PAC-Bayesian Bounds for Discrete Density Estimation and Co-clustering Analysis

Yevgeny Seldin[*,†]
[*]Max Planck Institute
for Biological Cybernetics
Tübingen, Germany
seldin@tuebingen.mpg.de

Naftali Tishby[†,‡]
[†]School of Computer Science and Engineering
[‡]Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem, Israel
tishby@cs.huji.ac.il

Abstract

We applied PAC-Bayesian framework to derive generalization bounds for co-clustering[1]. The analysis yielded regularization terms that were absent in the preceding formulations of this task. The bounds suggested that co-clustering should optimize a trade-off between its empirical performance and the mutual information that the cluster variables preserve on row and column indices. Proper regularization enabled us to achieve state-of-the-art results in prediction of the missing ratings in the MovieLens collaborative filtering dataset.

In addition a PAC-Bayesian bound for discrete density estimation was derived. We have shown that the PAC-Bayesian bound for classification is a special case of the PAC-Bayesian bound for discrete density estimation. We further introduced combinatorial priors to PAC-Bayesian analysis. The combinatorial priors are more appropriate for discrete domains, as opposed to Gaussian priors, the latter of which are suitable for continuous domains. It was shown that combinatorial priors lead to regularization terms in the form of mutual information.

1 Introduction

Co-clustering is a widely used approach to the analysis of data matrices by simultaneous clustering of "similar" rows and columns of the data matrix [2]. In [8] we identified two types of problems that are often solved by co-clustering. The first is discriminative prediction of the missing matrix entries and the second is estimation of a joint probability distribution of variables corresponding to rows and columns of the data matrix. Discriminative prediction corresponds to problems like collaborative filtering, where the missing ratings are discriminatively predicted given the ⟨viewer,movie⟩ pairs. Density estimation corresponds to problems such as analysis of word-document co-occurrence data, where the task is to learn the joint distribution of words and documents (rows and columns of a matrix).

For the purpose of analysis of generalization properties of co-clustering we found convenient to apply the PAC-Bayesian framework [4]. The key for success of PAC-Bayesian analysis lies in the ability to slice a hypothesis space in an intelligent way. For example, differentiation of separating hyperplanes by the size of the margin combined with PAC-Bayesian analysis enabled the derivation of state-of-the-art generalization bounds for Support Vector Machines [3, 5]. In [7] we suggested an intelligent partition of the space of co-clustering solutions that yielded meaningful and practically useful bounds for this problem. We defined a prior over this space by combinatorial counting of the hypotheses according to the partition and showed that this form of a prior leads to regularization terms in the form of mutual information. For the analysis of density estimation with co-clustering, we have extended the PAC-Bayesian framework and derived and PAC-Bayesian bounds for discrete density estimation.

2 Main Results

Due to space limitations we present only a subsample of the results, for further details refer to [7, 8, 6].

2.1 Discrete Density Estimation

Theorem 1. *Let $\mathcal{X}$ be the sample space and let $p(X)$ be an unknown distribution over $X \in \mathcal{X}$. Let $\mathcal{H}$ be a hypothesis class, such that each $h \in \mathcal{H}$ is a function from $\mathcal{X}$ to a finite set $\mathcal{Z}$. Let $p_h(Z) = P_{X\sim p(X)}\{h(X) = Z\}$ be the distribution over $\mathcal{Z}$ induced by $p(X)$ and h. Let $\mathcal{P}$ be a prior distribution over $\mathcal{H}$. Let $\mathcal{Q}$ be an arbitrary distribution over $\mathcal{H}$ and $p_{\mathcal{Q}}(Z) = \mathbb{E}_{\mathcal{Q}(h)} p_h(Z)$ a distribution over Z induced by $p(X)$ and $\mathcal{Q}$. Let S be an i.i.d. sample of size N generated according to $p(X)$ and let $\hat{p}(X)$*

[1]This abstract surveys the results developed in [7, 8, 6].

be the empirical distribution over $\mathcal{X}$ corresponding to S. Let $\hat{p}_h(Z) = P_{X\sim\hat{p}(X)}\{h(X) = Z\}$ be the empirical distribution over Z corresponding to h and S and $\hat{p}_{\mathcal{Q}}(Z) = \mathbb{E}_{\mathcal{Q}(h)}\hat{p}_h(Z)$. Then with a probability greater than $1-\delta$ for all distributions $\mathcal{Q}$ simultaneously:

$$D(\hat{p}_{\mathcal{Q}}\|p_{\mathcal{Q}}) \leq \frac{D(\mathcal{Q}\|\mathcal{P}) + (|Z|-1)\ln(N+1) - \ln\delta}{N}, \quad (1)$$

where[2] *$\hat{p}_{\mathcal{Q}} \equiv \hat{p}_{\mathcal{Q}}(Z)$ and $p_{\mathcal{Q}} \equiv p_{\mathcal{Q}}(Z)$ for brevity and $D(\cdot\|\cdot)$ is the KL-divergence.*

The PAC-Bayesian bound for classification [4] is a special case of the PAC-Bayesian bound for density estimation. In order to illustrate this, let Z be the error variable. Then each $h \in \mathcal{H}$ is a function from the sample space of pairs $\langle X, Y\rangle$ to the error variable Z and $|Z| = 2$. Furthermore, $\hat{L}(\mathcal{Q}) = P_{\hat{p}(X,Y)}\{Z=1\}$ and $L(\mathcal{Q}) = P_{p(X,Y)}\{Z=1\}$. Substituting this into (1) yields the PAC-Bayesian bound for classification.

The proof of theorem 1 is based on applying the law of large numbers to show that for a single hypothesis $\mathbb{E}_S e^{ND(\hat{p}_h(Z)\|p_h(Z))} \leq (N+1)^{|Z|-1}$ and then treating distributions $\mathcal{Q}$ over $\mathcal{H}$ by applying change of measure inequality (also called compression lemma [1]).

2.2 Co-clustering Analysis

We present the PAC-Bayesian bound for discriminative prediction with co-clustering only. For the PAC-Bayesian bound for density estimation with co-clustering refer to [8, 6]. We consider the following form of discriminative predictors:

$$q(Y|X_1,..,X_d) = \sum_{C_1,..,C_d} q(Y|C_1,..,C_d)\prod_{i=1}^{d} q(C_i|X_i).$$

In the collaborative filtering example $d=2$, Y is the rating, X_1 is viewer ID and X_2 is movie ID. The conditional probability distribution $q(C_i|X_i)$ represents the probability of assigning X_i to cluster C_i. The conditional probability $q(Y|C_1,..,C_d)$ represents the probability of assigning label Y to cell $\langle C_1,..,C_d\rangle$ in the cluster product space. We denote collectively the free parameters of the model by $\mathcal{Q} = \left\{\{q(C_i|X_i)\}_{i=1}^d, q(Y|C_1,..,C_d)\right\}$. In [7] it is shown that $\mathcal{Q}$ is a distribution over hypothesis space of hard partitions of the parameter space $\mathcal{X}_1\times..\times\mathcal{X}_d$. We denote $L(\mathcal{Q}) = \mathbb{E}_{p(X_1,..,X_d,Y)}\mathbb{E}_{q(Y'|X_1,..,X_d)}l(Y,Y')$ and $\hat{L}(\mathcal{Q}) = \mathbb{E}_{\hat{p}(X_1,..,X_d,Y)}\mathbb{E}_{q(Y'|X_1,..,X_d)}l(Y,Y')$, where $l(Y,Y')$ is a given loss function for predicting Y' instead of Y. We define $\bar{q}(c_i) = \frac{1}{|X_i|}\sum_{x_i} q(c_i|x_i)$ to be the marginal distribution over C_i corresponding to $q(C_i|X_i)$ and a *uniform* distribution over X_i. We define the mutual information corresponding to the joint distribution $\bar{q}(x_i,c_i) = \frac{1}{|X_i|}q(c_i|x_i)$ defined by $q(c_i|x_i)$ and the uniform distribution over X_i as $\bar{I}(X_i;C_i) = \frac{1}{|X_i|}\sum_{x_i,c_i} q(c_i|x_i)\ln\frac{q(c_i|x_i)}{\bar{q}(c_i)}$.

Theorem 2. *For any probability distribution $p(X_1,..,X_d,Y)$ over $\mathcal{X}_1\times..\times\mathcal{X}_d\times\mathcal{Y}$ and for any loss function l bounded by 1, with a probability of at least $1-\delta$ over selection of an i.i.d. sample of size N according to p, for all randomized classifiers $\mathcal{Q} = \left\{\{q(C_i|X_i)\}_{i=1}^d, q(Y|C_1,..,C_d)\right\}$:*[3]

$$D_b(\hat{L}(\mathcal{Q})\|L(\mathcal{Q})) \leq \frac{\sum_{i=1}^d |X_i|\bar{I}(X_i;C_i) + K}{N}, \quad (2)$$

where $K = \sum_i |C_i|\ln|X_i| + (\prod_i |C_i|)\ln|Y| + \ln\frac{N+1}{\delta}$.

Theorem 2 is obtained by defining a prior via combinatorial counting of the possible ways to cluster X_i-s and calculating the KL-divergence of $\mathcal{Q}$ from this prior, which yields the mutual information term [7, 6]. As shown previously [6], regularization by mutual information provides state-of-the-art predictions on the MovieLens dataset.

3 Discussion and Future Work

As shown in previous work [6], theorem 2 can be generalized to graphical models having tree shape. Interesting directions for future research include generalization of theorem 1 to continuous density estimation and theorem 2 to more general graphical models.

References

[1] A. Banerjee. On Bayesian bounds. In *ICML*, 2006.

[2] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. Modha. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *JMLR*, 8, 2007.

[3] J. Langford and J. Shawe-taylor. PAC-Bayes & margins. In *NIPS*, 2002.

[4] D. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37, 1999.

[5] D. McAllester. Simplified PAC-Bayesian margin bounds. In *COLT*, 2003.

[6] Y. Seldin. *A PAC-Bayesian Approach to Structure Learning*. PhD thesis, The Hebrew University, 2009.

[7] Y. Seldin and N. Tishby. Multi-classification by categorical features via clustering. In *ICML*, 2008.

[8] Y. Seldin and N. Tishby. PAC-Bayesian generalization bound for density estimation with application to co-clustering. In *AISTATS*, 2009.

[2] Throughout the abstract $|\cdot|$ stands for cardinality of a corresponding variable.

[3] $D_b(p\|q)$ stands for the KL-divergence between two Bernoulli variables with biases p and q.