Gaussian Similarity Clustering*

James Ting-Ho Lo
Department of Mathematics and Statistics
University of Maryland Baltimore County
Baltimore, MD 21250, U.S.A.
Email: jameslo@umbc.edu

Abstract

Key words: clustering; gradual deconvexification; local minima; Gaussian similarity; prototype; nonlinear regression; generalization; regularization.

1 Introduction

One of the most widely used methods of finding isotropic clusters in a dataset is the K-means algorithm (and its variants). This unsupervised hard clustering method is essentially a gradient-decent minimization procedure, which begins with an initial set of K cluster-centers and iteratively updates the set so as to decrease an error function (e.g., sum of squared errors). The complexity of an minimization iteration of the K-means algorithm applied on a sample size of m instances, each being characterized by N attributes, is O(K*m*N). This complexity, which is linear in K, m and N, is one of the reasons for the popularity of the K-means type of clusterin algorithm, because K, m and N are usually very large nowaday. Convergence of the K-means algorithm in finite iterations is proven in (Selim and Ismail, 1984). Low computational cost is undoubtedly the most attractive advantage of the K-means algorithm in comparison to other clustering methods, which are usually of non-linear complexity. Other advantages of the algorithm include its intuitive interpretation, straightforward implementation, and fast convergence.

However, the K-means algorithm has the following shortcomings:

1. It cannot determine the number K of clusters except by repeatedly using the algorithm for different values of K. It is usually difficulty to estimate from the nature of the dataset in the application.
2. It is sensitive to the selection of the K initial partition or prototypes/centroids, because the algorithm may get trapped into a nonglobal local minimum of the objective criterion.
3. It is sensitive to noises and outliers, which can increase the squared error dramatically.
4. The use of it is often limited to numerical attributes. It is applicable only when the numerical mean is defined.
5. In general, the less isotropical the clusters in the dataset are, the more its clustering performance suffers.

Remedies of these shortcomings have been proposed: Haung (1998) presented the K-prototypes algorithm, which is based on the K-means algorithm but removes numerical data limitations while preserving its efficiency. The algorithm clusters objects with numerical and categorical attributes in a way similar to the K-means algorithm. The similarity measure on numerical attributes is the square Euclidean distance;

*This material is based upon work supported in part by the National Science Foundation under Grant ECCS1508880, but does not necessarily reflect the position or policy of the Government.

the similarity measure on the categorical attributes is the number of mismatches between objects and the cluster prototypes.

A clustering algorithm, that attempts to reduce the sensitivity of the K-means algorithm toward noises and outliers is the K-medoids or PAM (partition around medoids – (Kaufmann and Rousseeuw, 1987)). It is similar to the K-means algorithm, but differs from the K-means in its representation of the different clusters. Each cluster is represented by the most centric object in the cluster, rather than by the mean that may not belong to the cluster. Another clustering algorithm, that attempts to reduce the sensitivity of the K-means algorithm toward noises and outliers, is obtained by using the sum of absolute errors instead of SSE as the minimization criterion (Estivill-Castro, 2000). Again, it requires more computation than the K-means algorithm.

This paper proposes yet another clustering algorithm that remedies shortcomings 1, 2, 3 and 5 listed above. Shortcoming 4 can be coped with for the proposed algorithm by using the same idea as that in Haung (1998). This is discussed in a forthcoming paper. The proposed algorithm determines the number K of clusters in the process, alleviates the local-minimum problem, reduces or eliminates the effects of noises and outliers, and can find elliptical clusters. The proposed algorithm is based on a novel similarity measure called the Gaussian similarity, and thus the new algorithm is called the Gaussian similarity clustering (GSC) or algorithm. As compared with the K-means algorithm, GSC converges as fast and lends intuitive interpretation and not much harder implementation, but each of its iteration involves more computation. Fortunately, the computation in an iteration can be greatly reduced by mathematical analysis and numerical consideration of the Gaussian similarity.

GSC is related with a method of convexifying the sum squared error (SSE) for avoiding its nonglobal local minima in data fitting and the method's successful application to training neural networks were reported in [3, 4, 2, 5, 6, 7]. The method transforms the SSE into a risk-averting error (RAE) by applying the operator $\exp(\lambda(\cdot))$ to each summand of the SSE, where the risk sensitivity index λ is a positive-valued variable. This RAE is inspired by but different from the error function with the same name that is used in deriving robust controllers and filters [1, 9, 10].

It is appropriate to mention here that GSC is a relatively inexpensive way to find a Gaussian mixture model of the dataset [**?**, **?**, **?**].

2 Sum of Gaussian Similarities (SGS) in an Isotropic Gaussian Cluster

Roughly speaking, for an isotropic cluster, the similarity measure used should be large between two data points with a small Enclidean distance and drops off quickly as the Euclidean distance increases beyond the "radius" of the "spherical" cluster. Gaussian similarity is such a similarity measure, which based on the Gaussian density function. For a given standard deviation (sd) σ, the Gaussian similarity between the data points, vectors x and y, is the value of the Gaussian density function $g(x, y, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\|x-y\|^2}{2\sigma^2}}$, where $\|\cdot\|$ is the Euclidean norm. The similarity measure $g(x, y, \sigma)$ is called the Gaussian similarity relative to the standare deviation (sd) σ or variance σ^2. Note that the smaller $\|x - y\|/\sigma$ is (or the closer x and y are relative to the scale σ), the greater the Gaussian similarity between x and y is and drops off quickly as $\|x - y\|$ increases to more than 2σ. For example, the Gaussian similarity $g(x, y, \sigma)$ is 0.39894σ, 0.24197σ, 0.05399σ, 0.004432σ, 0.000134σ, 0.0000015σ for $\|x - y\| = 0$, σ, 2σ, 3σ, 4σ, 5σ, respectively.

Let $X = \{x_k, k = 1, \ldots, K\} \subset R^N$ be a given set of data points and p be a prototype for X, the data points x_k and the prototype p being a vector from the Euclidean space R^N. The sum of Gaussian similarities (SGS) of all the data points $x_k \in X$ to p is

$$\begin{aligned} G(p, \sigma) &= \sum_{k=1}^{K} g(x_k, p, \sigma) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=1}^{K} e^{-\frac{\|x_k - p\|^2}{2\sigma^2}} \end{aligned} \tag{1}$$

As mentioned above, the Gaussian similarity drops off quickly as $\|x_k - p\|$ increases to more than 2σ and approaches 0 as $\|x_k - p\|$ increases to more than 3σ. This de-emphasizes the effect of data points more than 3σ away from p on SGS. Starting with a given p, we move it by minimizing (**??**) to $p^* = \arg\min_p G(p,\sigma)$. Those data points that are more than 3σ away from p^* are supposed to belong to a different cluster. This way, the prototype p would be at the "center" of the cluster without being pulled unduely by outliers or data points from other clusters. The size of the cluster "centered" at p^* depends on σ, which is to be selected to define the cluster separated from others.

It is proven that the convexity region of $G(p,\sigma)$ expands monotonically as σ increases. To make advantage of larger convexity regions, $G(p,\sigma)$ is started at a very small σ, which is gradually increased. If the cluster is isotropic, σ is increased to about 1/3 to 2/5 of the "radius" of the isotropic cluster, which is estimated in the iterative process of minimizing $G(p,\sigma)$. Equally important, throughout the process of minimizing $G(p,\sigma)$ and increasing σ, only those data points within the current 4σ to 5σ (Euclidean distance) of the current p need to be taken into consideration or computation, greatly saving the computation required.

3 Normalized Sum of Gaussian Similarities (NSGS) in an Isotropic Gaussian Cluster

For σ sufficiently small or $\|x_k - p\|^2$ sufficiently large, the computation of $e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}$ in $g(x_k,p,\sigma)$ incurs computer (or arithmetic) underflow. To avoid such a problem, we use the following normalized mean of Gaussian similarities (NSGS)

$$\begin{aligned} C(p,\sigma) &= -2\sigma^2 \ln\left(\frac{\sigma\sqrt{2\pi}}{K} G(p,\sigma)\right) \\ &= -2\sigma^2 \ln\left(\frac{1}{K}\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}\right) \end{aligned}$$

and evaluate $C(p,\sigma)$ as follows: For the vector p, let the set $S(p,\sigma) = \arg\max_{k\in\{1,...,K\}} g(x_k,p,\sigma)$, which may contain more than one element if a tie exists, and $M(p,\sigma) = \min_k \{k | k \in S(p,\sigma)\}$ which is the smallest index k among all values in $S(p,\sigma)$. It follows that for all $k \in \{1,...,K\}$,

$$\begin{aligned} g(x_k,p,\sigma) &\le g\left(x_{M(p,\sigma)},p,\sigma\right) \\ \|x_k - p\| &\ge \left\|x_{M(p,\sigma)} - p\right\| \end{aligned}$$

Defining the symbol

$$\eta_k(p) := e^{-\frac{\|x_k-p\|^2 - \left\|x_{M(p,\sigma)}-p\right\|}{2\sigma^2}} \tag{2}$$

we have, by straightforward calculation,

$$\begin{aligned} C(p,\sigma) &= -2\sigma^2 \ln\left(\frac{1}{K}\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}\right) \\ &= -2\sigma^2 \ln\left(\frac{1}{K} e^{-\frac{\left\|x_{M(p,\sigma)}-p\right\|^2}{2\sigma^2}} \sum_k e^{-\frac{\|x_k-p\|^2 - \left\|x_{M(p,\sigma)}-p\right\|^2}{2\sigma^2}}\right) \\ &= -2\sigma^2 \ln\left(\frac{1}{K} e^{-\frac{\left\|x_{M(p,\sigma)}-p\right\|^2}{2\sigma^2}} \sum_{k=1}^{K} \eta_k(p)\right) \\ &= -2\sigma^2 \ln\frac{1}{K} + \left\|x_{M(p,\sigma)} - p\right\|^2 - 2\sigma^2 \ln\left(\sum_{k=1}^{K} \eta_k(p)\right) \end{aligned} \tag{3}$$

Hence the NSGS $C(p,\sigma)$ is computed without computer underflow whatever value $\sigma > 0$ is.

4 Gradient and Hessian Matrices of SGS and NSGS in an Isotropic Gaussian Cluster

Recall the sum of Gaussian similarities (SGS) of p to all the data points $x_k \in X$:

$$G(p,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \sum_k e^{-\frac{\|x_k - p\|^2}{2\sigma^2}} \tag{4}$$

The derivative of $G(p,\sigma)$ with respect to p_j is

$$\begin{aligned}
\frac{\partial G(p,\sigma)}{\partial p_j} &= \frac{1}{\sigma\sqrt{2\pi}} \sum_k e^{-\frac{\|x_k - p\|^2}{2\sigma^2}} \left(\frac{1}{-2\sigma^2}\right) (-2(x_{kj} - p_j)) \\
&= \frac{1}{\sigma^3\sqrt{2\pi}} \sum_k e^{-\frac{\|x_k - p\|^2}{2\sigma^2}} (x_{kj} - p_j)
\end{aligned}$$

The second-order derivative of $G(p,\sigma)$ with respect to p_j and p_i is

$$\begin{aligned}
\frac{\partial^2 G(p,\sigma)}{\partial p_i \partial p_j} &= \frac{1}{\sigma^3\sqrt{2\pi}} \sum_k e^{-\frac{\|x_k - p\|^2}{2\sigma^2}} (x_{kj} - p_j) \left(\frac{1}{-2\sigma^2}\right) (-2(x_{ki} - p_i)) \\
&+ \frac{1}{\sigma^3\sqrt{2\pi}} \sum_k e^{-\frac{\|x_k - p\|^2}{2\sigma^2}} (-\delta_{ij}) \\
&= \frac{1}{\sigma^3\sqrt{2\pi}} \sum_k e^{-\frac{\|x_k - p\|^2}{2\sigma^2}} \left(\frac{(x_{ki} - p_i)(x_{kj} - p_j)}{\sigma^2} - \delta_{ij}\right)
\end{aligned}$$

Note that the computation of both the first- and second-order derivatives incur computer underflow. We show how derivatives of the NSGS $C(p,\sigma)$ can be computed without computer underflow in the following: Straightforward calculation yields

$$\begin{aligned}
\frac{\partial C(p,\sigma)}{\partial p_j} &= -2\sigma^2 \frac{\partial}{\partial p_j} \left(\ln \frac{1}{K} \sum_k e^{-\frac{\|x_k - p\|^2}{2\sigma^2}}\right) \\
&= \frac{-2\sigma^2}{\frac{1}{K}\sum_k e^{-\frac{\|x_k - p\|^2}{2\sigma^2}}} \frac{1}{K} \sum_k e^{-\frac{\|x_k - p\|^2}{2\sigma^2}} \left(\frac{1}{-2\sigma^2}\right) (-2(x_{kj} - p_j)) \\
&= \frac{-2\sum_k e^{-\frac{\|x_k - p\|^2 - \left\|x_{M(p,\sigma)} - p\right\|}{2\sigma^2}} (x_{kj} - p_j)}{\sum_k e^{-\frac{\|x_k - p\|^2 - \left\|x_{M(p,\sigma)} - p\right\|}{2\sigma^2}}} \\
&= \frac{-2\sum_k \eta_k(p)(x_{kj} - p_j)}{\sum_k \eta_k(p)}
\end{aligned} \tag{5}$$

which shows that evaluation of the gradient vector $\left[\frac{\partial C(p,\sigma)}{\partial p_j}\right]$ does not incur computer underflow. By the same idea, evaluation of the Hessian matrix $\left[\frac{\partial^2 C(p,\sigma)}{\partial p_i \partial p_j}\right]$ can also be evaluated without incurring computer

underflow.

$$\begin{aligned}
\frac{\partial^2 C(p,\sigma)}{\partial p_i \partial p_j} &= \frac{\partial}{\partial p_i}\frac{-2\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}(x_{kj}-p_j)}{\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}} \\
&= -2\frac{\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}\left[\left(\frac{-2}{-2\sigma^2}\right)(x_{ki}-p_i)(x_{kj}-p_j)-\delta_{ij}\right]}{\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}} \\
&- 2\left(\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}(x_{kj}-p_j)\right)\frac{\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}\left(\frac{-2}{-2\sigma^2}\right)(x_{ki}-p_i)}{\left(\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}\right)^2} \\
&= \frac{-2}{\sigma^2}\frac{\sum_k e^{-\frac{\|x_k-p\|^2-\left\|x_{M(p,\sigma)}-p\right\|^2}{2\sigma^2}}\left[(x_{ki}-p_i)(x_{kj}-p_j)-\delta_{ij}\right]}{\sum_k e^{-\frac{\|x_k-p\|^2-\left\|x_{M(p,\sigma)}-p\right\|^2}{2\sigma^2}}} \\
&- \frac{1}{2\sigma^2}\frac{\partial C(p,\sigma)}{\partial p_i}\frac{\partial C(p,\sigma)}{\partial p_j} \\
&= \frac{-2}{\sigma^2}\frac{\sum_k \eta_k(p)\left[(x_{ki}-p_i)(x_{kj}-p_j)-\delta_{ij}\right]}{\sum_k \eta_k(p)} - \frac{1}{2\sigma^2}\frac{\partial C(p,\sigma)}{\partial p_i}\frac{\partial C(p,\sigma)}{\partial p_j}
\end{aligned} \tag{6}$$

Limits of Normalized Mean of Gaussian Similarities

Because

$$\eta_k(p) \le 1$$
$$\ln\left(\sum_{k=1}^{K}\eta_k(p)\right) \le \ln K$$

it follows

$$\lim_{\sigma\to 0} C(p,\sigma) = \left\|x_{M(p.\sigma)}-p\right\|^2 \tag{7}$$

which shows that as σ tends to 0, $C(p,\sigma)$ converges to the minimum squared error.

Observing that

$$\begin{aligned}
\frac{1}{K}\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}} &= \frac{1}{K}\sum_{k=1}^{K}\left[1-\frac{\|x_k-p\|^2}{2\sigma^2}+O\left(\frac{1}{\sigma^4}\right)\right] \\
&= 1-\frac{1}{K}\sum_{k=1}^{K}\frac{\|x_k-p\|^2}{2\sigma^2}+O\left(\frac{1}{\sigma^4}\right)
\end{aligned}$$

Recalling the power expansion formula,

$$\ln(1+x) = x-\frac{1}{2}x^2+\frac{1}{3}x^3-\cdots, \text{ for } -1<x<1$$

we have, for λ sufficiently small,

$$\begin{aligned} C(p,\sigma) &= -2\sigma^2 \ln\left(\frac{1}{K}\sum_{k=1}^{K} e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}\right) \\ &= -2\sigma^2 \ln\left(1-\frac{1}{K}\sum_{k=1}^{K}\frac{\|x_k-p\|^2}{2\sigma^2}+O\left(\frac{1}{\sigma^4}\right)\right) \\ &= -2\sigma^2\left(-\frac{1}{K}\sum_{k=1}^{K}\frac{\|x_k-p\|^2}{2\sigma^2}+O\left(\frac{1}{\sigma^4}\right)\right) \\ &= \frac{1}{K}\sum_{k=1}^{K}\|x_k-p\|^2+O\left(\frac{1}{\sigma^2}\right) \end{aligned}$$

It follows that

$$\lim_{\sigma\to\infty} C(p,\sigma) = \frac{1}{K}\sum_{k=1}^{K}\|x_k-p\|^2 \tag{8}$$

which shows that as σ tends to ∞, $C(p,\sigma)$ converges to the mean squared error. Therefore, for σ large enough, $\arg\min_p C(p,\sigma) = \arg\min_p G(p,\sigma) \approx \arg\min_p \frac{1}{K}\sum_{k=1}^{K}\|x_k-p\|^2$.

5 Searching for an Isotropic Gaussian Cluster by Gradual Deconvexification

Recall that the K-means algorithm determines K modes of the sum of squared errors (or equivalently the K means, prototypes, or centroids) together in a minimization process. In contrast, GSC (Gaussian similarity clustering) searches for one mode (or prototype) at a time. If not much prior knowledge is available, as is usually the case, about the number and locations of the modes, a number of initial prototypes uniformly distributed in the space R^N of data points are first selected. It is better for the number to be larger than "expected". Instead of forcing K prototypes to represent K different clusters as by the K-means algorithm, GSC drives each initial prototype to a mode of the SGS to represent a cluster of data points that are the most similar to the initial prototype. It is possible for more than one initial prototype to be driven to the same mode, only one is kept to represent their common cluster.

In the event some data points are not included in the clusters obtained, additional initial prototypes can be selected to find additional clusters. An extremely small cluster so obtained may represent outliers. Since prototypes are placed separately, the prototypes that have been placed remain unchanged and are not involved in the computation for placing additional prototypes.

We note here that although we select a number of initial prototypes, they are adjusted one by one with their clusters formed. Whenever the adjustment of an initial prototype with the formation of its cluster is completed, the data points in the cluser are removed from the data sets.

Given a set $X = \{x_k, k = 1, \ldots, K\} \subset R^N$ of data points x_k, a cluster is found first by minimizing the Sum of Gaussian similarities (SGS) $G(p,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}}$ by the variation of p, starting with a selected initial prototype p_0. be a given set of data points and p be a prototype for X, the data points x_k and the prototype p being vectors from the Euclidean space R^N. The sum of Gaussian similarities (SGS) of p to all the data points $x_k \in X$ is

$$\begin{aligned} G(p,\sigma) &= \sum_k g(x_k,p,\sigma) \\ &= \frac{1}{\sigma\sqrt{2\pi}}\sum_k e^{-\frac{\|x_k-p\|^2}{2\sigma^2}} \end{aligned} \tag{9}$$

GSC (Gaussian similarity clustering) searches for a cluster by finding a prototype p that minimizes the SGS is represented by a prototype at the "center" of the cluster.

Before searching, we select a number of prototypes uniformly in an estimated region of the data points. These prototypes usually do not coincide with a data point. In order to move a prototype in the direction its most similar data point when σ is very small, we have to use (3) to compute $C(p, \sigma)$, (5) to compute the gradient vector of $C(p, \sigma)$, and if necessary, (6) to compute the Hessian matrix of $C(p, \sigma)$.

Moving each of the uniformly distributed initial prototypes into the "center" of its closest cluster at a succession of increasing values of σ^2 yields yields candidates of the prototypes that represent those clusters closest to the initial prototypes. Multiple prototypes may move into the same "center" of a cluster. Only a single one is kept to represent the cluster. If there are data points more than 3σ away from every obtained "center", more initial prototypes are created near them. In this manner, a guess of the number of prototypes is needed no longer.

It is proven that the convexity region of $G\left(p, \sigma^2\right)$ expands monotonically as the variance σ^2 decreases. The limit of the convexity region of $G\left(p, \sigma^2\right)$ is the entire prototype space (or data point space). For a very small σ^2, each prototype stays at the center of a very small set of data points. When σ^2 gradually increases, the cluster with p as the center becomes larger and larger in process of maximizing $G\left(p, \sigma^2\right)$. By eliminating multiple prototypes at a center of a cluster, we make sure that there is a prototype at each center. If not all the data points are within 3σ of a prototype, additional prototypes can be provided among such data points. If there are very few data points within 3σ of a prototype, we may consider dropping the prototype or the cluster.

In determining the position of prototype p, the process of increasing σ^2 stops whenever increasing σ^2 and performing maximization at the increased value of σ^2 does not increase $G\left(p, \sigma^2\right)$ much for a certain number of times, indicating no additional data point is close enough to p between the begining value and ending value of σ^2 over said certain number of times. The prototype is said to be mature and recorded.

6 Kernel Density Estimation for Clustering

The idea of using kernel density estimation for clustering is motivated by the following thought: Let the data points in each of a number of clusters be obtained by sampling from a certain distribution, any distribution. Assume that the density of the distribution of each cluster is known. Under this assumption, the clustering problem can be easily solved perfectly using the densities.

The idea is to estimate the density of the **entire** data set first, and then cluster the data set using the density estimate. A popular approach to estimating the density is kernel density estimation, which has been much studied. A good introduction to kernel density estimation can be found at https://en.wikipedia.org/wiki/Kernel_density_estim

If the kernel used is the Gaussian kernel, kernel density estimation is related to Gaussian similarity clustering (GSC). Hopefully, we can find synergism or cross-fertilization between GSC and kernel density estimation.

If the band width in the Gaussian kernel is large enough for the density estimate to be smooth enough, we can look for local minima of the density estimate and use them as the prototypes after proper combination of the local minima. On the other hand, GSC can help us in finding the local minima.

According to the article at https://en.wikipedia.org/wiki/Kernel_density_estimation, if the actual density function from which the data points of a cluster is Gaussian, the optimal choice of the bandwidth σ is rule of thumb $1.06sn^{-1/5}$ (for single variate data) or $0.9an^{-1/5}$ (for multivariate data), where s is the standard deviation of the data points in the cluster, $a = \min\left(s, \frac{IQR}{1.34}\right)$, IQR is the interquartile range (i.e., the 75th percentile minus the 25th percentile), and n is the number of data points in the cluster [**?**] [Silverman [1986], "Density estimation for statistics and data analysis"]. s and IQR are difficulty, if not impossible to estimate especially before the cluster is identified. Fortunately, the mode of the density estimate is not too sensitive to the bandwith as long as it is large enough for the density estimate to be unimodal, provided the data points from the true multivariate normal density are symmetric in each coordinate axis [8] [On Estimation of a Probability Density Function and Mode, Annals of Mathematical Statistics, Vol. 33, Issue 3, pp. 1065-1076, Sept. 1962].

GSC (Gaussian similarity clustering) algorithm can be looked upon as adaptively adjusting σ to obtain an estimate of the mode of the true Gaussian density of a cluster.

7 Sum of Gaussian Similarities (SGS) in a Nonisotropic Gaussian Cluster

The general Gaussian density $N(x,p,V)$ with mean vector p and covariance matrix V:

$$N(x,p,V)=\left(\sqrt{(2\pi)^N|V|}\right)^{-1}\exp\left(-\frac{1}{2}(x-p)^T V^{-1}(x-p)\right)$$

where its covariance matrix V can be estimated from the dataset S as follows:

$$\hat{V}=\sum_{x_k\in S}(x_k-p)(x_k-p)^T$$

A better estimate $\hat{V}^{new}$ of V can be obtained by iterating the following:

$$\hat{V}^{new}=\sum_{\left\{x_k:\left(\sqrt{(2\pi)^N|V^{old}|}\right)^{-1}\exp\left(-\frac{1}{2}(x_k-p)^T(V^{old})^{-1}(x_k-p)\right)>c\right\}}(x_k-p)(x_k-p)^T$$

where c is the number for which the probability of $\left\{x_k:\left(\sqrt{(2\pi)^N|V^{old}|}\right)^{-1}\exp\left(-\frac{1}{2}(x_k-p)^T\left(V^{old}\right)^{-1}(x_k-p)\right)>c\right\}$ is 99.73% or 95.45%, depending on how close the clusters are. A better estimate S^{new} can be obtained in accordance with $\hat{V}^{new}$.

With the obtained $\hat{V}$, the data point x_k is assigned to the cluster if $\left(\sqrt{(2\pi)^N|V|}\right)^{-1}\exp\left(-\frac{1}{2}(x_k-p)^T V^{-1}(x_k-p)\right)>c$, or equivalent $(x_k-p)^T V^{-1}(x_k-p)>-2\ln\left(c\sqrt{(2\pi)^N|V|}\right)$. Note that once V^{-1} and $-2\ln\left(c\sqrt{(2\pi)^N|V|}\right)$ are evaluated, the computation for checking the inequality is not too costly. In the event that x_k satisfies $(x_k-p)^T V^{-1}(x_k-p)>-2\ln\left(c\sqrt{(2\pi)^N|V|}\right)$ for more than one cluster (or prototype p), x_k is assigned to the cluster with the largest $(x_k-p)^T V^{-1}(x_k-p)$ (or p).

The sum of Gaussian similarities in a nonisotropic Gaussian cluster is

$$\begin{aligned}G(p,\sigma)&=\sum_k N(x_k,p,V)\\&=\left(\sqrt{(2\pi)^N|V|}\right)^{-1}\sum_k\exp\left(-\frac{1}{2}(x_k-p)^T V^{-1}(x_k-p)\right)\end{aligned}\tag{10}$$

As mentioned above, the Gaussian similarity drops off quickly as $\|x_k-p\|$ increases to more than 2σ and approaches 0 as $\|x_k-p\|$ increases to more than 3σ. This de-emphasizes the effect of data points that are 3σ away from p on MGS and the effect of outliers which are usually relatively farther away from a cluster.

To make advantage of larger convexity regions, $G(p,\sigma)$ is started at a very small σ, which is gradually increased to about 1/3 to 1/4 of the "radius" of the isotropic cluster, which is estimated in the process. This way, the prototype p would be at the "center" of the cluster without being pulled unduely by outliers or data points from other clusters. Equally important, after an initial stage in the process of minimizing $G(p,\sigma)$, only those data points within 4σ to 5σ (Euclidean distance) of the current p need to be taken into consideration or computation, greatly saving the computation required.

8 Gaussian Mixture Approach to Clustering Isotropic and Nonisotropic Clusters

The approach "probabilistic model-based clustering using a mixture of models" is related to GSC. It is computationally very expensive for a large data point dimension N. A large number of references can be found by googling "probabilistic model-based clustering using a mixture of models".

Decomposition of Superpositions of Distribution Functions – December 31, 1995, ISBN-10: 0306301199 ISBN-13: 978-0306301193
by Pal Medgyessy
Decomposition of Superpositions of Density functions and Discrete Distribution – 1977, ISBN 0 470 15017 3
by Pal Medgyessy
https://en.wikipedia.org/wiki/Mixture_model
http://projecteuclid.org/euclid.ssu/1272547280
http://smm.sagepub.com/content/5/2/107.abstract
Everitt, B.S.; Hand, D.J. (1981). Finite mixture distributions. Chapman & Hall. ISBN 0-412-22420-8
Titterington, D.; Smith, A.; Makov, U. (1985). Statistical Analysis of Finite Mixture Distributions. Wiley. ISBN 0-471-90763-4
McLachlan, G.J.; Peel, D. (2000). Finite Mixture Models. Wiley. ISBN 0-471-00626-2
Nielsen, Frank (23 March 2012). "k-MLE: A fast algorithm for learning statistical mixture models". arXiv:1203.5181free to read [cs.LG].

References

[1] D. H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic games. *IEEE Transactions on Automatic Control*, AC-18-2:124–131, 1973.

[2] J. T.-H. Lo. Convexification for data fitting. *Journal of Global Optimization*, 46-2:317, 2010.

[3] J. T.-H. Lo and D. Bassu. An adaptive method of training multilayer perceptrons. In *Proceedings of the 2001 International Joint Conference on Neural Networks*, Washington, D.C., July, 2001.

[4] J. T.-H. Lo and D. Bassu. Robust identification of dynamic systems by neurocomputing. In *Proceedings of the 2001 International Joint Conference on Neural Networks*, Washington, D.C., July, 2001.

[5] J. T.-H. Lo, Y. Gui, and Y. Peng. Overcoming the local-minimum problem in training multilayer perceptrons with the NRAE training method. In J. Wang, G. Yen, and M. Polycarpou, editors, *Advances in Neural Networks - ISNN 2012*, Shenyang, China, 2012. International Symposium on Neural Networks.

[6] J. T.-H. Lo, Y. Gui, and Y. Peng. Overcoming the local-minimum problem in training multilayer perceptrons with the NRAE-MSE training method. In C. Guo, Z.-G. Hou, and Z. Zeng, editors, *Advances in Neural Networks - ISNN 2013*, Dalian, China, 2013. International Symposium on Neural Networks.

[7] J. T.-H. Lo, Y. Gui, and Y. Peng. The normalized risk-averting error criterion for avoiding nonglobal local minima in training neural networks. *Neurocomputing*, 149:3–12, 2015.

[8] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[9] J. Speyer, J. Deyst, and D. H. Jacobson. Optimaization of stochastic linear systems with additive measurement and process noise using exponential performance criteria. *IEEE Transactions on Automatic Control*, AC-19:358–366, 1974.

[10] P. Whittle. *Risk Sensitive Optimal Control*. Wiley, New York, New York, 1990.