Multi-Classification by Categorical Features via Clustering

Yevgeny Seldin[†] SELDIN@CS.HUJI.AC.IL
Naftali Tishby[†,‡] TISHBY@CS.HUJI.AC.IL
[†]School of Computer Science and Engineering, [‡]Center for Neural Computation,
The Hebrew University of Jerusalem, Israel

Abstract

We derive a generalization bound for multi-classification schemes based on grid clustering in categorical parameter product spaces. Grid clustering partitions the parameter space in the form of a Cartesian product of partitions for each of the parameters. The derived bound provides a means to evaluate clustering solutions in terms of the generalization power of a built-on classifier. For classification based on a single feature the bound serves to find a globally optimal classification rule. Comparison of the generalization power of individual features can then be used for feature ranking. Our experiments show that in this role the bound is much more precise than mutual information or normalized correlation indices.

1. Introduction

Clustering is one of the basic tools for dimensionality reduction in categorical spaces. In this paper we study classifiers based on a soft grid clustering of categorical parameter product spaces. The grid clustering is defined by a set of stochastic mappings $\{q_i : \mathcal{X}_i \mapsto \{1, .., m_i\}\}$, one for each parameter i, which map the possible values $\mathcal{X}_i$ of the parameter X_i to a reduced set C_i of size m_i. A classifier based on the grid clustering then assigns a separate prediction strategy to each partition cell. For example, in collaborative filtering we can cluster a thousand by thousand space of viewers by movies into a five by five space of viewer clusters by movie clusters (here X_1 are the viewers and X_2 are the movies). Then we can predict a missing entry within some partition cell with an average of ratings in that cell.

Grid clustering or some other form of dimensionality reduction can be helpful and even essential when the sample size is limited. However, an appropriate choice of clustering resolution is crucial for good results. A coarse clustering may be highly imprecise - think of the extreme of putting all the data into one big cluster. On the other hand, a fine clustering may be statistically unreliable - at the opposite extreme, if we put every parameter value into a separate cluster, some parameter combinations may not occur in the training set at all. Thus, unification of parameter values amplifies the statistical reliability, but reduces the precision. In this paper we relate this tradeoff to generalization properties of a classifier based on the clustering.

Applications of grid clustering to data with intrinsically categorical features are abundant. Seldin, Slonim and Tishby (2007) consider grid clustering from an MDL perspective and demonstrate its success in predicting missing values in the context of collaborative filtering. The same work achieves state of the art performance in terms of coherence of obtained clusters with manual annotation in the context of gene expression and stock data analysis. Here we also suggest a new application of grid clustering for feature ranking.

We are not aware of any previous work on generalization properties of models based on grid clustering. A somewhat related work is (Srebro, 2004), which derives a generalization bound for matrix approximation with bounded norm factorization. However, matrix factorization is a different model and the proof is based on a different technique (Rademacher complexities).

The key point of this paper is a derivation of a generalization bound for classification based on grid clustering. The bound is derived by using the PAC-Bayesian technique (McAllester, 1999). The power of the PAC-Bayesian technique lies in its ability to handle heterogeneous hypothesis classes so that the generalization bound for a specific hypothesis depends on the complexity of that hypothesis rather than on the complexity of the whole class. A classical example of an appli-

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

cation of the PAC-Bayesian bound are SVMs (Langford, 2005). It is well known that the VC-dimension of separating hyperplanes in $\mathbb{R}^n$ is $n+1$. As well, the VC-dimension of separating hyperplanes with a margin γ, assuming all points are bounded in a unit sphere, is $min\{\frac{1}{\gamma^2}, n\} + 1$. The ability to slice the hypothesis space into infinitely many subspaces characterized by a finer notion of complexity (the size of the margin) rather than the coarse VC-dimension of the whole class makes it possible to derive a better bound that remains meaningful even for infinite dimensional spaces.

In this paper we propose a fine measure of complexity of a grid partition of a cardinal space. The proposed measure of complexity is related to the entropy of a partition along each dimension i. The bound enables us to consider all possible partitions of the product space and to choose one with better generalization properties. In the case of a single parameter it is easy to find a global optimum of the bound. The mapping rule achieving the optimum is shown to be the optimal classification rule from a generalization point of view. Although the bound is not perfectly tight, its shape follows an error on a validation set extremely well. In the experimental section we apply the bound to feature ranking and it is shown to be much more precise than standard mutual information or normalized correlation rankings.

2. A Brief Review of the PAC-Bayesian Generalization Bound

To set the stage, we start with a simplified version of the PAC-Bayesian bound, called Occam's razor. Let $\mathcal{H}$ be a countable hypothesis space. For a hypothesis $h \in \mathcal{H}$ denote by $L(h)$ an expected and by $\hat{L}(h)$ an empirical loss of h. We assume the loss is bounded by b.

Theorem 1 (Occam's razor)**.** *For any data generating distribution and for any "prior distribution" $P(h)$ over $\mathcal{H}$ with a probability greater than $1-\delta$ over drawing an i.i.d. sample of size N, for all $h \in \mathcal{H}$:*

$$L(h) \leq \hat{L}(h) + b\sqrt{\frac{-\ln P(h) - \ln\delta}{2N}}. \qquad (1)$$

Proof. The proof is fairly simple and provides a good illustration of what the "prior distribution" $P(h)$ is. By Hoeffding's inequality $P\{L(h) - \hat{L}(h) \geq \varepsilon(h)\} \leq e^{-2N\varepsilon(h)^2/b^2}$ for any given $h \in \mathcal{H}$. We require that $e^{-2N\varepsilon(h)^2/b^2} \leq P(h)\delta$ for some prior $P(h)$ that satisfies $\sum_{h\in\mathcal{H}} P(h) = 1$. Then, by the union bound $L(h) \leq \hat{L}(h) + \varepsilon(h)$ for all $h \in \mathcal{H}$ with a probability of $1-\delta$. The minimal value of ε that satisfies the requirement is $\varepsilon(h) = b\sqrt{\frac{-\ln P(h) - \ln\delta}{2N}}$, which completes the proof.

We now introduce the notion of a randomized classifier. Let Q be any (posterior) distribution over $\mathcal{H}$. A randomized classifier associated with Q works by choosing a new classifier h from $\mathcal{H}$ according to Q every time a classification is made. We denote the loss of a strategy Q by $L(Q) = \mathbb{E}_{h\sim Q}L(h)$ and similarly $\hat{L}(Q) = \mathbb{E}_{h\sim Q}\hat{L}(h)$. By taking an expectation of (1) over the choice of h and exploiting the concavity of the square root we obtain that with a probability greater than $1-\delta$:

$$L(Q) \leq \hat{L}(Q) + b\sqrt{\frac{-\mathbb{E}_{h\sim Q}\ln P(h) - \ln\delta}{2N}}. \qquad (2)$$

The PAC-Bayesian bound (McAllester, 1999) was derived to allow uncountably infinite hypothesis spaces, though in our case the hypothesis space is finite. We cite a slightly tighter version of the bound proved in (Maurer, 2004).

Theorem 2 (PAC-Bayesian Bound)**.** *For any data distribution and for any "prior" P over $\mathcal{H}$ fixed ahead of training with a probability greater than $1-\delta$ for all distributions Q over $\mathcal{H}$:*

$$L(Q) \leq \hat{L}(Q) + b\sqrt{\frac{D(Q\|P) + \frac{1}{2}\ln(4N) - \ln\delta}{2N}}, \qquad (3)$$

where $D(Q\|P) = \mathbb{E}_{h\sim Q}\ln\frac{Q(h)}{P(h)}$ is the Kullback-Leibler (KL) divergence between the distributions Q and P.

If the loss function is bounded by one (1), (2), and (3) may be written in the form of a bound on the KL-divergence between $\hat{L}(Q)$ and $L(Q)$. For example, (1) may be obtained as: $D(\hat{L}(h)\|L(h)) \leq \frac{-\ln P(h) - \ln\delta}{N}$ if we start from the $P\{L(h) - \hat{L}(h) \geq \varepsilon\} \leq e^{-ND(L(h)+\varepsilon\|L(h))}$ form of Hoeffding's inequality. This provides a better bound in cases when $\hat{L}(Q)$ is sufficiently small (less than $\frac{1}{8}$). The choice of the square root form of the bounds is based on their easier analytical tractability for subsequent minimization.

By writing $D(Q\|P) = -H(Q) - \mathbb{E}_{h\sim Q}\ln P(h)$ it is easy to see that (3) is an improvement over (2) when $H(Q) > \frac{1}{2}\ln(4N)$. In our experiments (2) is usually tighter. It is an open question whether $\frac{1}{2}\ln(4N)$ can be removed from (3), at least in the case of a countable $\mathcal{H}$. For example, (Blanchard & Fleuret, 2007) suggest a parameterized tradeoff $\frac{k+1}{k}D(Q\|P) + \ln(k+1) + 3.5 + \frac{1}{2k}$ instead of $D(Q\|P) + \frac{1}{2}\ln(4N)$. For our data the tradeoff does not improve the results and therefore we omit its discussion.

3. A Formal Definition of Grid Clustering

Before proceeding to the results we provide a formal definition of grid clustering as used in this paper.

Definition 1. Grid Clustering *of the parameter space $\mathcal{X}_1 \times .. \times \mathcal{X}_d$ is a set of distributions $q_i(C_i|X_i)$ defining the probability of mapping $X_i \in \mathcal{X}_i$ to $C_i \in \{1, .., m_i\}$. If each of $q_i(C_i|X_i)$ is deterministic, we call the clustering a* deterministic grid clustering*. Otherwise it is a* stochastic grid clustering*.*

In the following sections we assume some unknown joint probability distribution $p(Y, X_1, .., X_d)$ of the parameters and the label exists. The set of all possible labels is denoted by $\mathcal{Y}$ and its size is denoted by n_y. The size of $\mathcal{X}_i$ is denoted by n_i. The cardinality of C_i is m_i. The value of each m_i can vary in the range of $1 \leq m_i \leq n_i$ for different partitions. A hypothesis h in a *deterministic* grid clustering is comprised of a set of deterministic mappings $q_i(C_i|X_i)$, for simplicity denoted by $q_i(X_i) : \mathcal{X}_i \to \{1, .., m\}$, and a set of $\prod_i m_i$ labels, one for each partition cell. We denote the hypothesis space by $\mathcal{H}$ and decompose it as $\mathcal{H} = \mathcal{H}|_1 \times .. \times \mathcal{H}|_d \times \mathcal{H}|_{\mathcal{Y}|\bar{m}}$. Here $\mathcal{H}|_i$ is a space of all possible partitions of $\mathcal{X}_i$, or, in other words, a projection of $\mathcal{H}$ onto dimension i. $\bar{m} = (m_1, .., m_d)$ is a vector of cardinalities of the partitions along each dimension and $\mathcal{H}|_{\mathcal{Y}|\bar{m}}$ is a space of all possible labelings of $\prod_i m_i$ partition cells. Similarly, $h \in \mathcal{H}$ is decomposed as $h = h|_1 \times .. \times h|_d \times h|_{y|\bar{m}}$. It is assumed that a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is given. The loss of h, denoted by $L(h)$, is defined as an expectation over p of L: $L(h) = \mathbb{E}_p l(h(X_1, .., X_d), Y(X_1, .., X_d))$. The empirical loss of h on a sample S of size N is denoted by $\hat{L}(h)$ and equals the average loss on the sample.

For stochastic mappings $q_i(C_i|X_i)$ it is assumed that a random realization of the mapping is done prior to the prediction. In other words, we choose a hypothesis h at random by determining the values of $q_i(X_i)$ according to $q_i(C_i|X_i)$ before we make a prediction. $Q = \{\{q_i(C_i|X_i)\}_{i=1}^d, q(Y|C_1, .., C_d)\}$ collectively denotes a distribution over $\mathcal{H}$ associated with a randomized classifier called Q. The loss of Q is denoted by $L(Q)$ and equals $L(Q) = \mathbb{E}_{h \sim Q} L(h)$. The empirical loss of Q is denoted by $\hat{L}(Q)$.

We define $q_i(C_i) = \frac{1}{n_i} \sum_{x_i} q_i(C_i|x_i)$ to be a marginal distribution over C_i corresponding to a *uniform* distribution over $\mathcal{X}_i$ and the conditional distribution $q_i(C_i|X_i)$ of our choice. The entropy of a partition along a dimension i with respect to a uniform distribution over $\mathcal{X}_i$ is then $H_U(q_i) \equiv H_U(C_i) = -\sum_{c_i} q_i(c_i) \ln q_i(c_i)$. The mutual information between X_i and C_i with respect to a uniform distribution over $\mathcal{X}_i$ is $I_U(X_i; C_i) = \frac{1}{n_i} \sum_{x_i, c_i} q_i(c_i|x_i) \ln[q_i(c_i|x_i)/q_i(c_i)]$.

4. Generalization Bound for Multi-Classification with Grid Clustering

In this section we state and prove a generalization bound for multi-classification with stochastic grid clustering:

Theorem 3. *For any probability measure p over instances and for any loss function l bounded by b, with a probability of at least $1-\delta$ over a selection of an i.i.d. sample S of size N according to p, for all randomized classifiers $Q = \{\{q_i(C_i|X_i)\}_{i=1}^d, q(Y|C_1, .., C_d)\}$:*

$$L(Q) \leq \hat{L}(Q) + b\sqrt{\frac{\sum_i n_i H_U(C_i) + K}{2N}}, \qquad (4)$$

$$K = \sum_i (m_i \ln n_i + \frac{(\ln(n_i) + 1)^2}{4}) + (\prod_i m_i) \ln n_y - \ln \delta. \qquad (5)$$

It is also possible to replace (4) in the theorem with:

$$L(Q) \leq \hat{L}(Q) + b\sqrt{\frac{\sum_i n_i I_U(X_i; C_i) + \frac{1}{2}\ln(4N) + K}{2N}}. \qquad (6)$$

Proof. The bounds (4) and (6) are direct consequences of (2) and (3) respectively for an appropriate choice of a prior P over $\mathcal{H}$. The main part of the proof is to define a prior P that will provide a meaningful complexity-related slicing of $\mathcal{H}$ and then to calculate $-\mathbb{E}_Q \ln P(h)$ for (4) and $D(Q\|P)$ for (6).

To define the prior P over $\mathcal{H}$ we count the hypotheses in $\mathcal{H}$. For a fixed partition there are $n_y^{\prod_i m_i}$ possibilities to assign the labels to the partition cells. There are n_i possibilities to choose the number of clusters along a dimension i. There are at most $\binom{n_i+m_i-1}{m_i-1} \leq n_i^{m_i-1}$ possibilities to choose a cluster cardinality profile along a dimension i. (This is the number of possibilities to place $m_i - 1$ ones in a sequence of $n_i + m_i - 1$ ones and zeros, where ones symbolize a partition of zeros ("balls") into m_i bins.) We take the $n_i^{m_i-1}$ bound for simplicity. For a fixed cardinality profile $|c_{i1}|, .., |c_{im_i}|$ (over a single dimension) there are $\binom{n_i}{|c_{i1}|,..,|c_{im_i}|}$ possibilities to assign X_i-s to the clusters. This multinomial coefficient can be bounded from above by $e^{n_i H_U(C_i)}$ (see (Cover & Thomas, 1991, page 284) for an elegant proof). Putting all the combinatorial calculations together it is possible to define a distribution $P(h)$ over

$\mathcal{H}$ that satisfies:

$$P(h) \geq \frac{1}{\exp\left[\sum_i (n_i H_U(C_i) + m_i \ln n_i) + (\prod_i m_i) \ln n_y\right]}. \tag{7}$$

We pause to stress that unlike in most applications of the PAC-Bayesian bound, in our case the prior P and the posterior Q are defined over slightly different hypothesis spaces. The posterior Q is defined for *named* clusterings - we explicitly specify for each X_i the "name" of C_i it is mapped to. Whereas the prior P is defined over *unnamed* partitions - we only check the cardinality profile of C_i, but we cannot recover which X_i-s are mapped to a given C_i. Nevertheless, the "named" distribution Q induces a distribution over the "unnamed" space by summing up over all possible name permutations. This enables us to compute $-\mathbb{E}_Q \ln P(h)$ we need for the bound.

We now turn to bound $-\mathbb{E}_Q \ln P(h)$. This is done by showing that Q is concentrated around the hypotheses (hard partitions) h for which the entropies of the partitions are close to the entropies $H_U(q_i)$. By the decomposition property we can write: $P(h) = P(h|_1)..P(h|_d)P(h|_{y|\bar{m}})$, and similarly for Q. Then $-\mathbb{E}_Q \ln P(h) = -\sum_i \mathbb{E}_Q \ln P(h|_i) - \mathbb{E}_Q \ln P(h|_{y|\bar{m}})$, and similarly for $D(Q\|P)$. The last term is easy to compute since P is uniform over $\mathcal{H}_{\mathcal{Y}|\bar{m}}$ and Q is defined for a fixed $\bar{m}$. Therefore, $-\mathbb{E}_Q \ln P(h|_{y|\bar{m}}) = (\prod_i m_i) \ln n_y$. For the first d terms we need to compute or at least to bound $Q(h|_i)$.

Recall that $h|_i$ is obtained from Q by drawing a cluster C_i for each $X_i \in \mathcal{X}_i$ independently according to the distribution $q_i(C_i|X_i)$. Let $\hat{q}_i = \{\frac{|c_{i1}|}{n_i}, .., \frac{|c_{im_i}|}{n_i}\}$ denote an empirical cluster cardinality profile along a dimension i obtained by such assignment. Then:

$$\mathbb{E}_{q_i} H(\hat{q}_i) = H_U(q_i) - \mathbb{E}_{q_i} D(\hat{q}_i\|q_i) \leq H_U(q_i), \tag{8}$$

where $H(\hat{q}_i) = -\sum_{c_i} \hat{q}(c_i) \ln \hat{q}(c_i)$ and $D(\hat{q}_i\|q_i) = \sum_{c_i} \hat{q}_i(c_i) \ln \frac{\hat{q}_i(c_i)}{q_i(c_i)}$. And also:

$$P_{q_i}\left\{H(\hat{q}_i) - \mathbb{E}H(\hat{q}_i) \geq \varepsilon\right\} \leq e^{-2n_i\varepsilon^2/(\ln(n_i)+1)^2}. \tag{9}$$

The latter inequality follows from the fact that the empirical entropy $H(\hat{q}_i)$ satisfies a bounded differences property with a constant equal to $\frac{1}{\ln(n_i)+1}$. See (Paninski, 2003) for a more detailed proof of (8) and (9).

Now, if $\hat{q}_i$ is the cardinality profile of $h|_i$, then $Q(h|_i) = Q(\hat{q}_i) \equiv P_{q_i}\{\hat{q}_i\}$. Let $\varepsilon(\hat{q}_i) = \max\{0, H(\hat{q}_i) - H_U(q_i)\}$. Since $H(\hat{q}_i) - H_U(q_i) \leq H(\hat{q}_i) - \mathbb{E}H(\hat{q}_i)$ by (8), from (9) we have: $Q(\hat{q}_i) \leq e^{-2n_i\varepsilon(\hat{q}_i)^2/(\ln(n_i)+1)^2}$. Thus:

$$-\mathbb{E}_Q \ln P(h|_i) = -\sum_{h|_i \in \mathcal{H}|_i} Q(h|_i) \ln P(h|_i)$$

$$= \sum_{h|_i \in \mathcal{H}|_i} Q(\hat{q}_i)(n_i H(\hat{q}_i) + m_i \ln n_i)$$

$$= \sum_{h|_i \in \mathcal{H}|_i} Q(\hat{q}_i)[n_i H_U(q_i) + m_i \ln n_i + n_i(H(\hat{q}_i) - H_U(q_i))]$$

$$\leq n_i H_U(q_i) + m_i \ln n_i + \sum_{h|_i \in \mathcal{H}|_i} Q(\hat{q}_i) n_i \varepsilon(\hat{q}_i)$$

$$\leq n_i H_U(q_i) + m_i \ln n_i + \sum_{h|_i \in \mathcal{H}|_i} n_i \varepsilon(\hat{q}_i) e^{\frac{-2n_i\varepsilon(\hat{q}_i)^2}{(\ln(n_i)+1)^2}}$$

$$\leq n_i H_U(q_i) + m_i \ln n_i + \int_0^\infty n_i \varepsilon e^{-2n_i\varepsilon^2/(\ln(n_i)+1)^2} d\varepsilon$$

$$= n_i H_U(C_i) + m_i \ln n_i + \frac{1}{4}(\ln(n_i)+1)^2.$$

This completes the proof of (4).

For (6) what remains is to compute $\mathbb{E}_Q \ln Q(h|_i)$. To do so we bound $\ln Q(\hat{q}_i)$ from above. The bound follows from the fact that if we draw n_i values of C_i according to $q_i(C_i|X_i)$ the probability of the resulting type $\hat{q}_i$ is bounded from above by $e^{-n_i H_U(C_i|X_i)}$, where $H_U(C_i|X_i) = -\frac{1}{n_i}\sum_{x_i,c_i} q_i(c_i|x_i) \ln q_i(c_i|x_i)$ (see Theorem 12.1.2 in (Cover & Thomas, 1991)). Thus $\mathbb{E}_Q \ln Q(h|_i) \leq -n_i H_U(C_i|X_i)$, which together with the identity $I_U(X_i; C_i) = H_U(C_i) - H_U(C_i|X_i)$ completes the proof of (6).

5. An Optimal Solution for a Single Feature and Feature Ranking

In this section we show that if there is only one parameter X (i.e., $d = 1$) a globally optimal (from a generalization point of view) classification rule may be efficiently found by examining the "direct" mappings $q(Y|X)$. In other words, for a single parameter there is no need for intermediate clustering. The obtained result is used in the applications section for feature ranking. It is shown there that the bound follows extremely well the shape of the true error of a classifier based on a single feature and is much more precise than mutual information or normalized correlation indices.

To prove the optimality of direct mappings we start with the observation that for any clustering C a classification rule $q(Y|X)$ defined as

$$q(y|x) = \sum_c q(c|x)q(y|c) \tag{10}$$

achieves the same loss as the loss of a hypothesis h based on the clustering C. Therefore, the space of all direct mappings $q(Y|X)$ incorporates all possible solutions that may be achieved via intermediate clustering. It remains to show that the generalization power of the direct mappings is not worse than the generalization power of clustering-based solutions and that the global optimum may be efficiently found.

To analyze the generalization power of a direct mapping we define n_y clusters c_y, one for each label $y \in \mathcal{Y}$, i.e., $C_y = \{c_y : y \in \mathcal{Y}\}$. All instances x mapped to a cluster c_y obtain the label y. Thus the clustering C_y is identified with the labeling Y, in particular $q(C_y|X) = q(Y|X)$, and we can replace $H_U(C_y)$ in (4) with $H_U(Y)$ and $I_U(X;C_y)$ in (6) with $I_U(X;Y)$. Moreover, in our construction there are only $(n_y!)$ possibilities to assign the labels to the clusters and not $n_y^{n_y}$ as in the case of general clustering. In addition, the cardinality of C_y is fixed at n_y and does not change from 1 to n, where n is the cardinality of X. This further reduces a $\ln(n)$ factor from the bound. Thus, the definition of K in (5) is improved to:

$$K_y = \ln[\binom{n+n_y-1}{n_y-1}] + \frac{(\ln n+1)^2}{4} + \ln(n_y!) - \ln\delta. \tag{11}$$

(We used the tighter bound $\binom{n+n_y-1}{n_y-1}$ instead of n^{n_y-1} on the number of partitions.) And we get:

$$L(Q) \le \hat{L}(Q) + b\sqrt{\frac{nH_U(Y) + K_y}{2N}} \tag{12}$$

instead of (4) and

$$L(Q) \le \hat{L}(Q) + b\sqrt{\frac{nI_U(X;Y) + K'_y}{2N}} \tag{13}$$

instead of (6) for $K'_y = K_y + \frac{1}{2}\ln(4N)$.

For any other clustering C the direct mapping $q(Y|X)$ defined by (10) satisfies $I_U(X;C) \ge I_U(X;Y)$ by the information processing inequality (Cover & Thomas, 1991). Furthermore, since in $\mathcal{H}$ every partition cell gets a single label, $H_U(Y|C) = 0$. Therefore, $H_U(Y) \le H_U(C)$ because $H_U(Y) = H_U(Y) - H_U(Y|C) = I_U(C;Y) = H_U(C) - H_U(C|Y) \le H_U(C)$. Adding the fact that the empirical losses are equal for the clustering-based classification and the associated direct mapping we obtain that both (12) and (13) for the direct mapping are tighter than (4) and (6) for the corresponding clustering solution.

We can further optimize (13) by looking for an optimal classification rule $q^*(Y|X)$ that minimizes it. The minimum is achieved by iteration of the following self-consistent equations, where $\hat{p}(x,y)$ is the empirical joint distribution of X and Y (the derivation is done by taking a derivative of the bound with respect to $q(Y|X)$ and is omitted due to lack of space):

$$q(y|x) = \frac{q(y)}{Z(x)} e^{-\frac{2}{b}\left(\sum_{y'}\hat{p}(x,y')l(y',y)\right)\sqrt{2N(nI_U(X;Y)+K'_y)}}, \tag{14}$$

$q(y) = \frac{1}{n}\sum_x q(y|x)$, $Z(x) = \sum_y q(y|x)$, and $I_U(X;Y) = \frac{1}{n}\sum_{x,y} q(y|x)\ln[q(y|x)/q(y)]$. Although $\sqrt{I_U(X;Y)}$ is not necessarily convex, in our experiments the iterations always converged to a global optimum. It is also possible to optimize a parameterized tradeoff $\hat{L}(Q) + \beta I_U(X;Y)$, which is convex since both mutual information $I_U(X;Y)$ and the empirical loss $\hat{L}(Q)$ are convex with respect to $q(Y|X)$. A linear search over β then leads to a global optimum of (13).

Note that the direct mapping is no longer optimal when there is more than one parameter. For example, for two parameters X_1, X_2, each with a cardinality n, the conditional distribution $p(Y|X_1,X_2)$ is defined over the product space of size $n^2 n_y$. This requires at least an order of $n^2 n_y$ samples - a number quadratic in n - for the direct inference to be possible. However, from (4) and (6) it follows that with grid clustering for relatively small cluster cardinalities m_i it may be possible to achieve reliable estimations when the sample size N is linear in n. This is further discussed in the next section.

A related bound for generalization in prediction by a single feature is suggested in (Sabato & Shalev-Shwartz, 2007). Sabato and Shalev-Shwartz designed an estimator for the loss of a prediction rule based on the empirical frequencies $q_{emp}(y|x) = \hat{p}(y|x)$. They prove that their estimate is at most $O(\frac{\ln(N/\delta)\sqrt{\ln(1/\delta)}}{\sqrt{N}})$ far from the generalization error of q_{emp}. Compared to their work, a strong advantage of bounds (12) and (13) is that they hold for any prediction rule $q(Y|X)$. In particular, they hold for the maximum likelihood prediction $q_{ml}(x) = \arg\max_y \hat{p}(y|x)$ that performs much better than q_{emp} in practice.

6. A Bound for Estimation of a Joint Probability Distribution in Grid Clustering

For a fixed set of mappings $\{q_i(C_i|X_i)\}$ denote by $p(Y,\bar{C})$ the joint probability distribution of Y and $\bar{C}$, where $\bar{C}$ stays for $\langle C_1,..,C_d\rangle$ for brevity. Denote by $\hat{p}(Y,\bar{C})$ its empirical counterpart. Clearly, $p(Y,\bar{C})$ is determined by $p(Y,X_1,..,X_d)$ and the set $\{q_i(C_i|X_i)\}$. In this section we bound the deviation between $p(Y,\bar{C})$ and its empirical estimation.

Theorem 4. *For any probability measure p over instances and an i.i.d. sample S of size N according to p, with a probability of at least $1-\delta$ for all grid clusterings $Q = \{q_i(C_i|X_i)\}_{i=1}^d$ the following holds:*

$$D(\hat{p}(Y,\bar{C})\|p(Y,\bar{C})) \le \frac{\sum_i n_i H_U(C_i) + K_2}{N} \tag{15}$$

$$K_2 = \sum_i \left(m_i \ln n_i + \frac{(\ln(n_i)+1)^2}{4} \right)$$

$$+ n_y \left(\prod_i m_i \right) \ln(N+1) - \ln \delta. \qquad (16)$$

Proof. The proof is based on the law of large numbers cited below (Cover & Thomas, 1991).

Theorem 5 (The Law of Large Numbers)**.** *Let $Z_1, .., Z_N$ be i.i.d. distributed by $p(Z)$. Then:*

$$P\{D(\hat{p}(Z)\|p(Z)) > \varepsilon\} \leq e^{-N\varepsilon + |Z|\ln(N+1)}, \qquad (17)$$

where $|Z|$ stays for the cardinality of Z.

Note that the cardinality of the random variable $\langle Y, \bar{C}\rangle$ is $n_y \prod_i m_i$. For the proof of theorem 4 we require that the right hand side of (17) be smaller than $P(h)\delta$. Application of a union bound and reversion of the requirement on ε bounds $D(\hat{p}(Y,\bar{C})\|p(Y,\bar{C}))$ for the case of hard partitions by $\frac{n_y(\prod_i m_i)\ln(N+1) - \ln(P(h)\delta)}{N}$ for all h. Since $D(p\|q)$ is convex in the pair (p, q) (Cover & Thomas, 1991, Theorem 2.7.2), for soft partitions $D(\hat{p}(Y,\bar{C})\|p(Y,\bar{C}))$ is bounded from above by $\mathbb{E}_Q \frac{n_y(\prod_i m_i)\ln(N+1) - \ln(P(h)\delta)}{N}$. The calculation of $-\mathbb{E}_Q \ln P(h)$ done earlier completes the proof.

Applying the inequality relating the L_1 norm and the KL divergence $\|P_1 - P_2\|_1 \leq \sqrt{2D(P_1\|P_2)}$ (see (Cover & Thomas, 1991)) we obtain a bound on the variational distance.

Corollary 1. *Under the conditions of theorem 4:*

$$\|p(Y,\bar{C}) - \hat{p}(Y,\bar{C})\|_1 \leq \sqrt{\frac{2\left(\sum_i n_i H_U(C_i) + K_2\right)}{N}} \qquad (18)$$

7. Generalization Bound for the Logarithmic Loss in Grid Clustering

The goal of this section is to provide a bound on the logarithmic loss $-\mathbb{E}\ln \hat{p}(Y|\bar{C}) = -\sum_{y,\bar{c}} p(y,\bar{c}) \ln \hat{p}(y|\bar{c})$. This loss corresponds to the prediction (and compression) power of the hypothesis. Since *ln* is an unbounded function and $\hat{p}(y|c)$ is not bounded from zero, we define a smoothed distribution:

$$p^*(y|\bar{c}) = \frac{\hat{p}(y|\bar{c}) + \gamma}{n_y\gamma + 1},$$

where $\gamma > 0$ is the smoothing parameter. To complete the definition: $p^*(\bar{c}) = \hat{p}(\bar{c})$ and $p^*(y,\bar{c}) = p^*(\bar{c})p^*(y|\bar{c})$. Instead of proving the bound for $\hat{p}$ it will be proved for p^*:

$$-\mathbb{E}\ln p^*(Y|\bar{C}) = -\sum_{y,\bar{c}} p(y,\bar{c}) \ln p^*(y|\bar{c})$$

$$= \sum_{y,\bar{c}} (\hat{p}(y,\bar{c}) - p(y,\bar{c})) \ln p^*(y|\bar{c}) - \sum_{y,\bar{c}} \hat{p}(y,\bar{c}) \ln p^*(y|\bar{c})$$

$$\leq \frac{1}{2}\|p(Y,\bar{C}) - \hat{p}(Y,\bar{C})\|_1 \ln \frac{n_y\gamma + 1}{\gamma}$$

$$- \sum_{y,\bar{c}} \hat{p}(y,\bar{c}) \ln(\hat{p}(y|\bar{c}) + \gamma) + \ln(n_y\gamma + 1) \qquad (19)$$

$$\leq \varepsilon \ln \frac{n_y\gamma + 1}{\gamma} - \sum_{y,\bar{c}} \hat{p}(y,\bar{c}) \ln \hat{p}(y|\bar{c}) + \ln(n_y\gamma + 1)$$

$$= \hat{H}(Y|\bar{C}) + \varepsilon \ln \frac{1}{\gamma} + (\varepsilon + 1)\ln(n_y\gamma + 1), \qquad (20)$$

where inequality (19) is justified by (18) and ε is defined as half of its right hand side. $\hat{H}(Y|\bar{C})$ stays for the empirical estimation of the entropy of Y given $\bar{C}$. Equation (20) is minimized for $\gamma = \frac{\varepsilon}{n_y}$, when we get:

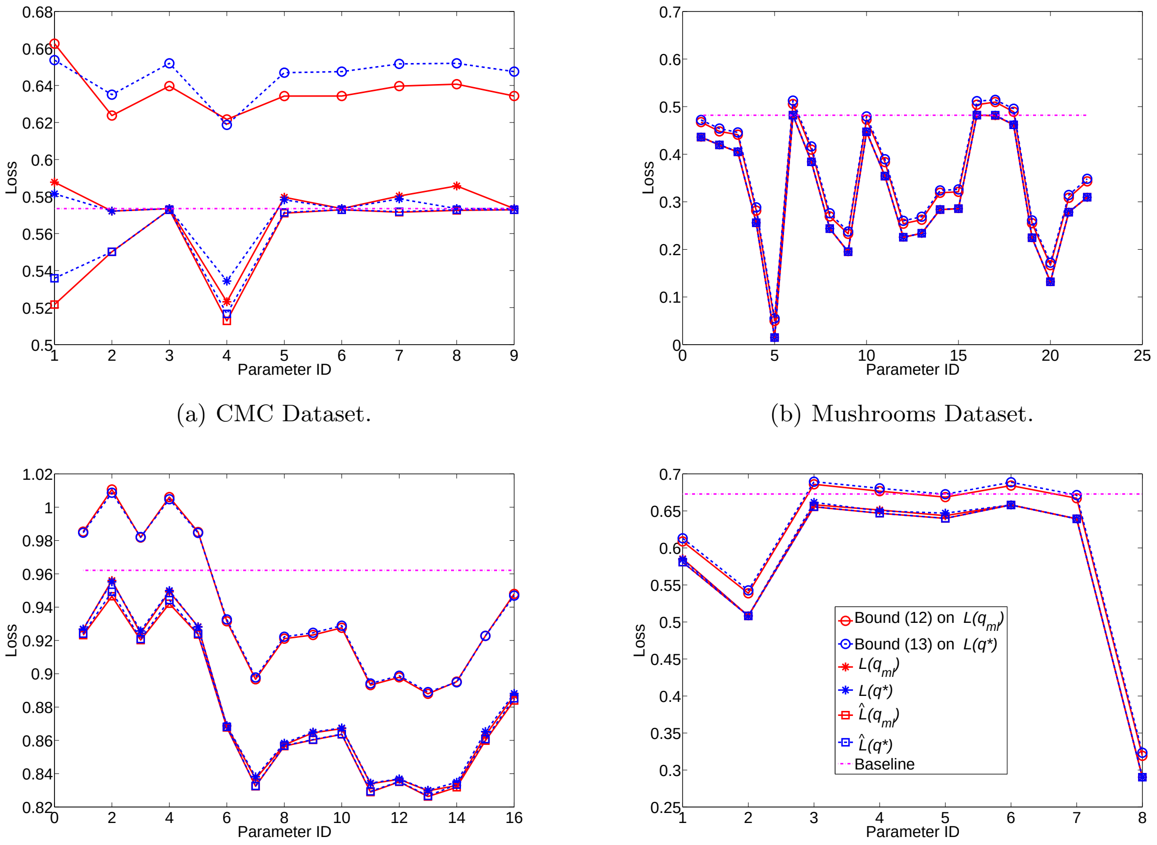
$$-\mathbb{E}\ln p^*(Y|\bar{C}) \leq \hat{H}(Y|\bar{C}) + \varepsilon \ln \frac{n_y}{\varepsilon} + (\varepsilon + 1)\ln(\varepsilon + 1). \qquad (21)$$

One natural application of the bound (21) to be studied in future work is to the broadly used "bag-of-words" models, where a decision is made based on multiple observations with the conditional independence assumption on the observations given the label. For example, in the bag of words model for document classification by topic we assume that the words are independent given a topic (the label Y). There is a single parameter X coming from the space of all words, but the classification is based on multiple observations of this parameter - all words in the document. Since we have a single parameter we can resort to the direct mappings $q(Y|X)$, as in section 5. Usually, a topic that maximizes the log likelihood of all the words in a document is assigned. After simple algebraic manipulations this can be translated to maximization of a sum of $\ln[q(y|x)]$ over the document (up to correct normalization by $\ln[q(y)]$), which is directly related to the expectation bounded in (21).

A related work in this context (Shamir, Sabato & Tishby, 2008) uses some different techniques to derive a bound for $|H(Y|C) - \hat{H}(Y|C)|$. We note that $H(Y|C) = -\mathbb{E}\ln p(Y|C)$ is the minimal logarithmic loss that could be achieved if we knew the true joint distribution $p(Y, C)$. Thus, (Shamir et al., 2008) give a lower bound on the performance of any prediction model based on grid clustering, whereas (21) is an upper bound on the performance of the prediction strategy $p^*(Y|\bar{C})$.

8. Applications

In this section we provide a series of applications of the bounds (12) and (13) to prediction by a single feature



(a) CMC Dataset.

(b) Mushrooms Dataset.

(c) Letters Dataset.

(d) Nursery Dataset

Figure 1. **Application of bounds (12) and (13).** This figure displays an application of bounds (12) and (13) to the four datasets discussed in text. The legend in subfigure (d) corresponds to all the graphs. The graphs contain the training loss $\hat{L}(q_{ml})$, the test loss $L(q_{ml})$ and the value of the bound (12) for the maximum likelihood prediction rule $q_{ml}(x) = \arg\max_y \hat{p}(y|x)$. A second triplet on the graphs corresponds to $\hat{L}(q^*)$, $L(q^*)$, and the value of the bound (13) for the prediction rule $q^*(Y|X)$ that minimizes (13). Baseline corresponds to the performance level that can be achieved by predicting the test labels using a marginal distribution of Y on the train set. All the calculations are done per parameter. For better visibility of the points they have been connected with lines, but the lines have no meaning.

and feature ranking, as suggested in section 5. We use (12) to bound the generalization error of the maximum likelihood classification rule. For zero-one loss the maximum likelihood rule $q_{ml}(X)$ returns for each value of x the most frequent value of Y that appeared with that x in the sample: $q_{ml}(x) = \arg\max_y \hat{p}(y|x)$. We also use the iterations (14) to find a classification rule $q^*(Y|X)$ that minimizes (13).

The experiments were conducted on four datasets obtained at the UCI Machine Learning Repository: Contraceptive Method Choice (CMC), Mushrooms, Letters and Nursery. In all the experiments we use 5 random partitions of the data into 80% train and 20% test subsets. Table 1 provides a short summary of the main parameters of the datasets. See (Asuncion & Newman, 2007) for a full description.

Figure 1 shows the training loss and the test loss of the maximum likelihood classification rule $q_{ml}(Y|X)$ for the four datasets considered. We stress that the maximum likelihood rule is calculated per parameter; actually there are d maximum likelihood rules $q_{ml}(Y|X_i)$, one for each parameter i of a given problem. Along with the test loss we draw the value of the bound (12). Note that the bound is quite tight and follows the shape of the test loss remarkably well in all the cases. The gap between the bound and the test loss is less than 0.1.

The same figure includes an additional triplet of lines - training loss, test loss, and the bound (13) value - corresponding to the $q^*(Y|X)$ classification rule that minimizes (13). The performance of q^* is very close to the performance of q_{ml} and the value of (13) is very

Table 1. **Description of the datasets:** for every set we give the number of features, d, a list of cardinalities of the features, n_i, the number of labels, n_y, and a train set size, N, which is 80% of a dataset size.

Data set	d	n_i-s	n_y	N
CMC	9	34, 4, 4, 15, 2, 2, 4, 4, 2	3	1,178
Mushrooms	22	6, 4, 10, 2, 9, 2, 2, 2, 12, 2, 5, 4, 4, 9, 9, 1, 4, 3, 5, 9, 6, 7, 2	2	6,499
Letters	16	16 for all n_i-s	26	16,000
Nursery	8	3, 5, 4, 4, 3, 2, 3, 3	5	10,368

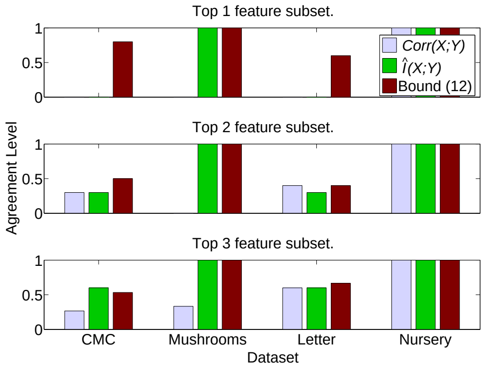


Figure 2. **Feature Ranking.** Agreement of $Corr(X;Y)$, $\hat{I}(X;Y)$, and the bound (12) with the test set on the top-1, top-2, and top-3 feature subsets.

close to the value of (12) with a small advantage to q_{ml} and (12) on average.

We conclude this section by comparing the bound (12) applied to feature ranking with the standard empirical mutual information $\hat{I}(X;Y) = \sum_{x,y} \hat{p}(x)\hat{p}(y|x) \ln \frac{\hat{p}(y|x)}{\hat{p}(y)}$ and the normalized correlation coefficient $Corr(X;Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ indices. We compare agreement between the top-1, top-2, and top-3 parameter subsets suggested by the indices with the corresponding test-based sets - Figure 2. For the top-1 choice (the best single parameter) our bound is clearly superior - it provides a significant level of success in two cases where the other two indices completely fail. For the top-2 choice there is a slight advantage over the mutual information and a clear advantage over the normalized correlation. In top-3 the bound performs similarly to the mutual information and is still superior to the normalized correlation.

9. Discussion

This paper derives generalization bounds for multi-classification based on grid clustering. The bounds enable evaluation of clustering solutions based on generalization properties of a built-on classifier. We acknowledge that the $(\prod_i m_i) \ln n_y$ term in the bounds limits their applicability to relatively few dimensional problems. Nevertheless, this domain contains enough challenges such as feature ranking, where our bounds are especially tight, collaborative filtering and many more. An interesting direction for future work would be to extend the applicability of the approach to higher dimensions by utilizing dependencies between the parameters.

Acknowledgements: We thank Ohad Shamir for useful discussions. This work was partially supported by Leibnitz Center for Research in Computer Science and the NATO SfP Programme.

References

Asuncion, A., & Newman, D. (2007). UCI machine learning repository. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Blanchard, G., & Fleuret, F. (2007). Occam's hammer. *COLT*.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons.

Langford, J. (2005). Tutorial on practical prediction theory for classification. *JMLR*, *6*.

Maurer, A. (2004). A note on the PAC-Bayesian theorem. www.arxiv.org.

McAllester (1999). Some PAC-Bayesian theorems. *Machine Learning*, *37*.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*.

Sabato, S., & Shalev-Shwartz, S. (2007). Prediction by categorical features: Generalization properties and application to feature ranking. *COLT*.

Seldin, Y., Slonim, N., & Tishby, N. (2007). Information bottleneck for non co-occurrence data. *NIPS*.

Shamir, O., Sabato, S., & Tishby, N. (2008). Learning and generalization with the information bottleneck method. Preprint.

Srebro, N. (2004). *Learning with matrix factorizations*. Doctoral dissertation, MIT.