaussian, then we want to use multiple centers to model the data properly.
- The algorithm will run k-means multiple times (up to k times when finding k centers), so the time complexity is at most O(k) times that of k-means.

G-Means Intricacies..

- The k-means algorithm implicitly assumes that the datapoints in each cluster are spherically distributed around the center.
- The Gaussian distribution test that the paper will present are valid for either covariance matrix assumption.
- The test also accounts for the number of datapoints n tested by incorporating n in the calculation of the critical value of the test . This prevents the G-means algorithm from making bad decisions about clusters with few datapoints.

Anderson-Darling statistic Test

To specify the G-means algorithm fully we need a test to detect whether the data assigned to a center are sampled from a Gaussian. The alternative hypotheses are

- H 0 : The data around the center are sampled from a Gaussian.
- H 1 : The data around the center are not sampled from a Gaussian.

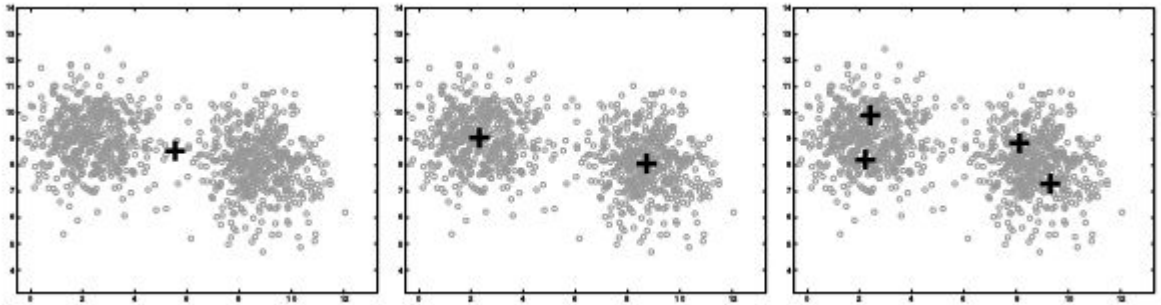
If we **accept the null hypothesis** H 0 , then we believe that the **one center is sufficient to model its data**, and we should not split the cluster into two sub-clusters. If we **reject H 0** and accept H 1 , then we **want to split the cluster.**

Hypothesis Test ...

- Choose a significance level α for the test.
- Initialize two centers, called “children” of c.
- Run k-means on these two centers in X. This can be run to completion, or to some early stopping point if desired. Let c1 , c2 be the child centers chosen by k-means.
- Let v = c 1 − c 2 be a d-dimensional vector that connects the two centers. This is the direction that k-means believes to be important for clustering. Then project X onto v: $x'_i = \langle x_i, v \rangle / ||v||^2$. .x’ is a 1-dimensional representation of the data projected onto v. Transform x’ so that it has mean 0 and variance 1.

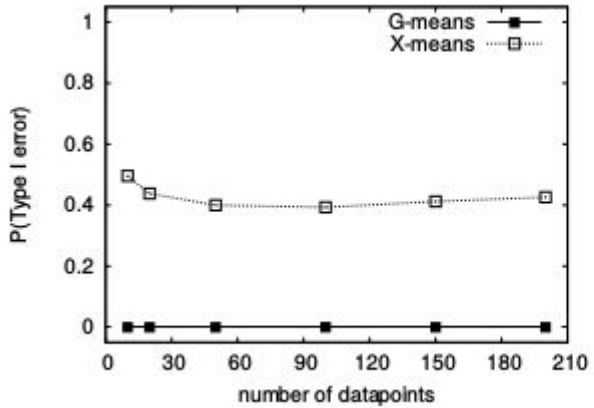
An Example

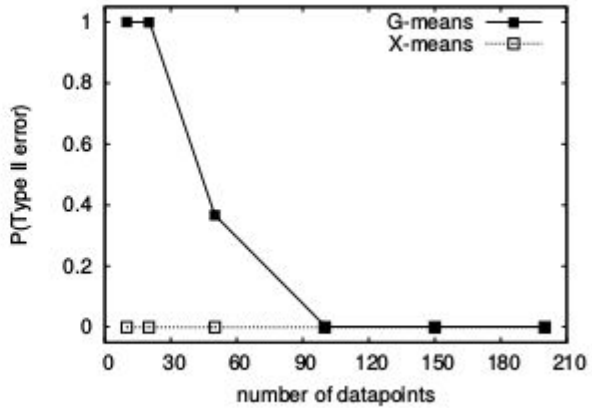
An example of running G-means for three iterations on a 2d dataset with two true clusters and 1000 points. Starting with one center (left plot), G-means splits into two centers (middle). The test for normality is significant, so G-means rejects H 0 and keeps the split. After splitting each center again (right), the test values are not significant, so G-means accepts H 0 for both tests and does not accept these splits. The middle plot is the G-means answer.



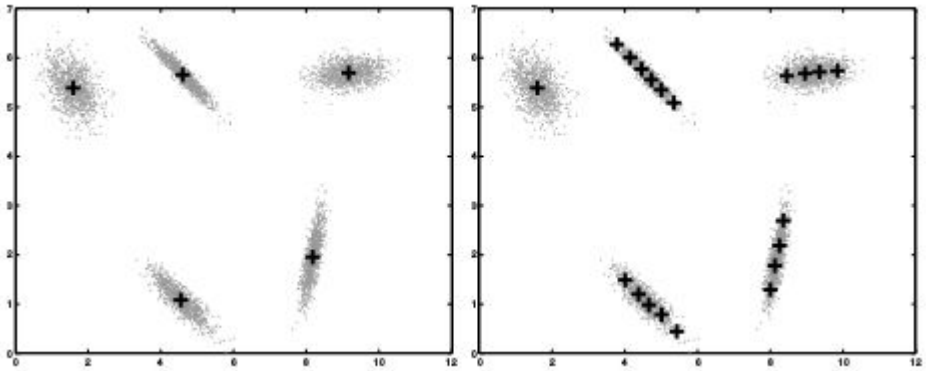
Statistical power

A comparison of the power of the Anderson-Darling test versus the BIC. For the AD test we fix the significance level (α = 0.0001)





2-d synthetic dataset with 5 true clusters. On the left, G-means correctly chooses 5 centers and deals well with non-spherical data. On the right, the BIC causes X-means to overfit the data, choosing 20 unevenly distributed clusters.



Conclusions

- The new G-means algorithm for learning K, uses dimension reduction and a powerful test for Gaussian fitness. G-means uses this statistical test to discover the number of clusters automatically.
- The only parameter supplied to the algorithm is the significance level of the statistical test.
- The G-means algorithm takes linear time and space in the number of datapoints and dimension, since k-means is itself linear in time and space.