
Generalization in Clustering with Unobserved Features

Eyal Krupka and Naftali Tishby
School of Computer Science and Engineering,
Interdisciplinary Center for Neural Computation

Table 1: Analogy with supervised learning

Training set	n randomly selected features (observed features)
Test set	Unobserved features
Learning algorithm	Cluster the <i>instances</i> into k clusters
Hypothesis class	All possible partitions of m instances into k clusters
Min generalization error	Max expected information on <i>unobserved</i> features
ERM	Maximize mean

Proof: For fixed cluster labels, t_1, \dots, t_m

of $(T; X_j)$ over the m^0 instances. This distribution is denoted by $\hat{P}(t; x_j)$ and the corresponding mutual information is denoted by $I_{\hat{P}}(T; X_j)$. Theorem 1 is build up from the following upper bounds, which are independent of m , but depend on the choice of m^0 . The first bound is on $E \sum_j I(T; X_j) \leq m I_{\hat{P}}(T; X_j)$, where the expectation is over random selection of the m^0 instances. From this bound we derive upper bounds on $J_{ob} \leq E \sum_j I(T; X_j) \leq m I_{\hat{P}}(T; X_j)$.

3 Empirical Evaluation

In this section we describe an experimental evaluation of the generalization properties of the *JobMax* algorithm for a finite large number of features. We examine the difference between l_{ob} and l_{un} as function of the number of observed features and the number of

