Journal of Machine Learning Research 9 (2008) 339-370 Submitted 11/06; Revised 9/07; Published 3/08

Generalization from Observed to Unobserved Features by Clustering

Eyal Krupka EYAL.KRUPKA@MAIL.HUJI.AC.IL
Naftali Tishby TISHBY@CS.HUJI.AC.IL
School of Computer Science and Engineering
Interdisciplinary Center for Neural Computation
The Hebrew University Jerusalem, 91904, Israel

Editor: Sanjoy Dasgupta

Abstract

We argue that when objects are characterized by many attributes, clustering them on the basis of a *random* subset of these attributes can capture information on the unobserved attributes as well. Moreover, we show that under mild technical conditions, clustering the objects on the basis of such a random subset performs almost as well as clustering with the full attribute set. We prove finite sample generalization theorems for this novel learning scheme that extends analogous results from the supervised learning setting. We use our framework to analyze generalization to unobserved features of two well-known clustering algorithms: *k*-means and the maximum likelihood multinomial mixture model. The scheme is demonstrated for collaborative filtering of users with movie ratings as attributes and document clustering with words as attributes.

Keywords: clustering, unobserved features, learning theory, generalization in clustering, information bottleneck

1. Introduction

Data clustering can be defined as unsupervised classification of objects into groups based on their similarity (see, for example, Jain et al., 1999). Often, it is desirable to have the clusters match some labels that are unknown to the clustering algorithm. In this context, good data clustering is expected to have homogeneous labels in each cluster, under some constraints on the number or complexity of the clusters. This can be quantified by mutual information (see, for example, Cover and Thomas, 1991) between the objects' cluster identity and their (unknown) labels, for a given complexity of clusters. However, since the clustering algorithm has no access to the labels, it is unclear how it can optimize the quality of the clustering. Even worse, the clustering quality depends on the specific choice of the unobserved labels. For example, good document clustering with respect to topics is very different from clustering with respect to authors.

In our setting, instead of attempting to cluster by some arbitrary labels, we try to predict unobserved features from observed ones. In this sense our target labels are simply other features that happened to be unobserved. For example, when clustering fruits based on their observed features such as shape, color and size, the target of clustering is to match unobserved features such as nutritional value or toxicity. When clustering users based on their movie ratings, the target of clustering is to match ratings of movies that were not rated, or not even created as yet.

In order to theoretically analyze and quantify this new learning scheme, we make the following assumptions. Consider a very large set of features, and assume that we observe only a *random*

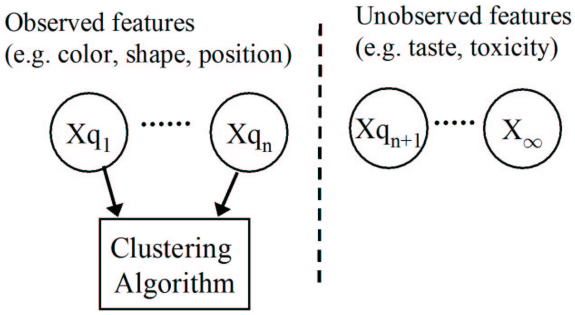


Figure 1: The learning scheme. The clustering algorithm has access to a random subset of features $(X_{q_1},...,X_{q_n})$ of m instances. The goal of the clustering algorithm is to assign a class label t_i to each instance, such that the expected mutual information between the class labels and a randomly selected *unobserved* feature is maximized.

subset of n features, called *observed features*. The other features are called *unobserved features*. We assume that the random selection of observed features is made from some unknown distribution $\mathcal{D}$ and each feature is selected independently.[1]

The clustering algorithm has access only to the observed features of m instances. After clustering, one of the *unobserved* features is randomly selected to be the target label. This selection is done using the same distribution, $\mathcal{D}$, of the observed feature selection. Clustering performance is measured with respect to this feature. Obviously, the clustering algorithm cannot be directly optimized for this specific feature.

The question is whether we can optimize the *expected* performance on the unobserved features, based only on the observed features. The expectation is over the *random* selection of the unobserved target features. In other words, can we find the clustering that is most likely to match a randomly selected unobserved feature? Perhaps surprisingly, for a large enough number of observed features, the answer is yes. We show that for any clustering algorithm, the average performance of the clustering with respect to the observed and unobserved features is similar. Hence we can indirectly optimize clustering performance with respect to unobserved features by analogy with generalization in supervised learning. These results are universal and do not require any additional assumptions such as an underlying model or a distribution that created the instances.

In order to quantify these results, we define two terms: the average observed information and the expected unobserved information. Let T be the variable which represents the cluster for each instance, and $\{X_1,...,X_L\}$ $(L \to \infty)$ the set of discrete random variables which denotes the features. The average observed information, denoted by I_{ob}, is the average mutual information between T and each of the observed features. In other words, if the observed features are $\{X_1,...,X_n\}$ then $I_{ob} = \frac{1}{n}\sum_{j=1}^{n} I(T;X_j)$. The expected unobserved information, denoted by I_{un}, is the *expected* value of the mutual information between T and a new *randomly* selected feature, that is, $E_{q\sim\mathcal{D}}\left\{I(T;X_q)\right\}$. We are interested in cases where this new selected feature is most likely to be one of the unobserved features, and therefore we use the term unobserved information. Note that whereas I_{ob} can be measured directly, this paper deals with the question of how to infer and maximize I_{un}.

1. For simplicity, we also assume that the probability of selecting the same feature more than once is near zero.

Our main results consist of two theorems. The first is a generalization theorem. It gives an upper bound on the probability of a large difference between I_{ob} and I_{un} for all possible partitions. It also states a *uniform convergence in probability* of $|I_{ob} - I_{un}|$ as the number of observed features increases. Conceptually, the average observed information, I_{ob}, is analogous to the training error in standard supervised learning (Vapnik, 1998), whereas the unobserved information, I_{un}, is similar to the generalization error.

The second theorem states that under constraints on the number of clusters, and a large enough number of observed features, one can achieve nearly the best possible performance, in terms of I_{un}. Analogous to the principle of Empirical Risk Minimization (ERM) in statistical learning theory (Vapnik, 1998), this is done by maximizing I_{ob}.

We use our framework to analyze clustering by the maximum likelihood of multinomial mixture model (also called Naive Bayes Mixture Model, see Figure 2 and Section 2.2). This clustering assumes a generative model of the data, where the *instances* are assumed to be sampled independently from a mixture of distributions, and for each such distribution all features are independent. These assumptions are quite different from our assumptions of fixed instances and randomly observed features.[2] Nevertheless, in Section 2.2 we show that this clustering achieves nearly the best possible clustering in terms of information on unobserved features.

In Section 3 we extend our framework to distance-based clustering. In this case the measure of the quality of clustering is based on some distance function instead of mutual information. We show that the *k*-means clustering algorithm (Lloyd, 1957; MacQueen, 1967) not only minimizes the observed intra-cluster variance, but also minimizes the unobserved intra-cluster variance, that is, the variance of unobserved features within each cluster.

Table 1 summarizes the similarities and differences of our setting to that of supervised learning. The key difference is that in supervised learning, the set of features is fixed and the training instances are assumed to be randomly drawn from some distribution. Hence, the generalization is to new instances. In our setting, the set of instances is fixed, but the set of observed features is assumed to be randomly selected. Hence, the generalization is to new features.

Our new theorems are evaluated empirically in Section 4, on two different data sets. The first is a movie ratings data set, where we cluster users based on their movie ratings. The second is a document data set, with words as features. Our main point in this paper, however, is the new conceptual framework and not a specific algorithm or experimental performance.

Section 5 discusses related work and Section 6 presents conclusions and ideas for future research. A notation table is available in Appendix B.

2. Feature Generalization of Information

In this section we analyze feature generalization in terms of mutual information between the clusters and the features. Consider a fixed set of m instances denoted by $\{\mathbf{x}[1],\ldots,\mathbf{x}[m]\}$. Each instance is represented by a vector of L features $\{x_1,\ldots,x_L\}$. The value of the qth feature of the jth instance is denoted by $x_q[j]$. Out of this set of features, n features are randomly and independently selected according to some distribution $\mathcal{D}$. The n randomly selected features are the *observed features* (variables) and their indices are denoted by $\tilde{\mathbf{q}} = (q_1,...,q_n)$, where $q_i \sim \mathcal{D}$. The ith observed feature of the jth instance is denoted by $x_{q_i}[j]$. After selecting the observed features, we also select *unobserved*

2. Note that in our framework, random independence refers to the *selection* of observed features, not to the feature values.

		Prediction of unobserved features	
	Supervised learning	Information	Distance-based
Training set	Randomly selected *instances*	*n* randomly selected features (observed features)	
Test set	Randomly selected *unlabeled instances*	Randomly selected *unobserved features*	
Hypothesis class	Class of functions from instances to labels	All possible partitions of *m* instances into *k* clusters	
Output of learning algorithm	Select hypothesis function	Cluster the *instances* into *k* clusters	
Goal	Minimize *expected* error on test set	Maximize *expected* information on *unobserved* features	Minimize *expected* intra-cluster variance of *unobserved* features
Assumption	Training and test instances are randomly and independently drawn from the same distribution	Observed and unobserved features are randomly and independently selected using the same distribution	
Strategy	Empirical Risk Minimization (ERM)	Observed Information Maximization (OIM)	Minimize observed intra-cluster variance
Related clustering algorithm		Maximum likelihood multinomial mixture model (Figure 2)	*k*-means
Good generalization	The training and test errors are similar	The observed and unobserved information are similar	The observed and unobserved intra-cluster variance are similar

Table 1: Analogy with supervised learning

features according to the same distribution $\mathcal{D}$. For simplicity, we assume that the total number of features, L, is large and the probability of selecting the same feature more than once is near zero (as in the case where $L \gg n^2$, where $\mathcal{D}$ is uniform distribution). This means that we can assume that a randomly selected unobserved feature is not one of the n observed features. It is important to emphasize that we have a fixed and finite set of instances; that is, we do not need to assume that the m instances were drawn from any distribution. Only the features are randomly selected.

We further assume that each of the features is a discrete variable with no more than s different values.[3] The clustering algorithm clusters the instances into k clusters. The clustering is denoted by the function $\mathbf{t} : [m] \to [k]$ that maps each of the m instances to one of the k clusters. The cluster label of the jth instance is denoted by $\mathbf{t}(j)$. Our measures for the quality of clustering are based on Shannon's mutual information. Let random variable Z denote a number chosen uniformly at random from $\{1,\ldots,m\}$. We define the quality of clustering with respect to a single feature, q, as $I(\mathbf{t}(Z); x_q[Z])$, that is, the empirical mutual information between the cluster labels and the feature.

Our measure of performance assumes that the number of clusters is predetermined. There is an obvious tradeoff between the preserved mutual information and the number of clusters. For example, one could put each instance in a different cluster, and thus get the maximum possible mutual information for all features. Obviously all clusters will be homogeneous with respect to all features but this clustering is pointless. Therefore, we need to have some constraints on the number of clusters.

Definition 1 *The **average observed information** of a clustering* $\mathbf{t}$ *and the observed features is denoted by* $I_{ob}(\mathbf{t}, \tilde{\mathbf{q}})$ *and defined by*

$$I_{ob}(\mathbf{t}, \tilde{\mathbf{q}}) = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{t}(Z); x_{q_i}[Z]).$$

*The **expected unobserved information** of a clustering is denoted by* $I_{un}(t)$ *and defined by*

$$I_{un}(\mathbf{t}) = \mathbf{E}_{q \sim \mathcal{D}} \left\{ I(\mathbf{t}(Z); x_q[Z]) \right\}.$$

In general, I_{ob} is higher when clusters are more coherent; that is, elements within each cluster have many identical observed features. I_{un} is high if there is a high probability that the clusters are informative on a randomly selected feature q (where $q \sim \mathcal{D}$). In the special case where the distribution $\mathcal{D}$ is uniform and $L \gg n^2$, I_{un} can also be written as the average mutual information between the cluster label and the unobserved features set; that is, $I_{un} \approx \frac{1}{L-n} \sum_{q \notin \{q_1,\ldots,q_n\}} I(\mathbf{t}(Z); x_q[Z])$. Recall that L is the total number of features, both observed and unobserved.

The goal of the clustering algorithm is to cluster the instances into k clusters that maximize the *unobserved* information, I_{un}. Before discussing how to maximize I_{un}, we first consider the problem of estimating it. Similar to the generalization error in supervised learning, I_{un} cannot be calculated directly in the learning algorithm, but we may be able to bound the difference between the observed information I_{ob}—our "training error"—and the unobserved information I_{un}—our "generalization error". To obtain generalization, this bound should be *uniform over all possible clusterings* with a high probability over the randomly selected features. The following lemma argues that *uniform convergence in probability* of I_{ob} to I_{un} always occurs.

3. Since we are exploring an empirical distribution of a finite set of instances, dealing with continuous features is not meaningful.

Lemma 2 *With the definitions above,*

$$\Pr_{\tilde{\mathbf{q}}=(q_1,\ldots,q_n)}\left\{\sup_{\mathbf{t}:[m]\to[k]}|I_{ob}(\mathbf{t},\tilde{\mathbf{q}})-I_{un}(\mathbf{t})|>\varepsilon\right\}\leq 2e^{-2n\varepsilon^2/(\log k)^2+m\log k},\quad\forall\varepsilon>0.$$

Proof For any q,

$$0\leq I(\mathbf{t}(Z);x_q[Z])\leq H(\mathbf{t}(Z))\leq\log k.$$

Using Hoeffding's inequality, for any specific (predetermined) clustering

$$\Pr_{\tilde{\mathbf{q}}=(q_1,\ldots,q_n)}\{|I_{ob}(\mathbf{t},\tilde{\mathbf{q}})-I_{un}(\mathbf{t})|>\varepsilon\}\leq 2e^{-2n\varepsilon^2/(\log k)^2}.$$

Since there are at most k^m possible partitions, the union bound is sufficient to prove Lemma 2.■

Note that for any $\varepsilon>0$, the probability that $|I_{ob}-I_{un}|>\varepsilon$ goes to zero, as $n\to\infty$. The convergence rate of I_{ob} to I_{un} is bounded by $O((\log k)/\sqrt{n})$. As expected, this upper bound decreases as the number of clusters, k, decreases.

Unlike the standard bounds in supervised learning, this bound increases with the number of instances (m), and decreases with increasing numbers of observed features (n). This is because in our scheme the training size is not the number of instances, but rather the number of observed features (see Table 1). However, in the next theorem we obtain an upper bound that is independent of m, and hence is tighter for large m.

Consider the case where n is fixed, and m increases infinitely. We can select a random subset of instances of size m'. For large enough m', the empirical distribution of this subset is similar to the distribution over all instances. By fixing m', we can get a bound which is independent of m. Using this observation, the next theorem gives a bound that is independent of m.

Theorem 3 ***(Information Generalization)*** *With the definitions above,*

$$\Pr_{\tilde{\mathbf{q}}=(q_1,\ldots,q_n)}\left\{\sup_{\mathbf{t}:[m]\to[k]}|I_{ob}(\mathbf{t},\tilde{\mathbf{q}})-I_{un}(\mathbf{t})|>\varepsilon\right\}\leq 8(\log k)e^{-n\varepsilon^2/\left(8(\log k)^2\right)+4sk\log k/\varepsilon-\log\varepsilon},\quad\forall\varepsilon>0.$$

The proof of this theorem is given in appendix A.1. In this theorem, the bound does not depend on the number of instances, but rather on s which is the maximum alphabet size of the features. The convergence rate here is bounded by $O((\log k)/\sqrt[3]{n})$. However, for relatively large n one can use the bound in Lemma 2, which converges faster.

As shown in Table 1, Theorem 3 is clearly analogous to the standard uniform convergence results in supervised learning theory (see, for example, Vapnik, 1998), where the random sample is replaced by our randomly selected features, the hypotheses are replaced by the clustering, and I_{ob} and I_{un} replace the empirical and expected risks, respectively. The complexity of the clustering (our hypothesis class) is controlled by the number of clusters, k.

We can now return to the problem of specifying a clustering that maximizes I_{un}, using only the observed features. For reference, we will first define I_{un} of the best possible clustering.

Definition 4 ***Maximally achievable unobserved information:*** *Let $I^*_{un,k}$ be the maximum value of I_{un} that can be achieved by any partition into k clusters,*

$$I^*_{un,k} = \sup_{\mathbf{t}:[m]\to[k]} I_{un}(\mathbf{t}).$$

The clustering that achieves this value is called ***the best clustering****. The average observed information of this clustering is denoted by $I^*_{ob,k}$.*

Definition 5 ***Observed information maximization algorithm:*** *Let* IobMax *be any clustering algorithm that, based on the values of the observed features, selects a clustering $\mathbf{t}^{opt,ob} : [m] \to [k]$ having the maximum possible value of I_{ob}, that is,*

$$\mathbf{t}^{opt,ob} = \arg\max_{\mathbf{t}:[m]\to[k]} I_{ob}(\mathbf{t}, \tilde{\mathbf{q}}).$$

Let $\tilde{I}_{ob,k}$ be the average observed information of this clustering and $\tilde{I}_{un,k}$ be the expected unobserved information of this clustering, that is,

$$\begin{aligned} \tilde{I}_{ob,k}(\tilde{\mathbf{q}}) &= I_{ob}\left(\mathbf{t}^{opt,ob}, \tilde{\mathbf{q}}\right), \\ \tilde{I}_{un,k}(\tilde{\mathbf{q}}) &= I_{un}\left(\mathbf{t}^{opt,ob}\right). \end{aligned}$$

The next theorem states that *IobMax* not only maximizes I_{ob}, but also maximizes I_{un}.

Theorem 6 ***(Achievability)*** *With the definitions above,*

$$\Pr_{\tilde{\mathbf{q}}=(q_1,\ldots,q_n)}\left\{\tilde{I}_{un,k}(\tilde{\mathbf{q}}) \le I^*_{un,k} - \varepsilon\right\} \le 8(\log k) e^{-n\varepsilon^2/\left(32(\log k)^2\right)+8sk\log k/\varepsilon-\log(\varepsilon/2)}, \quad \forall \varepsilon > 0. \qquad (1)$$

Proof We now define a *bad clustering* as a clustering whose expected unobserved information satisfies $I_{un} \le I^*_{un,k} - \varepsilon$. Using Theorem 3, the probability that $|I_{ob} - I_{un}| > \varepsilon/2$ for any of the clusterings is upper bounded by the right term of Equation 1. If for all clusterings $|I_{ob} - I_{un}| \le \varepsilon/2$, then surely $I^*_{ob,k} \ge I^*_{un,k} - \varepsilon/2$ (see Definition 4) and I_{ob} of all bad clusterings satisfies $I_{ob} \le I^*_{un,k} - \varepsilon/2$. Hence the probability that a bad clustering has a higher average observed information than the best clustering is upper bounded as in Theorem 6. ■

For small m, a tighter bound, similar to that of Lemma 2 can easily be formulated.

As a result of this theorem, when n is large enough, even an algorithm that knows the value of *all* features (observed and unobserved) cannot find a clustering which is significantly better than the clustering found by the *IobMax* algorithm. This is demonstrated empirically in Section 4.

Informally, this theorem means that for a large number of features we can find a clustering that is informative on unobserved features. For example, clustering users based on similar ratings of current movies are likely to match future movies as well (see Section 4).

In the generalization and achievability theorems (Theorems 3, 6) we assumed that we are dealing only with hard clustering. In Appendix A.2 we show that the generalization theorem is also applicable to soft clustering; that is, assigning a probability distribution among the clusters to each instance. Moreover, we show that soft clustering is not required to maximize I_{ob}, since its maximum value can be achieved by hard clustering.

2.1 Toy Examples

In the first two examples below, we assume that the instances are drawn from a given distribution (although this assumption is not necessary for the theorems above). We also assume that the number of instances is large, so the empirical and the actual distributions of the instances are about the same.

Example 1 *Let $X_1,...,X_\infty$ be Bernoulli($\frac{1}{2}$) random variables, such that all variables with an even index are equal to each other ($x_2 = x_4 = x_6 = ...$), and all variables with an odd index are independent of each other and of all other variables. If the number of randomly observed features is large enough we can find a clustering rule with two clusters such that $I_{ob} \cong \frac{1}{2}$. This is done by assigning the cluster labels based on the set of features that are correlated, for example, $\mathbf{t}(i) = x_2[i] + 1 \quad \forall i$, assuming that x_2 is one of the observed features. $I(\mathbf{t}(Z); x_i(Z))$ is one for even i, and zero for odd i. For large n, the number of randomly selected features with odd indices and even indices[4] is about the same (with high probability), and hence $I_{ob} \cong \frac{1}{2}$. For this clustering rule $I_{un} \cong \frac{1}{2}$, since half of the unobserved features match this clustering (all features with an even index).*

Example 2 *When $X_1,...,X_\infty$ are i.i.d. (independent and identically distributed) Bernoulli($\frac{1}{2}$) random variables, $I_{un} = 0$ for any clustering rule, regardless of the number of observed features. For a finite number of clusters, I_{ob} will necessarily approach zero as the number of observed features increases. More specifically, if we use two clusters, where the clustering is determined by one of the observed features (i.e., $\mathbf{t}(i) = x_j(i)$, where x_j is an observed feature), then $I_{ob} = \frac{1}{n}$ (because $I(\mathbf{t}(Z); x_j(Z)) = 1$ and $I(\mathbf{t}(Z); x_l(Z)) = 0$ for $l \neq j$).*

Example 3 *Clustering fruits based on the observed features (color, size, shape etc.) also matches many unobserved features. Indeed, people clustered fruits into oranges, bananas and others (by giving names in the language) long before vitamin C was discovered. Nevertheless, this clustering was very informative about the amount of vitamin C in fruits, that is, most oranges have similar amounts of vitamin C, which is different from the amount in bananas.*

Based on the generalization theorem, we now suggest a qualitative explanation of why clustering into bananas and oranges provides relatively high information on unobserved features, while clustering based on position (e.g., right/left in the field of view) does not. Clustering into bananas and oranges contains information on many observed features (size, shape, color, texture), and thus has relatively large I_{ob}. By the generalization theorem, this implies that it also has high I_{un}. By contrast, a clustering rule which puts all items that appeared in our right visual field in one cluster, and the others in a second cluster, has much smaller I_{ob} (since it does not match many observed features), and indeed it is not predictive about unobserved features.

Example 4 *As a negative example, if the type of observed features and the target unobserved features are very different, our assumptions do not hold. For example, when the observations are pixels of an image, and the target variable is the label of the image, we cannot generalize from information about the pixels to information about the label.*

2.2 Feature Generalization of Maximum Likelihood Multinomial Mixture Models

In the framework of Bayesian graphical models, the multinomial mixture model is commonly used. The assumption of this model is that all features are conditionally independent, given the value of

4. Note that the indices are arbitrary. The learning algorithm does not use the indices of the features.

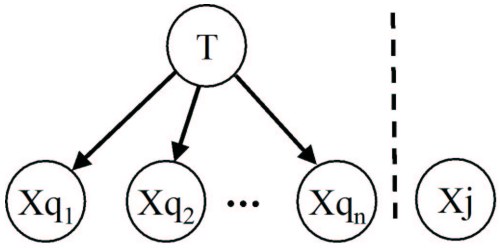


Figure 2: The Bayesian network (Pearl, 1988) of the multinomial mixture model. The observed random variables $\{X_{q_1},\ldots,X_{q_n}\}$ are statistically independent given the parent hidden random variable, T. This parent variable represents the cluster label. Although the basic assumptions of the multinomial mixture model are very different from ours, Theorem 7 tells us that this method of clustering generalizes well to unobserved features.

some hidden variable, that is,

$$\Pr(T=t,x_{q_1},\ldots,x_{q_n}) = \Pr(T=t)\prod_{r=1}^{n}\Pr(x_{q_r}|T=t),$$

where T denotes the hidden variable. The Bayesian network (Pearl, 1988) of this model is given in Figure 2. This standard model does not assume the existence of unobserved features, so we use the notation $x_{q_1},\ldots,x_{q_n}$ to denote the observed features which are used by the model. The set of instances are assumed to be drawn from such a distribution, with unknown parameters. Given the set of instances, the goal is to learn the distributions $\Pr(T=t)$ and $\Pr(x_{q_r}|T=t)$ that maximizes the probability of the observation, that is, values of the instances. This maximum-likelihood problem is typically solved using an EM algorithm (Dempster et al., 1977) with a fixed number of clusters (values of T). The output of this algorithm includes a soft clustering of all instances; that is, $P(T|Y)$, where Y denotes the index of the instance.

In the following theorem we analyze the feature generalization properties of soft clustering by the multinomial mixture model. We show that under some technical conditions, it pursues nearly the same goal as *IobMax* algorithm (Definition 5), that is, maximizing $\sum_j I\left(\mathbf{t}(Z);X_{q_j}(Z)\right)$.

Theorem 7 *Let $I_{ob,ML,k}$ be the observed information of clustering achieved by the maximum likelihood solution of a multinomial mixture model for k clusters. Then*

$$I_{ob,ML,k} \geq \tilde{I}_{ob,k} - \frac{2H(T)}{n},$$

where $\tilde{I}_{ob,k}$ is the observed information achieved by the IobMax clustering algorithm (Definition 5).

Proof

Elidan and Friedman (2003) showed that learning a hidden variable can be formulated as the multivariate information bottleneck (Friedman et al., 2001). Based on their work, in Appendix A.3 we show that maximizing the likelihood of observed variables is equivalent to maximizing $\sum_{j=1}^{n} I(T;X_{q_j}) - I(T;Y)$. Using our notations, this is equivalent to maximizing $I_{ob} - \frac{1}{n}I(T;Y)$. Since $I(T;Y) \leq H(T)$, the difference between maximizing I_{ob} and $I_{ob} - \frac{1}{n}I(T;Y)$ is at most $2H(T)/n$. ■

The meaning of Theorem 7 is that for large n, finding the maximum likelihood of mixture models is similar to finding the maximum unobserved information. Thus the standard EM-algorithm for maximum likelihood of mixture models can be viewed as a form of the *IobMax* algorithm.[5]

The standard mixture model assumes a generative model for generating the instances from some distribution, and finds the maximum likelihood of this model. This model does not assume anything about the selection of features or the existence of unobserved features. Our setup assumes that the instances are fixed and the observed features are randomly selected and we try to maximize information on unobserved features. Interestingly, while the initial assumptions are quite different, the results are nearly equivalent. We show that finding the maximum likelihood of the mixture model indirectly predicts unobserved features as well.

The maximum likelihood mixture model was used by Breese et al. (1998) to cluster users by their voting on movies. This clustering is used to predict the rating of new movies. Our analysis shows that for a large number of rated (observed) movies, it is nearly the best clustering method in terms of information on new movies.

The multinomial mixture model is also used for learning with labeled and unlabeled instances, and is considered a baseline method (see Section 2.3 in Seeger, 2002). The idea is to cluster the instances based on their features. Then, the prediction of a label for an unlabeled instance is estimated from the labels of other instances in the same cluster. From our analysis, this is nearly the best clustering method for preserving information on the label, assuming that the label is yet another feature that happened to be unobserved in some instances. This provides another interpretation regrading the hidden assumption of this clustering scheme for labeled and unlabeled data.

3. Distance-Based Clustering

In this section we extend the framework and include analysis of feature generalization bounds for distance-based clustering. We apply this to analyze feature generalization of the k-means clustering algorithm (See Table 1). The setup in this section is the same as the setup defined in Section 2 except as described below. We assume that we have a distance function, denoted by f, that measures the distance for every two values of any of the features. We assume that f has the following properties:

$$0 \leq f(x_q[j], x_q[l]) \leq c \quad \forall q, j, l, \tag{2}$$

$$f(a,a) = 0 \qquad \forall a, \tag{3}$$

$$f(a,b) = f(b,a) \qquad \forall a, b, \tag{4}$$

where c is some positive constant. An example of such a function is the square error, that is, $f(a,b) = (a-b)^2$, where we assume that the value of all features is bounded as follows $\left|x_q[j]\right| \leq \sqrt{c}/2$ ($\forall q, j$), for some constant c. The features themselves can be discrete or continuous. Although we do not directly use the function f in the definitions of the theorems in the following section, it is required for their proofs (Appendix A.4).

5. Ignoring the fact that achieving a global maximum is not guaranteed.

3.1 Generalization of Distance-Based Clustering

As in Section 2, we have a set of m fixed instances $\{\mathbf{x}[1],...,\mathbf{x}[m]\}$, and the clustering algorithm clusters these instances into k clusters. For better readability, in this section the partition is denoted by $\{C_1,...,C_k\}$. $|C_r|$ denotes the size of the rth cluster.

The standard objective of the k-means algorithm is to achieve minimum intra-cluster variance, that is, minimize the function

$$\sum_{r=1}^{k}\sum_{j\in C_r}|\mathbf{x}[j]-\mu_r|^2,$$

where μ_r is the mean point of all instances in the rth cluster.

In our setup, however, we assume that the clustering algorithm has access only to the *observed* features over the m instances. The goal of clustering is to achieve minimum intra-cluster variance of the *unobserved* features. To do so, we need to generalize from the observed to the unobserved intra-class variance. To formalize this type of generalization, let's first define these variances formally.

Definition 8 *The **observed intra-cluster variance** $D_{ob}\{C_1,...,C_k\}$ of a clustering $\{C_1,...,C_k\}$ is defined by*

$$D_{ob}\{C_1,...,C_k\}=\frac{1}{nm}\sum_{r=1}^{k}\sum_{j\in C_r}\sum_{i=1}^{n}\left(x_{q_i}[j]-\mu_{q_i}[r]\right)^2,$$

where $\mu_q[r]$ is the mean of feature q over all instances in cluster r, that is,

$$\mu_q[r]=\frac{1}{|C_r|}\sum_{l\in C_r}x_q[l].$$

In other words, D_{ob} is the average square distance of each observed feature from the mean of the value of the feature in its cluster. The average is over all observed features and instances. The k-means algorithm minimizes the observed intra-cluster variance.

Definition 9 *The **expected unobserved intra-cluster variance** $D_{un}\{C_1,...,C_k\}$ is defined by*

$$D_{un}\{C_1,...,C_k\}=\frac{1}{m}\sum_{r=1}^{k}\sum_{j\in C_r}\mathbf{E}_{q\sim\mathcal{D}}\left(x_q[j]-\mu_q[r]\right)^2.$$

D_{ob} and D_{un} are the distance-based variables analogous to I_{ob} and I_{un} defined in Section 2. In our setup, the goal of the clustering algorithm is to create clusters with minimal unobserved intra-class variance (D_{un}). As in the case of information-based clustering, we first consider the problem of estimating D_{un}. Before presenting the generalization theorem for distance-based clustering, we need the following definition.

Definition 10 *Let α be the ratio between the size of smallest cluster and the average cluster size, that is,*

$$\alpha(\{C_1,...,C_k\})=\frac{\min_r|C_r|}{m/k}.$$

Now we are ready for the generalization theorem for distance-based clustering.

Theorem 11 *With the above definitions, if* $\left|x_q[j]\right| \leq R$ *for every* q, j *then for every* $\alpha_c > 0$, $\varepsilon > 0$,

$$\Pr_{\{q_1,\ldots,q_n\}} \left\{ \sup_{\alpha(\{C_1,\ldots,C_k\})\geq\alpha_c} \left| D_{ob}\{C_1,\ldots,C_k\} - D_{un}\{C_1,\ldots,C_k\} \right| \leq \varepsilon \right\} \geq 1-\delta,$$

where

$$\delta = \frac{8k}{\alpha_c} e^{-n\varepsilon^2/8R^4 + \log\left(R^2/\varepsilon\right)}.$$

The proof of this theorem is given in Appendix A.4. Theorem 11 is a special case of a more general theorem (Theorem 14) that we present in the appendix. Theorem 14 can be applied to other distance-based metrics, beyond the intra-cluster variance defined in Definition 9.

Note that for any $\varepsilon > 0$, the probability that $|D_{ob} - D_{un}| \leq \varepsilon$ goes to one, as $n \to \infty$. The convergence rate of D_{ob} to D_{un} is bounded by $O(1/\sqrt{n})$. As expected, for a fixed value of δ the upper bound on $|D_{ob} - D_{un}|$ decreases as the number of clusters, k, decreases.

Theorem 11 bounds the difference between observed and unobserved variances. We now use it to find a clustering that minimizes the expected unobserved intra-cluster variance, using only the observed features.

Theorem 12 *Let* $\left\{C_1^{opt},\ldots,C_k^{opt}\right\}$ *be the clustering that achieves the minimum unobserved intra-cluster variance under the constraint* $\alpha(\{C_1,\ldots,C_k\}) \geq \alpha_c$ *for some constant* $0 < \alpha_c \leq 1$, *that is,*

$$\left\{C_1^{opt},\ldots,C_k^{opt}\right\} = \arg \min_{\{C_1,\ldots,C_k\}:\alpha\geq\alpha_c} D_{un}\{C_1,\ldots,C_k\},$$

and let D_{un}^{opt} *the best unobserved intra-cluster variance, be defined by* $D_{un}^{opt} = D_{un}\left\{C_1^{opt},\ldots,C_k^{opt}\right\}$.

Let $\left\{\hat{C}_1^{opt},\ldots,\hat{C}_k^{opt}\right\}$ *be the clustering with the minimum observed intra-cluster variance, under the same constraint on* $\alpha(\{C_1,\ldots,C_k\})$, *that is,*

$$\left\{\hat{C}_1^{opt},\ldots,\hat{C}_k^{opt}\right\} = \arg \min_{\alpha(\{C_1,\ldots,C_k\})\geq\alpha_c} D_{ob}\{C_1,\ldots,C_k\},$$

and let $\hat{D}_{un}^{opt}$ *be the unobserved intra-cluster variance of this clustering, that is,* $\hat{D}_{un}^{opt} = D_{un}\left\{\hat{C}_1^{opt},\ldots,\hat{C}_k^{opt}\right\}$.

For any $\varepsilon > 0$,

$$\Pr_{\{q_1,\ldots,q_n\}} \left\{\hat{D}_{un}^{opt} \leq D_{un}^{opt} + \varepsilon\right\} \geq 1-\delta,$$

where

$$\delta = \frac{16k}{\alpha_c} e^{-n\varepsilon^2/32R^4 + \log\left(R^2/\varepsilon\right)}. \tag{5}$$

Proof We now define a *bad clustering* as a clustering whose expected unobserved intra-cluster variance satisfies $D_{un} > D_{un}^{opt} + \varepsilon$. Using Theorem 11, the probability that $|D_{ob} - D_{un}| \leq \varepsilon/2$ for all possible clusterings (under the constraint on α) is at least $1-\delta$, where δ defined in Equation 5. If for all clusterings $|D_{ob} - D_{un}| \leq \varepsilon/2$, then surely $D_{ob}\left\{C_1^{opt},\ldots,C_k^{opt}\right\} \leq D_{un}^{opt} + \varepsilon/2$ and D_{ob} of all bad clusterings satisfies $D_{ob} > D_{un}^{opt} + \varepsilon/2$. Hence the probability that any of the bad clusterings has a lower observed intra-cluster variance than the best clustering is upper bounded by δ. Therefore, with a probability of at least $1-\delta$ none of the bad clusterings is selected by an algorithm that selects

the clustering with the minimum D_{ob}. ■

We cannot directly calculate the unobserved intra-cluster variance. However, Theorem 12 means that an algorithm that selects the clustering with the minimum observed intra-cluster variance indirectly finds the clustering with nearly minimum unobserved intra-cluster variance.

In general, minimizing observed intra-cluster variance is the optimization objective of *k*-means. Hence, *k*-means indirectly minimizes the unobserved intra-cluster variance. This means that in our context, *k*-means can be viewed as an analog to the empirical risk minimization (ERM) in the standard supervised learning context. We minimize the observed variance (training error) in order to indirectly minimize the expected unobserved variance (test error).

k-means is used in collaborative filtering such as movie rating predictions for grouping users based on similar ratings (see, for example, Marlin, 2004). After clustering, we can predict ratings of a new movie based on the ratings of a few users for this movie. If the intra-cluster variance of a new, previously unobserved movie is small, then we can estimate the rating of one user from the average ratings of other users in the same cluster.

An experimental illustration of the behavior of the observed and unobserved intra-cluster variances for *k*-means is available in Section 4.1.

4. Empirical Evaluation

In this section we test experimentally the generalization properties of *IobMax* and the *k*-means clustering algorithm for a finite number of features. For *IobMax* we examine the difference between I_{ob} and I_{un} as a function of the number of observed features, and number of clusters used. We also compare the value of I_{un} achieved by the *IobMax* algorithm to I^*_{un}, which is the maximum achievable I_{un} (see Definition 4). Similarly, for distance-based clustering we use *k*-means to examine the behavior of the observed and unobserved intra-cluster variances (see Definitions 8, 9).

The purpose of this section is *not* to suggest new algorithms for collaborative filtering or compare it to other methods, but simply to illustrate our new theorems on empirical data.

4.1 Collaborative Filtering

In this section, our evaluation uses a data set typically employed for collaborative filtering. Collaborative filtering refers to methods of making predictions about a user's preferences, by collecting the preferences of many users. For example, collaborative filtering for movie ratings can make predictions about the rating of movies by a user given a partial list of ratings from this user and many other users. Clustering methods are used for collaborative filtering by clustering users based on the similarity of their ratings (see, for example, Marlin, 2004; Ungar and Foster, 1998).

In our setting, each user is described as a vector of movie ratings. The rating of each movie is regarded as a feature. We cluster users based on the set of observed features, that is, rated movies. In our context, the goal of the clustering is to maximize the information between the clusters and unobserved features, that is, movies that have not yet been rated by any of the users. These can be movies that have not yet been made. By Theorem 6, given a large enough number of rated movies, we can achieve the best possible clustering of users with respect to unseen movies. In this region, no additional information (such as user age, taste, rating of more movies) beyond the observed features can improve the unobserved information, I_{un}, by more than some small ε.

For distance-based clustering, we cluster the users by the *k*-means algorithm based on a subset of features (movies). As we show in Section 3.1 the goal of *k*-means is to minimize the observed intra-cluster variance. From Theorem 12, this indirectly minimizes the unobserved intra-cluster variance as well. Here we empirically evaluate this type of generalization.

Data set. We use MovieLens (www.movielens.umn.edu), which is a movie rating data set. It was collected and distributed by GroupLens Research at the University of Minnesota. It contains approximately 1 million ratings of 3900 movies by 6040 users. Ratings are on a scale of 1 to 5. We use only a subset consisting of 2400 movies rated by 4000 users (or 2000 by 2000 for distance-based clustering). In our setting, each instance is a vector of ratings $(x_1,\ldots,x_{2400})$ by a specific user. Each movie is viewed as a feature, where the rating is the value of the feature.

Experimental Setup. We randomly split the 2400 movies into two groups, denoted by "A" and "B", of 1200 movies (features) each. We use a subset of the movies from group "A" as observed features and all movies from group "B" as the unobserved features. The experiment was repeated with 20 random splits and the results averaged. We estimate I_{un} by the mean information between the clusters and ratings of movies from group "B". We use a uniform distribution of feature selection ($\mathcal{D}$), and hence I_{un} can be estimated as the average information on the unobserved features, that is, $I_{un} = \frac{1}{1200}\sum_{j\in B} I(T;X_j)$. A similar setup is used for the distance-based clustering (with two groups of 1000 movies).

Handling Missing Values. In this data set, most of the values are missing (not rated). For information based-clustering, we handle this by defining the feature variable as 1,2,...,5 for the ratings and 0 for a missing value. We maximize the mutual information based on the empirical distribution of values that are present, and weight it by the probability of presence for this feature. Hence, $I_{ob} = \sum_{j=1}^{n} P(X_j \neq 0) I(T;X_j|X_j \neq 0)$ and $I_{un} = E_j\left\{P(X_j \neq 0) I(T;X_j|X_j \neq 0)\right\}$. The weighting prevents overfitting to movies with few ratings. Since the observed features are selected at random, the statistics of missing values of the observed and unobserved features are the same. Hence, all our theorems are applicable to these definitions of I_{ob} and I_{un} as well.

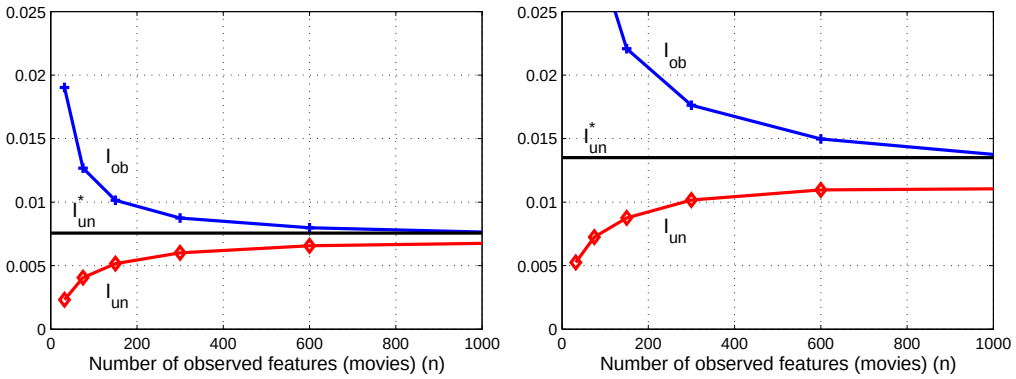
In order to verify that the estimated mutual information is not just an artifact of the finite sample size, we tested the mutual information after random permutation of ratings of each movie among users. Indeed, the resulting mutual information was significantly lower in the case of random permutation.

For the distance based clustering, we handle missing data by defining a default square distance between a feature and the cluster center where one (or two) of the values is missing. We select this default square distance to be the average variance of movie ratings (which is about 0.9).

4.2 Greedy *IobMax* Algorithm

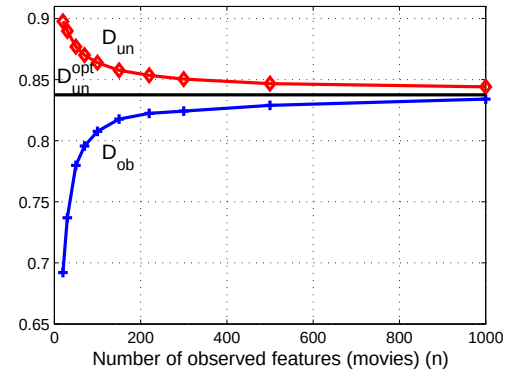
For information-based clustering, we cluster the users using a simple greedy clustering algorithm (see Algorithm 1). The input to the algorithm is all users, represented solely by the observed features. Since this algorithm can only find a local maximum of I_{ob}, we ran the algorithm 10 times (each used a different random initialization) and selected the results that had a maximum value of I_{ob}.

In our experiment, the number of observed features is large. Therefore, based on Theorem 7, the greedy *IobMax* can be replaced by the standard EM-algorithm which finds the maximum likelihood for multinomial mixture models. Although, in general, this algorithm finds soft clustering, in our

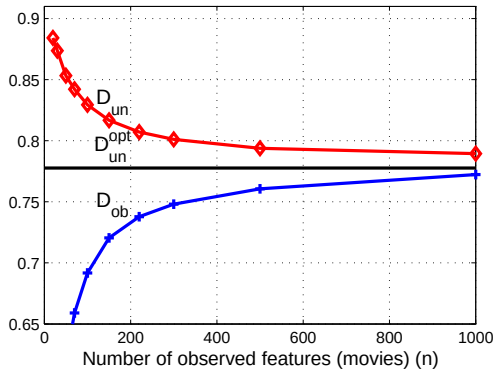


(a) 2 Clusters

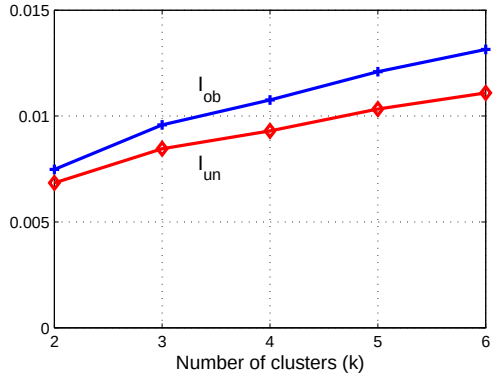
(b) 6 Clusters



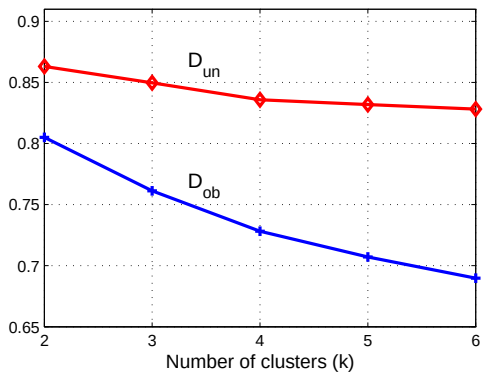
(c) 2 Clusters



(d) 6 Clusters



(e) Fixed number of features (n=1200)



(f) Fixed number of features (n=100)

Figure 3: Feature generalization as a function of the number of training features (movies) and the number of clusters. (a) (b) and (e) show the observed and unobserved information for various numbers of features and clusters (high is good). The overall mean information is low, since the rating matrix is sparse. (c) (d) and (f) shows the observed and unobserved intra-cluster variance (low is good). In these figures, the variance is only calculated on values which are not missing. Figures (e) and (f) show the effect of the number of clusters when the number of features is fixed.

case the empirical result clusterings are not soft, that is, one cluster is assigned to each instance (see Appendix A.2). As expected, the results of both algorithms are nearly the same.

Algorithm 1 A simple greedy *IobMax* algorithm

1. Assign a random cluster to each of the instances.
2. For $r = 1$ to R (where R is the upper limit on the number of iterations)
   (a) For each instance,
      i. Calculate I_{ob} for all possible clustering assignments of the current instance.
      ii. Choose the clustering that maximizes I_{ob}.
   (b) Exit if the clusters of all documents do not change.

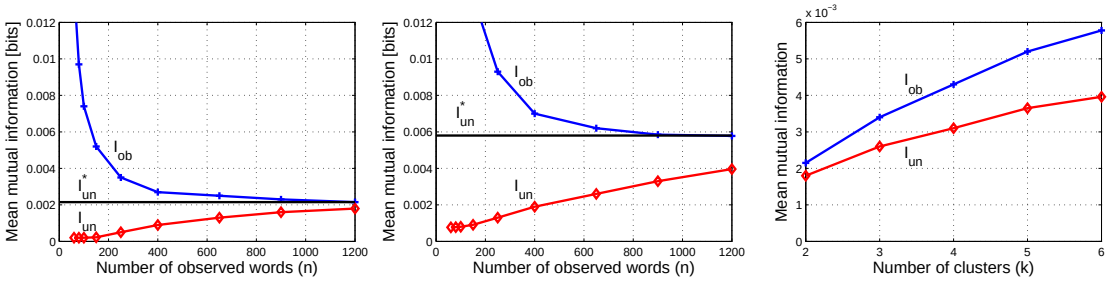
In order to estimate $I^*_{un,D}$ (see Definition 4), we also ran the same algorithm when all the features were available to the algorithm (i.e., also features from group "B"). In this case the algorithm tries directly to find the clustering that maximizes the mean mutual information on features from group "B".

4.3 Results

The results are shown in Figure 3. It is clear that as the number of observed features increases, I_{ob} decreases while I_{un} increases (see Figure 3(a,b)). When there is only one feature, two clusters can contain all the available information on this feature (e.g., by assigning $t(j) = x_{q_1}[j]$), so I_{ob} reaches its maximum value (which is $H(X_{q_1}[Z])$). As the number of observed features increases, we cannot preserve all the information on all the features in a few clusters, so the observed mutual information (I_{ob}) decreases. On the other hand, as the number of observed features increases, the cluster variable, $T = t(Z)$, captures the structure of the distribution (users' tastes), and hence contains more information on unobserved features. The generalization theorem (Theorem 3) tells us that the difference between I_{un} and I_{ob} will approach zero as the number of observed features increases. This is similar to the behavior of training and test errors in supervised learning. Informally, the achievability theorem (Theorem 6) tells as that for a large enough number of observed features, even though our clustering algorithm is based only on observed features, it can achieve nearly the best possible clustering, in terms of I_{un}. This can be seen in Figures 3 (a,b), where I_{un} approaches I^*_{un}, which is the unobserved information of the best clustering (Definition 4). As the number of clusters increases, both I_{ob}, I_{un} increase (Figure 3e), but the difference between them also increases.

Similar results were obtained for distance based clustering. The goal here is to minimize the unobserved intra-cluster variance (D_{un}), and this is done by minimizing the observed intra-cluster variance (D_{ob}). As discussed in Section 3.1, this can be achieved by k-means.[6] Again, for a small numbers of features (n) the clustering overfits the observed features, that is, the D_{ob} is relatively low but D_{un} is large. However, for large n, D_{un} and D_{ob} approach each other and both of them approach the unobserved intra-cluster variance of the best possible clustering (D^{opt}_{un}) as expected

6. Since k-means does not necessarily find the global optimum, we ran it 20 times with different initialization points, and chose the results with minimal observed intra-cluster variance. This does not guarantee a global optimum, but no other tractable algorithm is available today to achieve global optima.



(a) 2 Clusters (b) 6 Clusters (c) Fixed 1200 observed words

Figure 4: I_{ob}, I_{un} and I^*_{un} per number of training words and clusters. In (a) and (b) the number of words is variable, and the number of clusters is fixed. In (c) the number of observed words is fixed (1200), and the number of clusters is variable. The overall mean information is low, since a relatively small number of words contributes to the information (see Table 2)

from Theorem 12. When the number of clusters increases, both D_{ob} and D_{un} decrease, but the difference between them increases.

4.4 Words and Documents

In this section we repeat the information-based clustering experiment, but this time for document clustering with words as features. We show how clustering which is based on a subset of words (observed words) is also informative about the unobserved words. The obtained curves of information vs. number of features are similar to those in the previous section. However, in this section we also examine the resulting clustering (Table 2) to get a better intuition as to how this generalization occurs.

Data set. We use the 20-newsgroups (20NG) corpus, collected by Lang (1995). This collection contains about 20,000 messages from 20 Usenet discussion groups, some of which have similar topics.

Preprocessing. In order to prevent effects caused by different document lengths, we truncate each document to 100 words (by randomly selecting 100 words), and ignore documents which consist of fewer than 100 words. We use the "bag of words" representation: namely we convert each document into a binary vector $(x_1, x_2, \ldots)$, where each element in the vector represents a word, and equals one if the word appears in the document and zero otherwise. We select the 2400 words whose corresponding X_i has maximum entropy,[7] and remove all other words. After this preprocessing each document is represented by a vector $(x_1, \ldots, x_{2400})$.

Experimental setup. We randomly split the 2400 words into two groups of 1200 words (features) each. The groups were called "A" and "B". We use a variable number of words (1 to 1200) from group "A" as observed features. All the features from group "B" are used as the unobserved features. We repeat the test with 10 random splits and present the mean results.

7. In other words, the probability of the word appearing in a document is not near zero or near one.

	Word	t=1	t=2	t=3
Observed words	he	0.47	0.04	0.25
	game	0.23	0.01	0
	team	0.20	0	0
	x	0.01	0.15	0.01
	hockey	0.11	0	0
	jesus	0.01	0	0.09
	christian	0	0	0.08
	use	0.04	0.21	0.09
	file	0	0.08	0.01
Unobserved words	god	0.02	0.01	0.15
	players	0.13	0	0
	baseball	0.10	0	0
	window	0	0.10	0
	server	0	0.06	0

Table 2: Probability of a word appearing in a document from each cluster. Each column in the table represents a cluster (total of three clusters), and the numbers are the probabilities that a document from a cluster will contain the word (e.g., The word "he" appears in 47% of the documents from cluster 1). The results presented here are for learning from 1200 observed words, but only a few of the most informative words appear in the table.

4.5 Results

The results are shown in Figure 4 and Table 2. The qualitative explanation of the figure is the same as for collaborative filtering (see Section 4.1 and Figure 3). Table 2 presents a list of the most informative words, and their appearance in each cluster. This helps understand the way clustering learned from observed words matches unobserved words. We can see, for example, that although the word "player" is not part of the inputs to the clustering algorithm, it appears much more in the first cluster than in other clusters. Intuitively this can be explained as follows. The algorithm finds clusters that are informative on many observed words together, and thus matches the co-occurrence of words. This clustering reveals the hidden topics of the documents (sports, computers and religious), and these topics contain information on the unobserved words. We see that generalization to unobserved features can be explained from a standpoint of a generative model (a hidden variable which represents the topics of the documents) or from a statistical point of view (relationship between observed and unobserved information). In Section 6 we further discuss this dual view.

5. Related Work

In the framework of learning with labeled and unlabeled data (see, for example, Seeger, 2002), a fundamental issue is the link between the marginal distribution $P(\mathbf{x})$ over instances $\mathbf{x}$ and the conditional $P(y|\mathbf{x})$ for the label y (Szummer and Jaakkola, 2003). From this point of view our approach assumes that the label, y, is a feature in itself.

In the context of supervised learning, Ando and Zhang (2005) proposed an approach that regards features in the input data as labels in order to improve prediction in the target supervised problem. Their idea is to create many auxiliary problems that are related to the target supervised problem. They do this by masking some features in the input data, that is, making them unobserved features, and training classifiers to predict these unobserved features from the observed features. Then they transfer the knowledge acquired from these related classification tasks to the target supervised problem. A similar idea was used by Caruana and de Sa (1997) in supervised training of a neural net. The authors used some features as extra outputs of the neural net, rather than inputs, and show empirically that this can improve the classifier performance. In our framework, this could be interpreted as follows. We regard the label as a feature, and hence we can learn from prediction of these features to the prediction of the label. Loosely speaking, if we successfully predict many such features by the classifier, we expect to generalize better to the target feature (label).

The idea of an information tradeoff between complexity and information on target variables is similar to the idea of the information bottleneck (Tishby et al., 1999). But unlike the bottleneck method, here we are trying to maximize information on *unobserved* variables, using a finite sample.

In a recent paper, von Luxburg and Ben-David (2005) discuss the goal of clustering in two very different cases. The first is when we have complete knowledge about our data generating process, and the second is how to approximate an optimal clustering when we have incomplete knowledge about our data. In most current analyses of clustering methods, incomplete knowledge refers to getting a finite sample of instances rather than the distribution itself. Then, we can define the desired properties of a good clustering. An example of such a property is the stability of the clustering with respect to the sampling process, for example, the clusters do not change significantly if we add some data points to our sample. In our framework, even if the distribution of the instances is completely known, we assume that there are other features that we might not be aware of at the time of clustering. Another way to view this is that in our framework, incomplete knowledge refers to the existence of unobserved features rather than to an unknown distribution of the observed features. From this point of view, further research could concentrate on analyzing the feature stability of a clustering algorithm, for example, stability with respect to the adding of new features.

Another interesting work which addresses the difficulty of defining good clustering was presented by Kleinberg (2002). In this work the author states the desired properties a clustering algorithm should satisfy, such as scale invariances and richness of possible clusterings. Then he proves that it is impossible to construct a clustering that satisfies all the required properties. In his work the clustering depends on pairwise distances between data points. In our work, however, the analysis is feature oriented. We are interested in the information (or distance) per feature. Hence, our basic assumptions and analysis are very different.

The idea of generalization to unobserved features by clustering was first presented in a short version of this paper (Krupka and Tishby, 2005).

6. Discussion

We introduce a new learning paradigm: clustering based on observed features that generalizes to unobserved features. Our main results include two theorems that tell us how, without knowing the value of the unobserved features, one can estimate and maximize information between the clusters and the unobserved features. Using this framework we analyze feature generalization of the Maximum Likelihood Multinomial Mixture Model (Figure 2). The multinomial mixture model is a

generative probabilistic model which approximates the probability distribution of the observed data ($\mathbf{x}$), from a finite sample of instances. Our model does not assume any distribution that generated the instances, but instead assumes that the set of observed features is simply a random subset of features. Then, using statistical arguments we show that we can cluster by the unobserved features. Despite the very different assumptions of these models, we show that clustering by multinomial mixture models is nearly optimal in terms of maximizing information on unobserved features. However, to analyze and quantify this generalization our framework is required.

This dual view on the multinomial mixture model can also be applied to two different approaches that may explain our "natural clustering" of objects in the world (e.g., assigning object names in language). Let's return to our clustering of bananas and oranges (Example 3). From the generative point of view, we find a model with the cluster labels bananas and oranges as values of a hidden variable that created the distribution. This means that we have a mixture of two distributions, each related to one object type that is assigned to a different cluster. Since we have two types of objects (distributions), we expect that their unobserved features will correspond to these two types as well. However, the generative model does not quantify this expectation. In our framework, we view fruits in the world, and cluster them based on some kind of *IobMax* algorithm; that is, we find a representation (clustering) that contains significant information on as many observed features as possible, while still remaining simple. From our generalization theorem (Theorem 3), such a representation is expected to contain information on other rarely viewed salient features as well. Moreover, we expect this unobserved information to be similar to the information we have on the clustering on the observed features.

In addition to information-based clustering, we present similar generalization theorems for distance-based clustering, and use these to analyze generalization properties of *k*-means. Under some assumptions, *k*-means is also known as a solution for the maximum likelihood Gaussian mixture model. Analogous to what we show for information based clustering and multinomial mixture models, we show that this optimization goal of *k*-means is also optimal in terms of generalization to unobserved features.

The key assumption that enables us to prove these theorems is the *random independent selection* of the observed features. Note that a contrary assumption to random selection would be that given two instances $\{\mathbf{x}[1],\mathbf{x}[2]\}$, there is a correlation between the distance of a feature $\left|x_q[1]-x_q[2]\right|$ and the probability of observing this feature; for example, the probability of *observing* features that are similar is higher. If no such correlation exists, then the selection can be considered random in our context. Hence, we believe that in practice the random selection assumption is reasonable. However, in many cases, the assumption of complete independence in the selection of features is less natural. Therefore, we believe that further research on the effects of dependence in selection is required.

Another interpretation of the generalization theorem, without using the *random independence* assumption, might be combinatorial. The difference between the observed and unobserved information is large only for a small portion of all possible partitions into observed and unobserved features. This means that almost any arbitrary partition generalizes well.

The value of clustering which preserves information on unobserved features is that it enables us to learn new—previously unobserved—attributes from a small number of examples. Suppose that after clustering fruits based on their observed features (Example 3), we eat a chinaberry[8] and thus,

8. Chinaberries are the fruits of the Melia azedarach tree, and are poisonous.

we "observe" (by getting sick), the previously unobserved attribute of toxicity. Assuming that in each cluster, all fruits have similar unobserved attributes, we can conclude that all the fruits in the same cluster, that is, all chinaberries are likely to be poisonous.

Clustering is often used in scientific research, when collecting measurements on objects such as stars or neurons. In general, the quality of a theory in science is measured by its predictive power. Therefore, a reasonable measure of the quality of clustering, as used in scientific research, is its ability to predict unobserved features, or measurements. This is different from clustering that merely describes the observed measurements, and supports the rationale for defining the quality of clustering by its predictivity on unobserved features.

6.1 Further Research

Our clustering maximizes expected information on randomly selected features. Although on average this information may be high, there might be features the clustering has no information about. To address this problem, we could create more than one clustering, in such a way that each clustering contains information on other features. To achieve this, we want each new clustering to discover new information, that is, not to be redundant with previously created clusterings. This can be done based on the works of Gondek and Hofmann (2004) and Chechik and Tishby (2002) in the context of the Information Bottleneck. Another alternative is to represent each instance by a low dimensional vector, and then use this vector to predict unobserved features. Blitzer et al. (2005) represented words in a model called Distributed Binary Latent Variables, and used this representation to predict another word. Adopting this idea in our context, we can replace cluster labels by a vector of binary variables assigned to each instance, where each such variable encodes an independent aspect of the instance. Generalization in this case refers to the ability to predict unobserved features from these latent variables.

Our framework can also be extended beyond clustering by formulating a general question. Given the (empirical) marginal distribution of a random subset of features $P(X_{q_1},\ldots,X_{q_n})$, what can we say about the distribution of the full set $P(X_1,\ldots,X_L)$? In this paper we proposed a clustering based on a subset of features, and analyzed the information that the clustering yielded on features outside this subset. It would be useful to find more sophisticated representations than clustering, and analyze other theoretical aspects of the relationship between the distribution of the subset to that of the full set. This type of theoretical analysis can help in deriving prediction algorithms, where there are many instances for some of the variables (features), but other variables are rarely viewed, as in collaborative filtering. By relating the distribution of some variables to the distribution of others, we can also analyze and improve the estimation of $p(\mathbf{x})$ from a finite sample, even without assuming the existence of unobserved features. In a different context, Han (1978) analyzed the relationship between the average (per variable) entropy of random subsets of variables. He showed that the average entropy of a random subset of variables monotonically decreases with the size of the subset (see also Cover and Thomas, 1991). These results were developed in the context of information theory and compression, but may be applicable to learning theory as well.

In this paper, we assumed that we do not have additional prior information on the features. In practice, we often do have such information. For instance, in the movie ratings data set, we have some knowledge about each of the movies (genre, actors, year, etc.). This knowledge about the features can be regarded as meta-features. A possible extension of our framework is to use this knowledge to improve and analyze generalization as a function of the meta-features of the

unobserved features. The idea of learning along the features axis by using meta-features was implemented by Krupka et al. (submitted) for feature selection. They propose a method for learning to select features based on the meta-features. Using the meta-features we can learn what *types* of features are good, and predict the quality of unobserved features. They show that this is useful for feature selection out of a huge set of potentially extracted features; that is, features that are functions of the input variables. In this case all features can be observed, but in order to measure their quality we must calculate them for all instances, which might be computationally intractable. By predicting feature quality without calculating it, we can focus the search for good features on a small subset of the features. In a recent paper (Krupka and Tishby, 2007), we propose a method for learning the weights of a linear classifier based on meta-features. The idea is to learn weights as a function of the meta-features just as we learn labels as a function of features. Then, we can learn from feature to feature and not only from instance to instance. As shown empirically, this can significantly improve classification performance in the standard supervised learning setting.

In this work we focused on a new feature generalization analysis. Another research direction is to combine standard instance generalization with feature generalization. In problems like collaborative filtering or gene expression, there is an inherent symmetry between features and instances that have been used before in various ways (see, for example, Ungar and Foster, 1998). In the context of supervised learning, a recent work by Globerson and Roweis (2006) addresses the issue of handling differences in the set of observed features between training and test time. However, a general framework for generalization to both unobserved features and unobserved instances is still lacking.

Acknowledgments

We thank Aharon Bar-Hillel, Amir Globerson, Ran Bachrach, Amir Navot and Ohad Shamir for helpful discussions. We also thank the GroupLens Research Group at the University of Minnesota for use of the MovieLens data set. Our work is partly supported by the Center of Excellence grant from the Israeli Academy of Science and by the NATO SfP 982480 project.

Appendix A. Proofs

This appendix contains the proofs of Theorem 3 and Theorem 11. It also contains additional technical details that were used in the proof of Theorem 7.

A.1 Proof of Theorem 3

We start by introducing the following lemma, which is required for the proof of Theorem 3.

Lemma 13 *Consider a function g of two independent discrete random variables (U,V). We assume that $g(u,v) \leq c$, $\forall u,v$, where c is some constant. If $\Pr\{g(U,V) > \tilde{\varepsilon}\} \leq \delta$, then*

$$\Pr_V\{\mathbf{E}_u(g(u,V)) \geq \varepsilon\} \leq \frac{c-\tilde{\varepsilon}}{\varepsilon-\tilde{\varepsilon}}\delta, \quad \forall \varepsilon > \tilde{\varepsilon}.$$

Proof of lemma 13: Let $\mathcal{V}_L'$ be the set of values of V, such that for every $v' \in \mathcal{V}_L'$, $E_u(g(y,v')) \geq \varepsilon$. For every such v' we get,

$$\varepsilon \leq \mathbf{E}_u\left(g(u,v')\right) \leq c\Pr\left\{g(U,V) > \tilde{\varepsilon}|V=v'\right\} + \tilde{\varepsilon}\Pr\left\{g(U,V) \leq \tilde{\varepsilon}|V=v'\right\}.$$

Hence, $\Pr\{g(U,V) > \tilde{\varepsilon} | V = v'\} \geq \frac{\varepsilon - \tilde{\varepsilon}}{c - \tilde{\varepsilon}}$. From the complete probability formula,

$$\begin{aligned} \delta \geq \Pr\{g(U,V) > \tilde{\varepsilon}\} &= \sum_z \Pr\{g(U,V) > \tilde{\varepsilon} | V = v\} P(v) \\ &\geq \frac{\varepsilon - \tilde{\varepsilon}}{c - \tilde{\varepsilon}} \sum_{V : V \in V_L} P(v) \\ &= \frac{\varepsilon - \tilde{\varepsilon}}{c - \tilde{\varepsilon}} \Pr_V \{\mathbf{E}_u (g(u,V)) \geq \varepsilon\}. \end{aligned}$$

Lemma 13 follows directly from the last inequality. □

We first provide an outline of the proof of Theorem **3** and then provide a detailed proof.

Theorem 3—Proof outline: For the given m instances and any clustering $\mathbf{t}$, draw uniformly and independently m' instances (repeats allowed). For any feature index q, we can estimate $I(\mathbf{t}(Z); x_q[Z])$ from the empirical distribution of $(\mathbf{t}, x_q)$ over the m' instances. This empirical distribution is $p(\mathbf{t}(Z'), x_q[Z'])$ where Z' is a random variable denoting the index of instance chosen uniformly from the m' instances (defined formally below). The proof is built up from the following upper bounds, which are independent of m, but depend on the choice of m'. The first bound is on $\mathbf{E}\left\{\left|I(\mathbf{t}(Z); x_q[Z]) - I(\mathbf{t}(Z'); x_q[Z'])\right|\right\}$, where q is fixed and the expectation is over random selection of the m' *instances*. From this bound we derive an upper bound on $|I_{ob} - \mathbf{E}(\hat{I}_{ob})|$ and $|I_{un} - \mathbf{E}(\hat{I}_{un})|$, where $\hat{I}_{ob}$, $\hat{I}_{un}$ are the estimated values of I_{ob}, I_{un} based on the subset of m' instances, that is, the empirical distribution. The last required bound is on the probability that $\sup_{\mathbf{t}:[m]\to[k]} \left|\mathbf{E}\left(\hat{I}_{ob}\right) - \mathbf{E}\left(\hat{I}_{un}\right)\right| > \varepsilon_1$, for any $\varepsilon_1 > 0$. This bound is obtained from Lemmas 2 and 13. The choice of m' is independent of m. Its value should be large enough for the estimations $\hat{I}_{ob}$, $\hat{I}_{un}$ to be accurate, but not too large, so as to limit the number of possible clusterings over the m' instances.

Note that we do not assume the m instances are drawn from a distribution. The m' instances are drawn from the empirical distribution over the m instances. □

Theorem 3—Detailed proof: Let $\tilde{\mathbf{l}} = (l_1, \ldots, l_{m'})$ be indices of m' instances, where each index is selected randomly, uniformly and independently from $\{1, \ldots, m\}$. Let random variable Z' denote a number chosen uniformly at random from $\{1, \ldots, m'\}$. For any feature index q, we can estimate $I(\mathbf{t}(Z); x_q[Z])$ from $I(\mathbf{t}(l_{Z'}); x_q[l_{Z'}])$ as follows. The maximum likelihood estimation of entropy given a discrete empirical distribution $(\hat{p}_1, ..., \hat{p}_N)$, is defined as $\hat{H}_{MLE} = -\sum_{i=1}^{N} \hat{p}_i \log \hat{p}_i$. Note that N is the alphabet size of our discrete distribution. From Paninski (2003) (Proposition 1) the bias between the empirical and actual entropy $H(p)$ is bounded as follows:

$$-\log\left(1 + \frac{N-1}{m'}\right) \leq \mathbf{E}\left(\hat{H}_{MLE}(\hat{p})\right) - H(p) \leq 0.$$

where the empirical estimation $\hat{H}_{MLE}$ is based on m' instances drawn from the distribution p. The expectation is over random sampling of these m' instances. Since $I(\mathbf{t}(Z); x_q[Z]) = -H(\mathbf{t}(Z), x_q[Z]) + H(\mathbf{t}(Z)) + H(x_q(Z))$, we can upper bound the bias between the actual and the empirical estimation of the mutual information as follows:

$$\mathbf{E}_{\tilde{\mathbf{l}}=(l_1,\ldots,l_{m'})}\left\{\left|I(\mathbf{t}(Z); x_q[Z]) - I(\mathbf{t}(l_{Z'}); x_q[l_{Z'}])\right|\right\} \leq \log\left(1 + \frac{ks-1}{m'}\right) \leq \frac{ks}{m'}. \tag{6}$$

Recall that s is the upper bound on the alphabet size of x_q.

Let $\hat{I}_{ob}(\mathbf{t}, \tilde{\mathbf{q}}, \tilde{\mathbf{l}})$ and $\hat{I}_{un}\left(\mathbf{t}, \tilde{\mathbf{l}}\right)$ be the estimated values of $I_{ob}(\mathbf{t}, \tilde{\mathbf{q}})$, $I_{un}(\mathbf{t})$ based on $(l_1, \ldots, l_{m'})$, that is,

$$\hat{I}_{ob}\left(\mathbf{t}, \tilde{\mathbf{q}}, \tilde{\mathbf{l}}\right) = \frac{1}{n} \sum_{i=1}^{n} I\left(t(l_{z'}); x_{q_i}[l_{z'}]\right),$$

$$\hat{I}_{un}(\mathbf{t},\tilde{\mathbf{l}}) = \mathbf{E}_{q\sim\mathcal{D}}\left\{I\left(\mathbf{t}(l_{z'});x_q[l_{z'}]\right)\right\}.$$

From Equation 6 we obtain,

$$|I_{ob}(\mathbf{t},\tilde{\mathbf{q}}) - \mathbf{E}_{\tilde{\mathbf{l}}}\left(\hat{I}_{ob}\left(\mathbf{t},\tilde{\mathbf{q}},\tilde{\mathbf{l}}\right)\right)|,\ |I_{un}(\mathbf{t}) - \mathbf{E}_{\tilde{\mathbf{l}}}\left(\hat{I}_{un}\left(\mathbf{t},\tilde{\mathbf{l}}\right)\right)| \le ks/m',$$

and hence,

$$\begin{aligned} |I_{ob}(\mathbf{t},\tilde{\mathbf{q}}) - I_{un}(\mathbf{t})| &\le \left|\mathbf{E}_{\tilde{\mathbf{l}}}\left(\hat{I}_{ob}\left(\mathbf{t},\tilde{\mathbf{q}},\tilde{\mathbf{l}}\right)\right) - \mathbf{E}_{\tilde{\mathbf{l}}}\left(\hat{I}_{un}\left(\mathbf{t},\tilde{\mathbf{l}}\right)\right)\right| + 2ks/m' \\ &\le \mathbf{E}_{\tilde{\mathbf{l}}}\left(\left|\hat{I}_{ob}\left(\mathbf{t},\tilde{\mathbf{q}},\tilde{\mathbf{l}}\right) - \hat{I}_{un}\left(\mathbf{t},\tilde{\mathbf{l}}\right)\right|\right) + 2ks/m'. \end{aligned} \tag{7}$$

Using Lemma 2 we have an upper bound on the probability that

$$\sup_{\mathbf{t}:[m]\to[k]} \left|\hat{I}_{ob}\left(\mathbf{t},\tilde{\mathbf{q}},\tilde{\mathbf{l}}\right) - \hat{I}_{un}\left(\mathbf{t},\tilde{\mathbf{l}}\right)\right| > \varepsilon$$

over the random selection of *features*, as a function of m'. However, the upper bound we need is on the probability that

$$\sup_{\mathbf{t}:[m]\to[l]} \left\{\mathbf{E}_{\tilde{\mathbf{l}}}\left(\hat{I}_{ob}\left(\mathbf{t},\tilde{\mathbf{q}},\tilde{\mathbf{l}}\right)\right) - \mathbf{E}_{\tilde{\mathbf{l}}}\left(\hat{I}_{un}\left(\mathbf{t},\tilde{\mathbf{l}}\right)\right)\right\} > \varepsilon_1.$$

Note that the expectations $\mathbf{E}_l(\hat{I}_{ob})$, $\mathbf{E}_l(\hat{I}_{un})$ are done over a random selection of the subset of m' *instances*, for a set of features that is randomly selected *once*. In order to link these two probabilities, we use Lemma 13.

From Lemmas 2 and 13 it is easy to show that

$$\Pr_{\tilde{\mathbf{q}}}\left\{\mathbf{E}_{\tilde{\mathbf{l}}}\left(\sup_{\mathbf{t}:[m]\to[k]}\left|\hat{I}_{ob}\left(\mathbf{t},\tilde{\mathbf{q}},\tilde{\mathbf{l}}\right) - \hat{I}_{un}\left(\mathbf{t},\tilde{\mathbf{l}}\right)\right|\right) > \varepsilon_1\right\} \le \frac{4\log k}{\varepsilon_1} e^{-n\varepsilon_1^2/\left(2(\log k)^2\right)+m'\log k}. \tag{8}$$

Lemma 13 is used, where V represents the random selection of features, U represents the random selection of m' instances, $g(u,v) = \sup_{\mathbf{t}:[m]\to[k]}|\hat{I}_{ob} - \hat{I}_{un}|$, $c = \log k$, and $\tilde{\varepsilon} = \varepsilon_1/2$. Since

$$\mathbf{E}_{\tilde{\mathbf{l}}}\left(\sup_{\mathbf{t}:[m]\to[k]}\left|\hat{I}_{ob}\left(\mathbf{t},\tilde{\mathbf{q}},\tilde{\mathbf{l}}\right) - \hat{I}_{un}\left(\mathbf{t},\tilde{\mathbf{l}}\right)\right|\right) \ge \sup_{\mathbf{t}:[m]\to[k]} \mathbf{E}_{\tilde{\mathbf{l}}}\left(\left|\hat{I}_{ob}\left(\mathbf{t},\tilde{\mathbf{q}},\tilde{\mathbf{l}}\right) - \hat{I}_{un}\left(\mathbf{t},\tilde{\mathbf{l}}\right)\right|\right),$$

and from Equations 7 and 8 we obtain

$$\Pr_{\tilde{\mathbf{q}}}\left\{\sup_{\mathbf{t}:[m]\to[k]}|I_{ob}(\mathbf{t},\tilde{\mathbf{q}}) - I_{un}(\mathbf{t})| > \varepsilon_1 + \frac{2ks}{m'}\right\} \le \frac{4\log k}{\varepsilon_1} e^{-n\varepsilon_1^2/\left(2(\log k)^2\right)+m'\log k}.$$

By selecting $\varepsilon_1 = \varepsilon/2$, $m' = 4ks/\varepsilon$, we obtain Theorem 3. □

Note that the selection of m' depends on s (maximum alphabet size of the features). This reflects the fact that in order to accurately estimate $I(\mathbf{t}(Z);x_q[Z])$, we need a number of instances, m', which is much larger than the product of k and the alphabet size of x_q.

A.2 Information Generalization for Soft Clustering

In Section 2 we assumed that we are dealing with hard clustering. Here we show that the generalization theorem (Theorem 3) is also applicable to soft clustering. Nevertheless, we also show that soft clustering is not required, since the maximum value of I_{ob} can be achieved by hard clustering. Hence, although *IobMax*, as appears in Definition 5, is a hard clustering algorithm, it also achieves maximum I_{ob} (and nearly maximum I_{un}) of all possible soft clusterings.

Theorem 3 is applicable to soft clustering from the following arguments. In terms of the distributions $P(\mathbf{t}(Z), x_q(Z))$, assigning a soft clustering to an instance can be approximated by a second empirical distribution, $\hat{P}$, achieved by duplicating each of the instances, and then using hard clustering. Consider, for example, a case where we create a new set of instances by duplicating each of the original instances by 100 identical instances. Using hard clustering on the $\times 100$ larger set of instances, can approximate any soft clustering of the original set with quantization of $P(T|\mathbf{X})$ in steps of $1/100$. Obviously, for any $\varepsilon > 0$ we can create $\hat{P}$ that satisfies $max\left|P - \hat{P}\right| < \varepsilon$.

Now we show that for any soft clustering of an instance, we can find a hard clustering of the same instance that has the same or a higher value of I_{ob} (without changing the cluster identity of other instances). This is enough to show that soft clustering is not required to achieve the maximum value of I_{ob}, since any soft clustering can be replaced by hard clustering instance by instance. Let $P_\lambda(T|X_{q_1},\ldots,X_{q_n})$ define the distribution of any soft clustering. It can be written as the weighted sum of k distributions as follows

$$P_\lambda(T|X_{q_1},\ldots,X_{q_n}) = \sum_{i=1}^{k} \lambda_i \tilde{P}_i^j(T|X_{q_1},\ldots,X_{q_n}), \quad 0 \le \lambda_i \le 1, \sum_{i=1}^{k} \lambda_i = 1.$$

where $\tilde{P}_i^j$ is created by keeping the same soft clustering of instances $\{1,\ldots,j-1,j+1,...,m\}$, and replacing the soft clustering of the jth instance by a hard clustering $\mathbf{t}(j) = i$. Since $I(T;X_q)$ is a convex function of $P(T|X_q)$ for a fixed $P(X_q)$ for any q (Cover and Thomas, 1991), we get

$$I_{P_\lambda}(T;X_q) \le \sum_{i=1}^{k} \lambda_i I_{\tilde{P}_i^j}(T;X_q).$$

Taking the sum over all observed features $(q_1,\ldots q_n)$, we get

$$\sum_q I_{P_\lambda}(T;X_q) \le \sum_{i=1}^{k} \lambda_i \sum_q I_{\tilde{P}_i^j}(T;X_q),$$

and hence at least one of the distributions $\tilde{P}_1^j,\ldots,\tilde{P}_k^j$ has the same or higher I_{ob} then P_λ. In other words, we can replace the soft clustering of any instance j by a hard clustering without decreasing I_{ob}.

A.3 Maximum Likelihood Mixture Model and IobMax

In the proof of Theorem 7 we claimed that maximizing the likelihood of observed variables is equivalent to maximizing $\sum_{j=1}^{n} I(T;X_j) - I(T;Y)$. In this section we show this based on the work of Elidan and Friedman (2003). For the purpose of better readability in the context of their paper, we use the same notations as in their paper, and review them briefly here. Let Y be a variable that denotes the instance identity, that is, $Y[i] = i$ where $i \in \{1,\ldots,m\}$. Let $Q(Y,\mathbf{X})$ be the empirical distribution

of the features $\mathbf{X}$ in the instances, augmented by the distribution of Y. Let $P(\mathbf{X},T)$ be the maximum likelihood mixture model of the joint distribution $Q(\mathbf{X})$, that is, $P(\mathbf{X},T) = P(T)\prod_j \Pr(x_j|T)$.

From Propositions 4.1, 4.3 in Elidan and Friedman (2003), finding local maxima of the likelihood function is equivalent to minimizing the following Lagrangian

$$\mathcal{L}_{EM} = I_Q(T;Y) - (\mathbf{E}_Q[\log P(\mathbf{X},T)] - \mathbf{E}_Q[\log Q(T)]),$$

as a function of $Q(T|Y)$ and $P(\mathbf{X},T)$. In the stationary point of the EM-algorithm (see Propositions 4.4 and 4.5 in Elidan and Friedman, 2003), $Q(x_j,T) = P(x_j,T)$. Minimizing $\mathcal{L}_{EM}$ is equivalent to minimizing $I(T;Y) - \sum I(T;X_j)$ as shown below:

$$\begin{aligned}
\mathcal{L}_{EM} &= I_Q(T;Y) - (\mathbf{E}_Q[\log P(\mathbf{X},T)] - \mathbf{E}_Q[\log Q(T)]) \\
&= I_Q(T;Y) - \sum_{X,T} Q(\mathbf{X},T)\log\left[P(T)\prod_j P(x_j|T)\right] - H(T) \\
&= I_Q(T;Y) + H(T) - \sum_j \sum_{T,x_j} Q(x_j,T)\log\frac{P(x_j,T)}{P(T)} - H(T) \\
&= I_Q(T;Y) - \sum_j \sum_{T,x_j} Q(x_j,T)\log\frac{P(x_j,T)}{P(x_j)P(T)} + \sum_j \sum_{T,x_j} Q(x_i,T)\log P(x_j) \\
&= I_Q(T;Y) - \sum_j I(T;X_j) + \sum_j H(X_j).
\end{aligned}$$

Since $\sum_j H(X_j)$ is independent of $Q(T|Y)$, and $P(T,Y) = Q(T,Y)$ minimizing $\mathcal{L}_{EM}$ is equivalent to maximizing

$$\sum_j I(T;X_j) - I(T;Y).$$

A.4 Proof of Theorem 11

Before proving Theorem 11, we write generalized definitions of D_{ob}, D_{un} and prove a generalization bound for these generalized definitions (Theorem 14). Then we show that Theorem 11 is a special case of Theorem 14.

The quality of the clustering with respect to a single variable, X_q, is defined by a (weighted) average distance of all pairs of instances within the same cluster (large distance means lower quality). This measure is denoted by D_q which is defined by

$$D_q\{C_1,...,C_k\} = \frac{1}{m}\sum_{r=1}^{k}\frac{1}{|C_r|}\sum_{j,l\in C_r} f(x_q[j],x_q[l]).$$

Using these definitions, we define a generalized observed intra-cluster variable, denoted by $\tilde{D}_{ob}$, as the average of D_q over the observed features within the cluster, that is,

$$\tilde{D}_{ob}\{C_1,...,C_k\} = \frac{1}{n}\sum_{i=1}^{n} D_{q_i}\{C_1,...,C_k\}.$$

When $f(a,b) = \frac{1}{2}(a-b)^2$, we get

$$\begin{aligned} D_{ob}\{C_1,...,C_k\} &= \frac{1}{nm}\sum_{r=1}^{k}\sum_{j\in C_r}\sum_{i=1}^{n}\left(x_{q_i}[j] - \frac{1}{|C_r|}\sum_{l\in C_r} x_{q_i}[l]\right)^2 \\ &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{r=1}^{k}\frac{1}{2|C_r|}\sum_{j\in C_r}\sum_{l\in C_r}\left(x_{q_i}[j] - x_{q_i}[l]\right)^2 \\ &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{r=1}^{k}\frac{1}{|C_r|}\sum_{j\in C_r}\sum_{l\in C_r} f\left(x_{q_i}[j], x_{q_i}[l]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} D_{q_i}\{C_1,...,C_k\} \\ &= \tilde{D}_{ob}\{C_1,...,C_k\}, \end{aligned} \tag{9}$$

which means that D_{ob} is a special case of $\tilde{D}_{ob}$.

Similarly we define the generalized unobserved intra-cluster variance, denoted by $\tilde{D}_{un}$, as follows:

$$\tilde{D}_{un}\{C_1,...,C_k\} = \mathbf{E}_{q\sim\mathcal{D}}\left\{D_q\{C_1,...,C_k\}\right\}.$$

Again, when $f(a,b) = \frac{1}{2}(a-b)^2$, we get

$$D_{un}\{C_1,...,C_k\} = \tilde{D}_{un}\{C_1,...,C_k\}. \tag{10}$$

Theorem 14 *With the above definitions, for every function, f satisfies Equations 2, 3 and 4 (see Section 3) and for every $\varepsilon > 0$,*

$$\Pr_{\{q_1,...,q_n\}}\left\{\sup_{\alpha(\{C_1,...,C_k\})\geq\alpha_c}\left|\tilde{D}_{ob}\{C_1,...,C_k\} - \tilde{D}_{un}\{C_1,...,C_k\}\right| > \varepsilon\right\} \leq \frac{2k}{\alpha_c}e^{-n\varepsilon^2/2c^2+\log\frac{c}{\varepsilon}},$$

where α is defined in Definition 10.

Before proving Theorem 14, we introduce the following lemma that is required for the proof.

Lemma 15 *Let $\{Z_1,...,Z_S\}$ be a set of jointly S distributed random binary variables, where $z_i \in \{0,1\}$. If $\Pr(z_i = 1) \leq \delta$ for every i then for any $N_b \geq 1$*

$$\Pr\left\{\sum_{i=1}^{S} z_i \geq N_b\right\} \leq \frac{S}{N_b}\delta.$$

Lemma 15 follows directly from Markov's inequality.

The proof of Theorem 14 (given below) is based on the observation that $\tilde{D}_{ob}$ and $\tilde{D}_{un}$ are the weighted average of distance functions over a subset of the pairs of instances. This subset includes only pairs that are within the same cluster. In other words, the calculated inter-cluster variances of a clustering is based on the weighted average of $\frac{1}{2}\sum_{r=1}^{k}|C_r|(|C_r|-1)$ pairs out of the $\frac{1}{2}m(m-1)$ pairs of instances. We define "bad pairs" as pairs of instances with a large difference between the observed and unobserved distances. We use Hoeffding's inequality and Lemma 15 to bound the

probability that we have a large number of "bad pairs". Then we show that if the number of "bad pairs" is small, all clusterings are "good", that is, $\left|\tilde{D}_{ob} - \tilde{D}_{un}\right| \leq \varepsilon$.

Proof of Theorem 14

For each pair of instances j,l $(l > j)$ we define a random variable d_{jl} as the difference between the average observed distance and the expected distance,

$$d_{jl} = \left| \frac{1}{n} \sum_{i=1}^{n} f\left(x_{q_i}[j], x_{q_i}[l]\right) - \mathbf{E}_q \left\{ f\left(x_q[j], x_q[l]\right) \right\} \right|.$$

We also define a binary random variable Z_{jl} $(l > j)$ by

$$z_{jl} = \begin{cases} 1 & \text{if } d_{jl} > \tilde{\varepsilon} \\ 0 & \text{otherwise} \end{cases}$$

where $\tilde{\varepsilon}$ is a positive constant. In other words, z_{jl} is one for "bad" pairs, that is, pairs that have a large difference between the average observed distance and the expected distance. From Hoeffding's inequality,

$$\Pr_{\{q_1,\ldots,q_n\}} \left(z_{jl} = 1\right) \leq \tilde{\delta}, \tag{11}$$

where

$$\tilde{\delta} = 2e^{-2n\tilde{\varepsilon}^2/c^2}. \tag{12}$$

We have $\frac{1}{2}m(m-1)$ of these random binary variables. Let N_{bad} be the number of "bad" pairs, that is, $N_{bad} = \sum_{j,l:l>j} z_{jl}$. First we calculate an upper bound on $\left|\tilde{D}_{ob} - \tilde{D}_{un}\right|$ as a function of N_{bad}, and later we prove an upper bound on the probability of large N_{bad}.

By definition of $\tilde{D}_{ob}, \tilde{D}_{un}, d_{jl}$ and the properties of f (Equations 3 and 4) we can bound the difference between $\tilde{D}_{ob}$ and $\tilde{D}_{un}$ (for any clustering) as follows

$$\begin{aligned} &\left|\tilde{D}_{ob}\left\{C_1,\ldots,C_k\right\} - \tilde{D}_{un}\left\{C_1,\ldots,C_k\right\}\right| \\ &= \left| \frac{1}{mn} \sum_{i=1}^{n} \sum_{r=1}^{k} \frac{1}{|C_r|} \sum_{j,l \in r} f\left(x_{q_i}[j], x_{q_i}[l]\right) - \frac{1}{m} \sum_{i=1}^{k} \frac{1}{|C_r|} \sum_{j,l \in C_r} \mathbf{E}_q \left\{ f\left(x_q[j], x_q[l]\right) \right\} \right| \\ &\leq \frac{2}{m} \sum_{r=1}^{k} \frac{1}{|C_r|} \sum_{j,l \in C_k : l > j} d_{jl}. \end{aligned}$$

By defining ε_d as the following function of the clustering

$$\varepsilon_d\left(\left\{C_1,\ldots,C_k\right\}\right) = \frac{2}{m} \sum_{r=1}^{k} \frac{1}{|C_r|} \sum_{j,l \in C_k : l > j} d_{jl},$$

we have

$$\left|\tilde{D}_{ob}\left\{C_1,\ldots,C_k\right\} - \tilde{D}_{un}\left\{C_1,\ldots,C_k\right\}\right| \leq \varepsilon_d. \tag{13}$$

Recall that r_t is the number of instances in cluster t. Now we calculate an upper bound on ε_d as a function of $\tilde{\varepsilon}$ and N_{bad}. The total number of pairs in the rth cluster is $\frac{1}{2}|C_r|\left(|C_r| - 1\right)$. We have N_{bad}

pairs with a difference above $\tilde{\varepsilon}$ (but not more than c, since f is bounded). The error of each of the other pairs is upper bounded by $\tilde{\varepsilon}$. Hence we get

$$\begin{aligned}\varepsilon_d &\leq \frac{2}{m}\sum_{r=1}^{k}\frac{1}{|C_r|}\sum_{j,l\in C_r:l>j}\left(\tilde{\varepsilon}+z_{jl}\left(c-\tilde{\varepsilon}\right)\right)\\ &\leq \frac{2}{m}\left(\frac{1}{2}\sum_{r=1}^{k}\frac{1}{|C_r|}|C_r|\left(|C_r|-1\right)\tilde{\varepsilon}+\frac{N_{bad}\left(c-\tilde{\varepsilon}\right)}{\min_r|C_r|}\right)\\ &\leq \frac{2}{m}\left(\frac{1}{2}\sum_{r=1}^{k}|C_r|\tilde{\varepsilon}+\frac{N_{bad}c}{\min_r|C_r|}\right).\end{aligned}$$

Note that $\sum_r |C_r| = m$ (the sum of the size of all clusters is m). Hence,

$$\varepsilon_d \leq \tilde{\varepsilon}+\frac{2N_{bad}}{m\min_r|C_r|}c. \tag{14}$$

Let N_b be defined as follows

$$N_b = \frac{m\min_r|C_r|\tilde{\varepsilon}}{2c}. \tag{15}$$

If $N_{bad} \leq N_b$ then $\varepsilon_d \leq 2\tilde{\varepsilon}$ (from Equations 14 and 15). Hence,

$$\Pr_{\{q_1,\ldots,q_n\}}\left\{\varepsilon_d > 2\tilde{\varepsilon}\right\} = \Pr_{\{q_1,\ldots,q_n\}}\left\{N_{bad} > N_b\right\}. \tag{16}$$

From Lemma 15 and Equation 11 for any $N_b \geq 1$

$$\Pr_{\{q_1,\ldots,q_n\}}\left\{N_{bad} \geq N_b\right\} \leq \frac{m(m-1)}{2N_b}\tilde{\delta} \leq \frac{m^2}{2N_b}\tilde{\delta}. \tag{17}$$

Combining Equations 15, 16, 17 and the definition of $\tilde{\delta}$ (Equation 12) we get

$$\Pr_{\{q_1,\ldots,q_n\}}\left\{\varepsilon_d > 2\tilde{\varepsilon}\right\} \leq \frac{mc}{\min_r|C_r|\tilde{\varepsilon}}\tilde{\delta} = \frac{2mc}{\min_r|C_r|\tilde{\varepsilon}}e^{-2n\tilde{\varepsilon}^2/c^2}. \tag{18}$$

By selecting $\varepsilon = \tilde{\varepsilon}/2$ and using Equation 13 and 18 we get

$$\Pr_{\{q_1,\ldots,q_n\}}\left\{\sup_{\alpha(\{C_1,\ldots,C_k\})\geq\alpha_c}\left|\tilde{D}_{ob}\left\{C_1,\ldots,C_k\right\}-\tilde{D}_{un}\left\{C_1,\ldots,C_k\right\}\right| > \varepsilon\right\} \leq 4\frac{m}{\min_r|C_r|}e^{-n\varepsilon^2/2c^2+\log\frac{c}{\varepsilon}}. \tag{19}$$

Using the definition of α we have $m/\min_r|C_r| \leq k/\alpha_c$. Together with Equation 19 we get Theorem 14.

■

Now we are ready to prove Theorem 11 by showing it is a special case of Theorem 14.

Proof of Theorem 11

Let $f(a,b) = \frac{1}{2}(a-b)^2$. In this case f satisfies Equation 2, 3 and 4 where $c = 2R^2$ (since $0 \leq f(x_q[j],x_q[l]) \leq 2R^2$ for all q, j, l). In addition, $\tilde{D}_{ob} = D_{ob}$ and $\tilde{D}_{un} = D_{un}$ (Equations 9 and 10). Therefore, Theorem 11 is a special case of Theorem 14.

■

Appendix B. Notation Table

The following table summaries the notation and definitions used in the paper for quick reference.

Notation	Short Description
m	Number of instances
L	Number of features (both observed and unobserved)
$\{X_1,\ldots,X_L\}$	Random variables (features)
n	Number of observed features
$\tilde{\mathbf{q}}=(q_1,\ldots,q_n)$	Indices of observed features
$\mathcal{D}$	Probability distribution of selecting features
$\{\mathbf{x}[1],\ldots,\mathbf{x}[m]\}$	Instances (each instance is a vector of L elements, where n are observed)
$x_q[j]$	The qth feature of the jth instance
$x_{q_i}[j]$	The ith observed feature of the jth instance
k	Number of clusters
$\mathbf{t}:[m]\rightarrow[k]$	Function that maps instances to clusters
$\{C_1,\ldots,C_k\}$	Clusters (C_r is the set of instances in the rth cluster)
$\lvert C_r\rvert$	Size of rth cluster
T	A variable that represents the cluster label
Z	A random variable taking values uniformly from $\{1,2,\ldots,m\}$ (Random selection of instance index)
s	Upper bound on the number of values a discrete feature can have
I_{ob}	Average observed information (Definition 1)
I_{un}	Expected unobserved information (Definition 1)
$I^*_{un,k}$	Maximum possible value of I_{un} for k clusters (Definition 4)
$I^*_{ob,k}$	Value of I_{ob} for clustering with maximum I_{un} (Definition 4)
$\tilde{I}_{un,k}$	Value of I_{un} for clustering that achieves $\tilde{I}_{ob,k}$ (Definition 5)
$\tilde{I}_{ob,k}$	Maximum possible value of I_{ob} for k clusters (Definition 5)
$f(\cdot,\cdot)$	Distance function between two feature values
c	Constant - upper bound on the distance function
$D_{ob}\{C_1,\ldots,C_k\}$	Observed intra-cluster variance (Definition 8)
$D_{un}\{C_1,\ldots,C_k\}$	Expected unobserved intra-cluster variance (Definition 9)
D^{opt}_{un}	Minimum possible intra-cluster variance (Theorem 12)
$\alpha(\{C_1,\ldots,C_k\})$	Ratio between smallest to average cluster size (Definition 10)

References

R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6, 2005.

J. Blitzer, A. Globerson, and F. Pereira. Distributed latent variable models of lexical co-occurrences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2005.

J.S. Breese, D. Heckerman, and K. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 1998.

R. Caruana and V. R. de Sa. Promoting poor features to supervisors: Some inputs work better as outputs. In *Advances in Neural Information Processing Systems (NIPS)*, 1997.

G. Chechik and N. Tishby. Extracting relevant structures with side information. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

T. M. Cover and J. A. Thomas. *Elements Of Information Theory*. Wiley Interscience, 1991.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 1977.

G. Elidan and N. Friedman. The information bottleneck EM algorithm. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2003.

N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2001.

A. Globerson and S. Roweis. Nightmare at test time: robust learning by feature deletion. In *International Conference on Machine Learning (ICML)*, 2006.

D. Gondek and T. Hofmann. Non-redundant data clustering. In *Fourth IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society, 2004.

T. S. Han. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36:133–156, 1978.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31 (3):264–323, September 1999.

J. Kleinberg. An impossibility theorem for clustering. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

E. Krupka and N. Tishby. Generalization in clustering with unobserved features. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

E. Krupka and N. Tishby. Incorporating prior knowledge on features into learnning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

E. Krupka, A. Navot, and N. Tishby. Learning to select features using their properties. *Journal of Machine Learning Research*, submitted.

K. Lang. Learning to filter netnews. *Proc. 12th International Conf. on Machine Learning*, pages 331–339, 1995.

S. P. Lloyd. Least squares quantization in pcm. Technical report, Bell Laboratories, 1957. Published in 1982 in IEEE Transactions on Information Theory 28, 128-137.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

B. Marlin. Collaborative filtering: A machine learning perspective. Master's thesis, University of Toronto, 2004.

L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1101–1253, 2003.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

M. Seeger. Learning with labeled and unlabeled data. Technical report, University of Edinburgh, 2002.

M. Szummer and T. Jaakkola. Information regularization with partially labeled data. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. *Proc. 37th Allerton Conf. on Communication and Computation*, 1999.

L. H. Ungar and D. P. Foster. Clustering methods for collaborative filtering. In *Workshop on Recommender Systems at the 15th National Conference on Artificial Intelligence.*, 1998.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL Workshop on Statistics and Optimization of Clustering*, 2005.