
An Analysis of Model-based Clustering, Competitive Learning, and Information Bottleneck

Shi Zhong

Department of Computer Science and Engineering
Florida Atlantic University
777 Glades Road, Boca Raton, FL 33431
zhong@cse.fau.edu

Abstract

This paper provides a general formulation of probabilistic model-based clustering with deterministic annealing (DA), which leads to a unifying analysis of k-means, EM clustering, soft competitive learning algorithms (e.g., self-organizing map), and information bottleneck. The analysis points out an interesting yet not well-recognized connection between the k-means and EM clustering—they are just two different stages of a DA clustering process, with different temperatures. Demonstrated relationships between model-based clustering, competitive learning, and information bottleneck, can potentially generate a series of new algorithms.

1 Introduction

Simulated annealing (Kirkpatrick et al., 1983) is a stochastic optimization technique motivated by the annealing processes in physical chemistry. Certain chemical systems can be driven to their low-energy states by annealing, a gradual temperature-decreasing process. The annealing schedule, i.e., the rate at which the temperature is lowered, is critical to reaching the global optimum. Geman and Geman (1984) have theoretically shown that the global optimum can be achieved by a schedule following $T \propto \frac{1}{\log m}$, where m is the current iteration number. Such schedules, however, are very slow and unrealistic in many applications.

Deterministic annealing (DA), as the name suggests, is a deterministic version of simulated annealing (Rose, 1998). Derived from an information theoretic view of optimization problems, DA is not guaranteed to reach global optimum. Rather, it is a heuristic strategy that avoids many locally optimal solutions and enjoys a faster temperature schedule. It has been successfully used in a wide range of applications (Rose, 1998).

But its applications to clustering have been largely restricted to vector data (Rose et al., 1993; Hofmann & Buhmann, 1997). This paper provides a probabilistic model-based DA formulation, which helps us make insightful connections between mixture models, competitive learning, and information bottleneck.

Mixture models (McLachlan & Basford, 1988) have been widely used for density estimation and clustering applications. The EM algorithm (Dempster et al., 1977) is usually used to estimate maximum likelihood (or maximum *a posteriori*) parameters for mixture models. When used for clustering, it is called EM clustering. The standard k-means algorithm (Forgy, 1965; MacQueen, 1967) is another well-known clustering technique that has its root in statistical analysis and that is closely related to vector quantization in pattern recognition and signal processing. The standard K-means algorithm is often viewed as a special case of spherical Gaussian model-based EM clustering with variance $\sigma^2 \rightarrow 0$ (Mitchell, 1997), yet more sophisticated analyses exist. For example, an information-theoretic analysis was provided by Kearns et al. (1997), who use the term “k-means” to mean a generic algorithm that is not limited to vector space with Gaussian models. We follow this conceptual notation in our discussion. In this paper, we shall introduce a more intuitive explanation for the relationship between k-means and EM clustering, under a model-based deterministic annealing framework.

The online version of k-means also has close connection to competitive learning techniques (Ahalt et al., 1990), in which the k-means algorithm is viewed to use the Winner-Take-All learning rule while EM clustering a SoftMax learning rule. Self-organizing maps (SOM) (Kohonen, 1997) and Neural-Gas (Martinetz et al., 1993) are two sophisticated competitive learning methods that use an annealing mechanism to avoid locally optimal solutions. Connections between SOM, Neural-Gas, and DA have been made for vector quantization applications by Hofmann and Buhmann (1998), who

showed that the three types of algorithms are three different implementations of a continuation method (Allgower & Georg, 1990) for vector quantization, with different competitive learning rules. We shall show that SOM and Neural-Gas can be interpreted as variations of DA clustering and can be extended to employ different probabilistic models.

Recently, information bottleneck (IB) (Tishby et al., 1999) has attracted a lot of interests and gained popularity in text clustering applications. An explanation of its connection to maximum likelihood algorithms has been attempted by Slonim and Weiss (2003). By deriving IB formulation from multinomial model-based deterministic annealing, we show that IB text clustering explored recently by many authors implicitly assumes a multinomial distribution for clusters.

In short, this paper extends our previous work (Zhong & Ghosh, 2003) with a more general formulation of model-based deterministic annealing and a thorough discussion. Through the formulation, we demonstrate some useful connections between maximum likelihood estimation of mixture models, competitive learning, and information bottleneck. We believe these connections may lead to a series of new algorithms.

Section 2 presents model-based deterministic annealing. Section 3 demonstrates k-means and EM clustering as special stages of a model-based DA clustering process. Section 4 and 5 discuss the relationships between model-based DA clustering and competitive learning as well as information bottleneck. Section 6 presents possible new applications and algorithms based on our general formulation of model-based DA clustering. Finally, Section 7 summarizes this paper with remarks on future work.

2 Model-based Deterministic Annealing

In this section, we present a general formulation of model-based deterministic annealing and demonstrate the connection to Neal and Hinton’s EM interpretation (1998). Although the EM and DA algorithms apply to general optimization problems, for simplicity, here we restrain our discussion to probabilistic model-based clustering.

Consider a central clustering problem, in which $\mathcal{X} = \{x_1, \dots, x_N\}$ is a set of N data objects, $\mathcal{Z} = \{z_1, \dots, z_N\}$ a set of N (hidden or unobserved) cluster indices, and Θ a set of parameters associated with a probabilistic model of X and Z , two random variables governing the distribution of x and z . Note $z \in \{1, \dots, K\}$ is a discrete variable and K is the number of clusters. For each cluster z , the corresponding model for X is

$P(x|\Theta, z) = P(x|\theta_z)$, which measures the likelihood of object x coming from cluster z . The parameter θ_z is the set of parameters for cluster z . Our objective is to maximize the following average log-likelihood:

$$\begin{aligned} E &= \sum_x \sum_z P(x, z|\Theta) \log P(x|\theta_z) \\ &= \sum_x P(x|\Theta) \sum_z P(z|x, \Theta) \log P(x|\theta_z) \\ &\propto \sum_{x,z} P(z|x, \Theta) \log P(x|\theta_z), \end{aligned} \quad (1)$$

where the (parameterized) data prior $P(x|\Theta)$ is unavoidably set to be constant $1/N$ in practice as we always deal with a finite sample (but as $N \rightarrow \infty$, using $1/N$ is asymptotically correct). For simplicity, we ignore this constant and use the simplified notation in (1) in later analysis.

Maximizing E directly over $P(z|x, \Theta)$ leads to a k-means type partitioning of data objects, i.e., hard data assignment. That is, for each data object x , the posterior $P(z|x, \Theta)$ has a value of 1 for $z = \arg \max_{z'} P(x|\theta_{z'})$ and 0 otherwise. To consider soft partitioning, one can introduce entropy constraints or penalties into (scaled) objective E :

$$\begin{aligned} F &= \sum_{x,z} P(z|x, \Theta) \log P(x|\theta_z) \\ &\quad - \gamma \cdot H(Z|\Theta) + T \cdot H(Z|X, \Theta), \end{aligned} \quad (2)$$

where $H(Z|\Theta) = -\sum_z P(z|\Theta) \log P(z|\Theta)$ is the cluster prior entropy,

$$H(Z|X, \Theta) = -\sum_{x,z} P(z|x, \Theta) \log P(z|x, \Theta)$$

is the average cluster posterior entropy, and γ and T are coefficients used to tradeoff between maximizing E , minimizing $H(Z|\Theta)$, and maximizing $H(Z|X, \Theta)$. Intuitively, minimizing $H(Z|\Theta)$ has an effect of favoring smaller number of clusters while maximizing $H(Z|X, \Theta)$ favors soft posterior assignment of data objects to different clusters (to avoid hard decisions, leading to so-called maximum entropy clustering). The $H(Z|\Theta)$ term is not in the original DA formulation (Rose, 1998) but was mentioned as a possible entropy-constrained extension that has been used in vector quantization. We shall see how it is useful in making connections to EM clustering. The parameter T has a temperature interpretation, as explained later in this section.

The function F can be rewritten as

$$\begin{aligned} F &= \sum_{x,z} P(z|x, \Theta) [\log P(x|\theta_z) \\ &\quad + \gamma \log P(z|\Theta) - T \log P(z|x, \Theta)] . \end{aligned} \quad (3)$$

Using classic Lagrange optimization method, together with constraint $\sum_z P(z|x, \Theta) = 1$, one can derive the maximum entropy solution of posterior $P(z|x, \Theta)$ (for maximizing F) to be the well-know Gibbs distribution

$$\begin{aligned} P(z|x, \Theta) &= \frac{\exp\left(\frac{\log P(x|\theta_z) + \gamma \log P(z|\Theta)}{T}\right)}{\sum_{z'} \exp\left(\frac{\log P(x|\theta_{z'}) + \gamma \log P(z'|\Theta)}{T}\right)} \\ &= \frac{P(x|\theta_z)^{\frac{1}{T}} P(z|\Theta)^{\frac{\gamma}{T}}}{\sum_{z'} P(x|\theta_{z'})^{\frac{1}{T}} P(z'|\Theta)^{\frac{\gamma}{T}}}. \end{aligned} \quad (4)$$

Several interesting choices of γ are discussed in the next section. In general, when temperature T is high, the posterior entropy is high, meaning the probability of each data object being assigned to different clusters is balanced (or uniform). Conversely, as T decreases to 0, the posterior probability $P(z|x, \Theta)$ will be close to either 0 or 1, leading to a hard data assignment. A DA algorithm naturally starts at a high temperature and then gradually decreases the temperature towards 0. At each temperature, the EM algorithm is used to maximize the objective (2), with a E-step in (4) and a M-step

$$\theta_z^{(new)} = \arg \max_{\theta} \sum_x P(z|x, \Theta) P(x|\theta). \quad (5)$$

A generic model-based DA clustering algorithm was constructed in Zhong and Ghosh (2003). Note that in this formulation, the temperature T only affects the calculation of posterior $P(z|x, \Theta)$, i.e., the E-step of the EM algorithm. Annealing M-step is possible but computationally much more expensive than the DA formulation.

The relationship to maximum likelihood estimation of mixture models with the EM algorithm is demonstrated as follows. According to the view of Neal and Hinton (1998), the objective function that the EM algorithm maximizes is

$$\begin{aligned} F_1 &= E[\log P(\mathcal{X}, \mathcal{Z}|\Theta)|\mathcal{X}, \Theta] + H(Z|X, \Theta) \quad (6) \\ &= \sum_{x,z} P(z|x, \Theta) \log P(x, z|\Theta) + H(Z|X, \Theta) \quad (7) \\ &= \sum_{x,z} P(z|x, \Theta) \log P(x|\theta_z) - H(Z|\Theta) \\ &\quad + H(Z|X, \Theta), \end{aligned} \quad (8)$$

where \mathcal{Z} is the (hidden) cluster indices of all data objects, the first term in (6) is an expectation of complete data log-likelihood over $P(z|x, \Theta)$, and the second term in (6) is the posterior entropy. Derivation of (7) from (6) can be found in Blimes (1998), thus is omitted here. From (7) to (8) is straightforward. Therefore, the objective function in (2) is the same as that used in Neal and Hinton (1998) if we set

$\gamma = T = 1$. Neal and Hinton (1998) showed that the optimal value of F_1 over $P(z|x, \Theta)$ is exactly the incomplete data log-likelihood $\log P(\mathcal{X}|\Theta)$.

3 K-means vs. EM Clustering

Let us now discuss several choices of the γ parameter, which elucidates an intuitive relationship between k-means and EM clustering.

- $\gamma = 0$:

The objective F becomes

$$\sum_{x,z} P(z|x, \Theta) \log P(x|\theta_z) + T \cdot H(Z|X, \Theta),$$

which is the same as the original DA formulation with the distortion function set to be $-\log P(x|\theta_z)$. As $T \rightarrow \infty$, $P(z|x, \Theta)$ becomes uniform, which makes sense when there is no prior knowledge about the distribution of z . As T tends to 0, the clustering algorithm becomes k-means, which maximizes the first term in (2).

- $\gamma = 1$:

The objective F becomes

$$\sum_{x,z} P(z|x, \Theta) \log P(x, z|\Theta) + T \cdot H(Z|X, \Theta),$$

which further reduces to standard EM clustering (7) if $T = 1$. As $T \rightarrow \infty$, $P(z|x, \Theta)$ becomes uniform, same as the $\gamma = 0$ case. As T tends to 0, the clustering algorithm becomes a hard clustering algorithm that is slightly different from k-means. Specifically, $P(z|x, \Theta)$ is 1 for $z = \arg \max_{z'} P(z'|\Theta) P(x|\theta_{z'})$ and 0 otherwise. This hard clustering version was observed and analyzed in Banerjee et al. (2003). It seems to maximize a lower bound of incomplete data (log-)likelihood.

- $\gamma = T$:

This is the scenario we are mainly interested in. The objective F becomes

$$\sum_{x,z} P(z|x, \Theta) \log P(x|\theta_z) - T \cdot I(X; Z|\Theta), \quad (9)$$

where one can easily see that an interpretation on maximizing F is to minimize the mutual information between X and Z , i.e., to compress X into Z as much as possible, and meanwhile to maximize the average log-likelihood. The temperature T is used to adjust such a tradeoff. The E-step now becomes

$$P(z|x, \Theta) = \frac{P(z|\Theta) P(x|\theta_z)^{\frac{1}{T}}}{\sum_{z'} P(z'|\Theta) P(x|\theta_{z'})^{\frac{1}{T}}}.$$

As $T \rightarrow \infty$, the cluster posteriors become equal to priors, i.e., $P(z|x, \Theta) = P(z|\Theta)$. This is useful when one intends to enforce some prior knowledge on the distribution of cluster sizes (which can not be achieved for the previous two scenarios). As T lowers to 1, the clustering algorithm is equivalent to EM clustering; as T further goes to 0, the algorithm becomes k-means. Therefore, we can view EM clustering and k-means as two different stages of a model-based DA clustering process, with $T = 1$ and $T = 0$, respectively. This is a better interpretation than the traditional one in Mitchell (1997), where one has to artificially diminish the variance of Gaussian models, destroying reasonable Gaussian model descriptions for the clusters generated by the standard k-means algorithm.

4 Relation to Competitive Learning

In this section, we discuss two related competitive learning algorithms that have an annealing flavor—self-organizing map (Kohonen, 1997) and Neural-Gas (Martinetz et al., 1993), both have been used for clustering. To see the connection clearly, we first discuss batch versions of the two algorithms.

For the original SOM algorithm, the model parameters are just the means of each cluster, i.e., $\Theta = \{\mu_z\}_{z=1, \dots, K}$. A distinct feature of SOM is the use of a topological map, in which each cluster has a fixed coordinate. Let the map location of cluster z be ϕ_z and $K_\alpha(\phi_1, \phi_2) = \exp\left(-\frac{\|\phi_1 - \phi_2\|^2}{2\alpha^2}\right)$ a neighborhood function. Let $z(x) = \arg \max_z \|x - \mu_z\|$, where $\|\cdot\|$ is L_2 norm. The batch SOM algorithm amounts to iterating between the following two steps:

$$P(z|x, \Theta) = \frac{K_\alpha(\phi_z, \phi_{z(x)})}{\sum_{z'} K_\alpha(\phi_{z'}, \phi_{z(x)})} \quad (10)$$

and

$$\mu_z^{(new)} = \frac{1}{N} \sum_x P(z|x, \Theta) x, \quad (11)$$

where α is a parameter controlling the width of the neighborhood function and decreases gradually during the clustering process. As α goes to 0, $P(z|x, \Theta)$ becomes either 0 or 1 and the algorithm reduces to k-means.

It is immediately noted that the α has the same functionality of a temperature parameter in deterministic annealing. The difference from standard model-based deterministic annealing is that here the calculation of $P(z|x, \Theta)$ is constrained by a topological map structure, which gives SOM the advantage that all resulting clusters are structurally related according to the

pre-specified topological map. Obviously, this modified E-step makes SOM not a strict EM algorithm; maximum likelihood statement, as well as the convergence theorem of the EM algorithm (Dempster et al., 1977), do not apply. Loosely speaking, however, the convergence of SOM is guaranteed by the convergence of k-means during the final stage of annealing process where α becomes close to 0 (thus $P(z|x, \Theta)$ becomes hard and the algorithm becomes k-means).

Comparing the batch SOM algorithm with model-based DA clustering, we notice that a natural extension of SOM is to use probabilistic models in the M-step (11). Heskes (2001) discussed this direction and explored multinomial model-based SOM for analyzing market basket data.

The Neural-Gas algorithm differs from the SOM clustering, only in the E-step, i.e., how $P(z|x, \Theta)$ is calculated. It uses a rank-based scheme

$$P(z|x, \Theta) = \frac{e^{-r(x,z)/\alpha}}{\sum_{z'} e^{-r(x,z')/\alpha}}, \quad (12)$$

where α is again an equivalent temperature parameter and $r(x, z)$ a rank function that takes value $k - 1$ if μ_z is the k -th closest cluster centroid to data vector x (according to L_2 distance). Convergence of this algorithm has been analyzed and interpreted using diffusion dynamics (Martinetz et al., 1993). Similar to SOM, however, the convergence is obvious at the final stage of annealing as α lowers to 0. Though it seems natural to extend Neural-Gas to use probabilistic models and rank clusters based on the log-likelihood measure $P(x|\theta_z)$, we have not seen any model-based Neural-Gas algorithms so far. So model-based Neural-Gas algorithms can be an interesting future work and useful in analyzing non-Gaussian data.

Competitive learning techniques are usually presented in online form—as each data item is presented, multiple clusters (in the clustering context) compete for it and the cluster centroids get updated according to the competing results. The updating part basically follows a gradient descent approach and the learning rate needs to be carefully selected (based on stochastic approximation principles, see Cherkassky & Mulier, 1998). Online algorithms are better equipped than batch versions to handle very large data sets that can not fit into memory. The connection demonstrated above indicates that an online version of any model-based DA clustering algorithm can be readily constructed following the practice used in competitive learning methods.

5 Relation to Information Bottleneck

In this section, we show an equivalence between IB clustering and multinomial model-based DA clustering. This result indicates that, when applied to clustering, the IB framework implicitly assume a discrete multinomial distribution for the representation of data. Even though the original IB framework does not specify discrete distributions, continuous applications have not been seen yet and may involve considerable computational difficulty.

Let $P(x|\theta_z)$ be a multinomial distribution parameterized by $\theta_z = \{P(y|z)\}$, where Y is a random variable associated with the set of possible symbols from which x is drawn. $P(y|z)$ is the probability of drawing symbol y in cluster z . Assume that each x is formed by drawing n times from the set of symbols and $n_x(y)$ is the number of times symbol y is drawn. We then have

$$\begin{aligned} \log P(x|\theta_z) &= \log \prod_y P(y|z)^{n_x(y)} \\ &= \sum_y n_x(y) \log P(y|z) \\ &= n \left(\sum_y P(y|x) \log \frac{P(y|z)}{P(y|x)} - H_{y|x} \right) \\ &= -n (D_{KL}(P(y|x)||P(y|z)) - H_{y|x}), \quad (13) \end{aligned}$$

where $D_{KL}(P||Q)$ is the Kullback-Leibler (KL) divergence between probability distributions P and Q , and $H_{y|x} = -\sum_y P(y|x) \log P(y|x)$ can be seen as a constant w.r.t. Z and Θ . On the other hand, assume that Y is independent of Z given X (which is true for clustering since Z is just a compressed version of X), it is easy to show that (Tishby et al., 1999; Dhillon et al., 2002)

$$I(X; Y) - I(Z; Y) \propto \sum_{x,z} P(z|x) D_{KL}(P(y|x)||P(y|z)).$$

Therefore, when multinomial models are used, plugging (13) into (9), we can show that

$$\begin{aligned} F &= \sum_{x,z} P(z|x, \Theta) \log P(x|\theta_z) - T \cdot I(Z; X|\Theta) \\ &= -n \sum_{x,z} P(z|x, \Theta) D_{KL}(P(y|x)||P(y|z)) \\ &\quad - n \cdot H_{y|x} - T \cdot I(Z; X|\Theta) \\ &= -n' (I(X; Y|\Theta) - I(Z; Y|\Theta)) \\ &\quad - n \cdot H_{y|x} - T \cdot I(Z; X|\Theta) \\ &= -T (I(Z; X|\Theta) - \beta I(Z; Y|\Theta)) + C, \quad (14) \end{aligned}$$

where $C = -n' \cdot I(X; Y|\Theta) - n \cdot H_{y|x}$ is a constant w.r.t. Z and Θ . Note that $I(Z; X|\Theta) - \beta I(Z; Y|\Theta)$ is a parameterized version of IB objective (to be minimized),

and $\beta = \frac{n'}{T}$ is the tradeoff coefficient and analogous to an inverse temperature parameter. This is a more accurate relationship than the one defined in Slonim and Weiss (2003) who attempted a mapping between information bottleneck and maximum likelihood estimation of mixture models. As we showed that the latter is simply one specific stage of a model-based DA process, the mapping defined in Slonim and Weiss (2003) would be exact only when $\beta = 1$, with multinomial models. It is not surprising to see the equivalence between multinomial model-based DA clustering and IB clustering since deterministic annealing was explicitly mentioned to be related in the original IB paper (Tishby et al., 1999).

It is worth mentioning that some recent works on KL clustering (Dhillon et al., 2002; Dhillon & Guan, 2003) are just (hard) k-means versions of IB clustering with $\beta \rightarrow \infty$ or multinomial model-based DA clustering with $T \rightarrow 0$. As β goes to infinity, minimizing $I(Z; X) - \beta I(Z; Y)$ is equivalent to maximizing $I(Z; Y)$, or same as minimizing information loss $I(X; Y) - I(Z; Y)$ (since $I(X; Y)$ is a constant).

6 Clustering with An Auxiliary Space

In this section, we discuss how the framework of model-based DA clustering can be useful in an interesting clustering scenario, in which each data item has two representations, one of them characterizes some auxiliary information of the data item. For example, in gene expression data analysis, auxiliary functional category information may be available for genes. One may desire to cluster genes according to the auxiliary representation while still using the expression information during the clustering process (Sinkkonen & Kaski, 2001). Similar situation exists for text clustering, where large amount of auxiliary ontology information of text documents exist and can be exploited.

Sinkkonen and Kaski (2001) proposed to minimize KL-divergence-based loss function in the auxiliary space and while using primary space models to calculate posterior probabilities $P(z|x, \Theta)$. In their work, multinomial models were used in the discrete auxiliary space and von Mises-Fisher models used for the continuous primary space. Rewritten in our notation, the objective they are maximizing is

$$\sum_{x,z} P(z|x, \Theta) \log P(x|\lambda_z)$$

where Θ is a set of von Mises-Fisher (vMF) model parameters and λ 's are parameters for the multinomial models in the auxiliary space. This interpretation immediately suggest many variations of clustering data with two representations. A general objective function

could be

$$\sum_{x,z} P(z|x, \Theta, \Lambda) \log P(x|\theta_z, \lambda_z),$$

which extends the parameter Θ in (2) to include models in both spaces. Sinkkonen and Kaski (2001) chose $P(x|\theta_z, \lambda_z) = P(x|\lambda_z)$, i.e., to cluster in the auxiliary space. A design that considers both spaces is

$$\log P(x|\theta_z, \lambda_z) = \log P(x|\theta_z) + \eta \log P(x|\lambda_z),$$

where η adjust the tradeoff between modeling data in two spaces. Similarly, many choices exist for the form of posterior probability $P(z|x, \Theta, \Lambda)$. It can be based on Θ alone, on Λ alone, on a topological map, or on model ranks in either space. Which choice to pick will depend on one's intention and practical applications.

7 Conclusion

We have presented a general formulation for model-based deterministic annealing, which takes k-means and EM clustering as special cases. The formulation shows some interesting connections to learning mixture models with the EM algorithm, to competitive learning techniques such as self-organizing map and Neural-Gas, and to information bottleneck clustering. These connections may suggest a series of possible new algorithms that are useful in handling large volumes of data and data with multiple representations.

In the future, we shall mainly investigate the efficacy of the suggested new algorithms in text clustering and gene expression data analysis applications.

The information bottleneck as a general principle, has been extended by Friedman et al. (2001) to address interesting bi-clustering or co-clustering (Dhillon, 2001; Dhillon et al., 2003) and multi-clustering problems. How the model-based DA clustering can be adapted (or combined with IB principle) for these problems seems to be another promising future direction.

References

Ahalt, S. C., Krishnamurthy, A. K., Chen, P., & Melton, D. E. (1990). Competitive learning algorithms for vector quantization. *Neural Networks*, 3, 277–290.

Allgower, E. L., & Georg, K. (1990). *Numerical continuation methods: An introduction*. Berlin Heidelberg: Springer-Verlag.

Banerjee, A., Dhillon, I., Sra, S., & Ghosh, J. (2003). Generative model-based clustering of directional data. *Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (pp. 19–28).

Blimes, J. A. (1998). *A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models* (Technical Report). University of California at Berkeley.

Cherkassky, V., & Mulier, F. (1998). *Learning from data: Concepts, theory, and methods*. New York: John Wiley & Sons.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1–38.

Dhillon, I., & Guan, Y. (2003). Information theoretic clustering of sparse co-occurrence data. *Proc. IEEE Int. Conf. Data Mining* (pp. 517–520). Melbourne, FL.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (pp. 269–274).

Dhillon, I. S., Mallela, S., & Kumar, R. (2002). Enhanced word clustering for hierarchical text classification. *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (pp. 446–455).

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. *Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (pp. 89–98).

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics*, 21, 768.

Friedman, N., Mosenzon, O., Slonim, N., & Tishby, N. (2001). Multivariate information bottleneck. *Proc. 17th Conf. Uncertainty in Artificial Intelligence*.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6, 721–741.

Heskes, T. (2001). Self-organizing maps, vector quantization, and mixture modeling. *IEEE Trans. Neural Networks*, 12, 1299–1305.

Hofmann, T., & Buhmann, J. (1997). Pairwise data clustering by deterministic annealing. *IEEE Trans. Pattern Anal. Machine Intell.*, 19, 1–14.

Hofmann, T., & Buhmann, J. (1998). Competitive learning algorithms for robust vector quantization. *IEEE Trans. Signal Processing*, 46, 1665–1675.

- Kearns, M., Mansour, Y., & Ng, A. Y. (1997). An information-theoretic analysis of hard and soft assignment methods for clustering. *Proc. 13th Conf. Uncertainty in Artificial Intelligence* (pp. 282–293).
- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.
- Kohonen, T. (1997). *Self-Organizing Map*. New York: Springer-Verlag.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Symp. Math. Statistics and Probability* (pp. 281–297).
- Martinetz, T. M., Berkovich, S. G., & Schulten, K. J. (1993). “Neural-Gas” network for vector quantization and its application to time-series prediction. *IEEE Trans. Neural Networks*, *4*, 558–569.
- McLachlan, G., & Basford, K. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.
- Mitchell, T. (1997). *Machine learning*. McGraw Hill.
- Neal, R., & Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse and other variants. In M. I. Jordan (Ed.), *Learning in graphical models*, 355–368. Kluwer Academic Publishers.
- Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of The IEEE*, *86*, 2210–2239.
- Rose, K., Gurewitz, E., & Fox, G. C. (1993). Constrained clustering as an optimization method. *IEEE Trans. Pattern Anal. Machine Intell.*, *15*, 785–794.
- Sinkkonen, J., & Kaski, S. (2001). Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, *14*, 217–239.
- Slonim, N., & Weiss, Y. (2003). Maximum likelihood and the information bottleneck. *Advances in Neural Information Processing Systems 15* (pp. 335–342). Cambridge, MA: MIT Press.
- Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. *Proc. 37th Annual Allerton Conf. Communication, Control and Computing* (pp. 368–377).
- Zhong, S., & Ghosh, J. (2003). A unified framework for model-based clustering. *Journal of Machine Learning Research*, *4*, 1001–1037.