Statistical Mechanics and Phase Transitions in Clustering

Kenneth Rose, Eitan Gurewitz,[a] and Geoffrey C. Fox
Caltech Concurrent Computation Program, California Institute of Technology, Mail Stop 206-49, Pasadena, California 91125
(Received 2 May 1990)

A new approach to clustering based on statistical physics is presented. The problem is formulated as fuzzy clustering and the association probability distribution is obtained by maximizing the entropy at a given average variance. The corresponding Lagrange multiplier is related to the "temperature" and motivates a deterministic annealing process where the free energy is minimized at each temperature. Critical temperatures are derived for phase transitions when existing clusters split. It is a hierarchical clustering estimating the most probable cluster parameters at various average variances.

PACS numbers: 05.70.Fh, 02.50.+s, 89.70.+c

Many natural phenomena are in fact optimization processes, and the drive to understand and analyze them yielded powerful mathematical methods over the years. Thus, when wishing to solve a hard optimization problem, it is advantageous to apply these methods through a physical analogy. In this work we apply methods of statistical physics to the problem of clustering, which is an important optimization problem in a large variety of fields, such as astrophysics, pattern recognition, and learning and data compression, to name but a few. Clustering methods are major tools for the analysis of data without knowledge of *a priori* distributions. In its most basic form, the problem of clustering is that of partitioning a given set of data points into subgroups, each of which should be as homogeneous as possible. This is usually made mathematically precise by defining a cost function or criterion to be minimized. All useful cost functions to our knowledge are not convex and have several local minima. Traditional clustering techniques are "descent" algorithms in the sense that at each iteration the cost is reduced. For this reason they tend to get trapped in a local minimum.

The interest in the application of statistical mechanics to nonconvex optimization problems has been growing recently. A known technique for nonconvex optimization is stochastic relaxation or simulated annealing,[1] based on the Metropolis algorithm.[2] However, one must be very careful with the annealing schedule, the rate at which the temperature is lowered. In their work on image restoration, Geman and Geman[3] have shown that, in theory, the global minimum can be achieved if the schedule obeys $T \propto 1/\ln n$, where n is the number of the current iteration. Such schedules are not realistic in many applications. Unlike stochastic relaxation where random moves are made on the given cost surface, deterministic annealing can be viewed as incorporating the "randomness" into the cost function. This cost function is deterministically optimized at each temperature sequentially, starting at high temperature and going down. Differing methods based on this concept have been applied to the traveling-salesman and other problems.[4-6]

The essential approach in this study is that no assumption is made on the data distribution. The only measurable quantity, for a given clustering configuration, is the energy (cost) function.[7] This meets the basic philosophy of information theory[8] and statistical mechanics[9] and leads to applying the maximum-entropy principle.

(1) Maximum-entropy clustering.—The formulation of clustering within a probabilistic framework is advantageous, as can be seen from the growing interest in fuzzy clustering[10,11] in the past two decades. The main principle is that each point is associated *in probability* with each cluster. The state of the system is given by the set of probability distributions for associating points with clusters. In the fuzzy-clustering literature these association probabilities are called "fuzzy membership in clusters." Hard clustering is a marginal special case, where each point is deterministically associated with a single cluster, and is therefore a more restrictive approach. The objective within our probabilistic framework will thus be to find an optimal probability distribution of associations.

Let $E_j(x)$ denote the energy (cost) associated with assigning a data point x to the cluster C_j. The average total cost for a given configuration of clusters is

$$\langle E \rangle = \sum_x \sum_j P(x \in C_j) E_j(x) , \tag{1}$$

where the summation is over all the given data points and all clusters. This quantity is our clustering measure. Since we do *not* make any assumption about the data distribution, we apply the principle of maximum entropy. In particular, of all possible probability distributions which yield a given average total cost, we choose the one which maximizes the entropy. As is well known, the association probabilities which maximize the entropy under the constraint (1) are Gibbs distributions, i.e.,

$$P(x \in C_j) = e^{-\beta E_j(x)}/Z_x , \tag{2}$$

where Z_x is the partition function

$$Z_x = \sum_k e^{-\beta E_k(x)} . \tag{3}$$

The parameter β is the Lagrange multiplier determined

by the given value of $\langle E \rangle$ in (1). In our physical analogy, β is inversely proportional to the temperature. As β gets larger, the associations become less fuzzy. In fact, for $\beta=0$ each point is equally associated with all clusters, while for $\beta \to \infty$ each point belongs to exactly one cluster with probability 1.

For a given set of clusters, it is assumed that the probabilities relating different x's to their clusters are independent. Hence the total partition function is

$$Z=\prod_x Z_x . \tag{4}$$

(2) Effective cost or free energy.—The previous section was derived for a fixed set of clusters. We shall now extend the derivation to include optimization over the cluster parameters. Let y_j be the parameter vector which specifies the cluster C_j (e.g., centroid and variance). An instance of the system is given by a set of cluster-parameter vectors $Y=\{y_j\}$, and a set of associations $V=\{v_{xj}\}$, where

$$v_{xj}=\begin{cases}1 \text{ if } x \in C_j,\\ 0 \text{ otherwise.}\end{cases}$$

By the same reasoning as before, in order to avoid assumptions on the distribution of the data, the probability of this instance is given by the Gibbs distribution

$$P(Y,V)=\frac{e^{-\beta E(Y,V)}}{\sum_{Y',V'} e^{-\beta E(Y',V')}} , \tag{5}$$

where

$$E(Y,V)=\sum_x \sum_j v_{xj} E_j(x) . \tag{6}$$

The most probable instance is the one which maximizes the probability in (5), i.e., the instance of lowest energy. This is the result one wishes to obtain for hard clustering. However, if one is more interested in estimating the most probable set of cluster parameters, as is the case when one tries to generalize from a given set of training samples, then the following marginal probability should be considered:

$$P(Y)=\sum_V P(Y,V) , \tag{7}$$

where the summation is performed over *all legal* associations. A legal association V is such that each data point is assigned to *exactly* one cluster. By using (6) we obtain the identity

$$\sum_V e^{-\beta E(Y,V)}=\prod_x \sum_k e^{-\beta E_k(x)} .$$

The right-hand side allows us to reintroduce into (5) the partition function Z as defined in (3) and (4), which is a function of the vector set Y. The marginal probability of (7) now becomes

$$P(Y)=\frac{Z}{\sum_{Y'} Z} . \tag{8}$$

Based on the partition function we derive the free energy as

$$F=-\frac{1}{\beta}\ln Z , \tag{9}$$

which is also a function of Y. We can now rewrite (8) as

$$P(Y)=\frac{e^{-\beta F}}{\sum_{Y'} e^{-\beta F}} . \tag{10}$$

It is evident from (10) that the most probable set of vectors Y is the one which minimizes the free energy (F). Therefore, the effective cost to be minimized for estimating the cluster parameters is indeed the free energy. In order to proceed and minimize the effective cost, we have to specify $E_j(x)$, the cost for associating a point with a given cluster.

(3) The squared-distance cost.—As an important example we chose the cost for associating x with C_j as

$$E_j(x)=|x-y_j|^2 , \tag{11}$$

where y_j is the centroid of C_j. The point association probability is by (2)

$$P(x \in C_j)=\frac{e^{-\beta|x-y_j|^2}}{\sum_k e^{-\beta|x-y_k|^2}} , \tag{12}$$

and the corresponding free energy (9) is

$$F=-\frac{1}{\beta}\sum_x \ln\left(\sum_k e^{-\beta|x-y_k|^2}\right) . \tag{13}$$

The set of vectors Y which optimizes the effective cost satisfies

$$\frac{\partial}{\partial y_j} F=0, \quad \forall j .$$

It should be noted that y_j are vectors and this is shorthand notation implying differentiation with respect to each component to yield the 0 vector. Differentiating (13) we obtain

$$\sum_x \frac{(x-y_j)e^{-\beta|x-y_j|^2}}{\sum_k e^{-\beta|x-y_k|^2}}=0, \quad \forall j . \tag{14}$$

This means that the optimal y_j is indeed the center of mass or the average of the samples in C_j. In fact, it satisfies

$$y_j=\frac{\sum_x x P(x \in C_j)}{\sum_x P(x \in C_j)} ; \tag{15}$$

i.e., we assign to each data point its relative weight in the cluster. Clearly, (15) is an implicit equation in y_j through (12). This naturally leads us to propose fixed-point iterations

$$Y^{(n+1)}=f(Y^{(n)}) , \tag{16}$$

where f is based on (12) and (15) for all clusters.

One may notice the similarity between the above re-

sults and the maximum-likelihood estimate of means in normal mixtures.[12,13] An important distinction to keep in mind is that here there is no prior knowledge or assumption on the probability densities of the data. The parametrized probability distribution involved is derived directly from the cost function to be minimized and is indeed the corresponding Gibbs distribution. Since we chose the squared-distance cost, our association probability distribution becomes Gaussian.

(4) Annealing and phase transitions.—The procedure described above determines a set of vectors $\{y_j\}_{j=1,\ldots,n}$ for each fixed β. In principle, changing n, the assumed number of clusters, will modify the resulting set of vectors. However, there exists some n_c such that for all $n > n_c$, one gets only n_c *distinct* vectors while the remaining $n - n_c$ vectors are repetitions from this set. Thus, at a given β we get at most n_c clusters. Therefore, we shall assume here that we have enough vectors to produce the maximal number of clusters (n_c) at a given β, and we shall only consider them without repetitions (e.g., if all vectors are identical then we have one cluster).

The free energy F and the β defined in section (1) are Legendre transform images of each other. Fixing one of them determines the other. When we extend the formulation to include the cluster parameters as in section (2), then for a given β we locate the cluster parameters which minimize F. For $\beta=0$, each data point is uniformly associated with all clusters (12), and thus by (15) all the parameter vectors will be identical, and will point to the center of mass of the data. Clearly, for $\beta=0$ we have a single minimum (which is the global minimum) for F, and the entire data set is interpreted as one cluster. At higher β, the free energy may have many local minima, and the concept of annealing emerges here can be viewed as tracking the global minimum while gradually increasing β. Moreover, at $\beta=0$ there is only one cluster ($n_c=1$), but at some positive β we shall have $n_c > 1$. In other words, this cluster will split into smaller clusters, and will thus undergo a phase transition. The new clusters will then split at higher β, so that the process may be viewed as a sequence of phase transitions. Figure 1 shows the phase diagram for a clustering example.

Let us derive the critical value β_c at which the first phase transition occurs. We have one cluster centered at the data center of mass. Without loss of generality we shall take this point to be the origin. The phase transition occurs when we have nonzero vectors which minimize F, i.e., nonzero solutions to (14). It can be shown that our process is continuous at the first phase transition. Hence, $|y_k| \approx 0$, $\forall\ k$ on a small neighborhood of the first phase transition, and by series expansion in y_j we get

$$\sum_x (y_j - x)[1 - \beta|y_j|^2 + 2\beta x^t y_j] = 0\,,$$

where the superscript t denotes transposition. Discarding higher powers of $|y_j|$ we obtain in matrix notation

$$(I - 2\beta C_{xx})y_j = 0\,, \tag{17}$$

where C_{xx} is the covariance matrix of the data, and I is the identity matrix. The critical value for β is thus

$$\beta_c = 1/2\lambda_{\max}\,, \tag{18}$$

where $\lambda_{\max}$ is the largest eigenvalue of C_{xx}. We conclude that the critical temperature is determined by the variance along the largest principal axis of the distribution. Moreover, y_j is an eigenvector of C_{xx}, which means that the split will be initiated in the direction of this principal axis. It is easy to see that as long as we may neglect intercluster influences, this derivation will hold for the following phase transitions, and every cluster will split at the critical temperature corresponding to its variance. In Fig. 2 we show the resulting clustering at the phases given in Fig. 1. The "explosion" in Fig. 2(d) is explained by the isotropic distribution of the clusters that are being split. In such a case, all vectors are eigenvectors, and at the critical temperature they all become possible solutions.

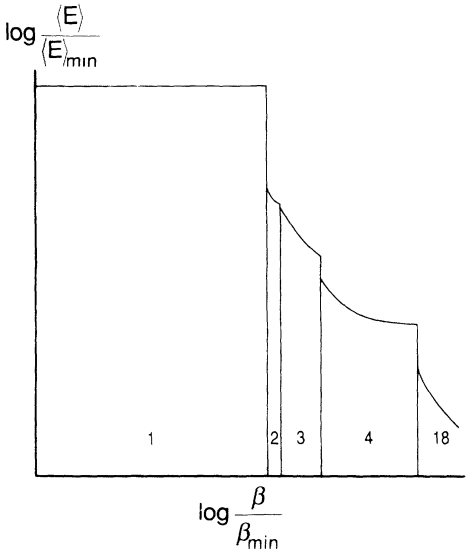

FIG. 1. Phase diagram for the distribution shown in Fig. 2. The number of actual clusters is shown for each phase.

In conclusion, our annealing process with its phase transitions gives us a natural hierarchical clustering. What we have is not a single clustering solution, but a hierarchy of solutions (or a multiscale solution) where each level corresponds to a different scale, from coarse to fine detail. This is also equivalent to the basic philosophy of rate-distortion theory, where one derives a set of optimal rates for representing the data at different levels of distortion. An open question on which we are currently working is under what circumstances, if at all, do first-order transitions occur, as these are not predictable by the derivation of the previous section. In all our ex-

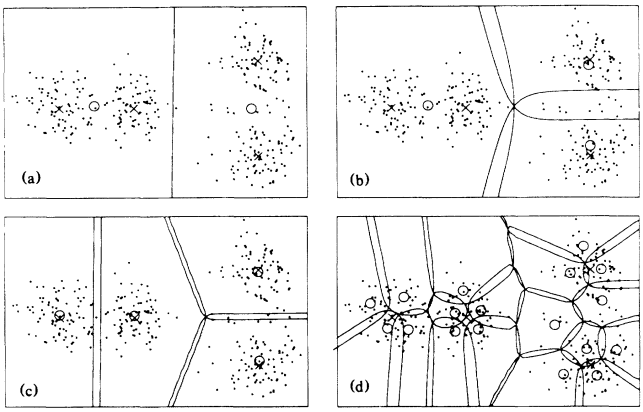

FIG. 2. Clustering at different phases corresponding to Fig. 1. The data are generated from four Gaussian distributions centered at the location marked by ×. The calculated cluster centroids are marked by ○. (a) Two clusters ($\beta=0.005$), (b) three clusters ($\beta=0.01$), (c) four clusters ($\beta=0.05$), and (d) eighteen clusters ($\beta=0.1$).

periments we have not come across such a phase transition.

This work was supported by the Joint Tactical Fusion Program Office, Reference No. A:49-288-82201-0-8420.

[a]On leave of absence from the Department of Physics, Nuclear Research Centre–Negev, P.O. Box 9001, Beer-Sheva, Israel.

[1]S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).

[2]N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[3]S. Geman and D. Geman, IEEE Trans. Pattern Anal. Machine Intell. **6**, 721 (1984).

[4]R. Durbin and D. Willshaw, Nature (London) **326**, 689 (1987).

[5]P. D. Simic, Network **1**, 89 (1990).

[6]A. L. Yuille, Neural Computation **2**, 1 (1990).

[7]K. Rose, E. Gurewitz, and G. C. Fox, California Institute of Technology Technical Report No. C3P-857, 1990 (to be published).

[8]C. Shannon and W. Weaver, *The Mathematical Theory of Communication* (Univ. of Illinois Press, Urbana, IL, 1949).

[9]E. T. Jaynes, in *Papers on Probability, Statistics and Statistical Physics,* edited by R. D. Rosenkrantz (Kluwer, Dordrecht, The Netherlands, 1989).

[10]J. C. Dunn, J. Cybern. **3**, 32 (1974).

[11]J. C. Bezdek, IEEE Trans. Pattern Anal. Machine Intell. **2**, 1 (1980).

[12]R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1974).

[13]A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data* (Prentice Hall, Englewood Cliffs, NJ, 1988).