APPROXIMATION SET CODING FOR INFORMATION THEORETIC MODEL VALIDATION

Alberto Giovanni Busetto[1,2], *Morteza Haghir Chehreghani*[1], *Joachim M. Buhmann*[1,2]

[1] Department of Computer Science, ETH Zurich, Zurich, Switzerland
[2] Competence Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland

ABSTRACT

Models can be seen as mathematical tools aimed at prediction. The fundamental modeling question is: which model best generalizes the available data? We discuss the central ideas of a recently introduced principle for model validation: Approximation Set Coding (ASC). The principle is inspired by concepts from statistical physics and it is based on information theory. There exists a central analogy between communication and learning which can be used to evaluate informativeness by designing codes based on sets of solutions. These sets are called approximation sets; they should be small enough to be informative and large enough to be stable under noise fluctuations. We present the application of ASC to two tasks: clustering and learning of logical propositions. The two modeling tasks highlight the generality of the principle and its main properties. Experimental results are discussed in the biological application domain.

1. INTRODUCTION

In the context of modeling, validation constitutes a fundamental step. The central question is: which model should be selected given the data? A justified answer to this question requires a precise assessment of the predictive capability of candidate models.

Our problem definition explicitly considers the case in which models are defined in terms of cost functions. This setting is in contrast to the more restrictive (yet still interesting) one in which a specific cost is given *a priori* and the estimation process solely consists of selecting the best parameters from a set. In our case, model selection consists of finding the most informative cost. To do that, we must define and estimate informativeness.

Let us start by introducing cluster model selection as a motivating example. We define a solution of a clustering analysis as an assignments of labels to samples. Clustering, hence, produces partitions of the available sample points. Alternative partitions are evaluated and selected on the basis of a cost function. The cost function (that is, the model) is often made explicit, but may also be implicitly defined in terms of outputs of an algorithmic process. In applications, the cost function is typically chosen according to human intuition and remains fixed for the analysis. For simplicity, let us now consider a clustering procedure based on an explicit cost function $R(\cdot|X)$, which evaluates solutions on the basis of the dataset X. Given X, the learning process terminates as soon as a (globally or locally) optimal solution is found. At this point, two important issues remain open. Is the result informative? Is the model justified? In order to answer these questions, we need a precise definition of the modeling goal in terms of predictive capabilities. There already exist theoretical and practical answers to these questions. At present, the set of established principles and procedures for predictive modeling include Minimum Description Length [1], Kolmogorov Structure Function [2], BIC [3] & AIC [4], Minimum Message Length [5], Solomonoff's Induction [6, 7], PAC [8] and PAC-Bayesian generalization bounds [9]. These approaches are based on convincing justifications from information theory, algorithmic information theory, probability and statistical learning theory.

The discussion of the individual merits of these approaches is certainly of great interest and value but goes beyond the scope of this contribution. We focus on the recently introduced idea of Approximation Set Coding [10]. ASC shares the spirit of the mentioned approaches, but with a rather different goal: selecting models by measuring the informativeness of equivalence classes of solutions.

2. APPROXIMATION SET CODING

ASC selects the optimal quantization of the hypothesis class to find the set of hypotheses constituting the best tradeoff between informativeness and stability. The informal justification is the following. On the one hand, selecting very few solutions exposes the modeler to the danger of instability with respect to fluctuations induced by noise [11]. On the other hand, selecting many solutions yields stable but rather uninformative results. With minimalistic assumptions about the nature of the noise, it is possible to select the set of solutions which provides the best tradeoff between informativeness and stability. This optimal set constitutes the best approximation available for a model. Models are then compared in terms of their informativeness, finally yielding the optimal approximation set.

Let us now start by formalizing the central concepts. Consider a cost model $R(c|X)$, which evaluates the cost of choosing solution $c \in \mathcal{C}(X)$ to generalize the given dataset $X \in \mathcal{X}$. As conventional in statistical learning theory, the smaller the cost, the better is the quality of the solution. The set of all candidate solutions is defined as the hypothesis class $\mathcal{C}(X)$, which is given to the modeler. Depending on the application, individual solutions might be parametric (with variable parameters) or simple elements from a set. In both cases, each element c of the hypothesis class indicates a particular and fixed candidate solution. Different cost functions define different models (for instance $R_1(c|X)$ and $R_2(c|X)$); for the rest of the manuscript, we identify models with their respective cost function. Our task is then to evaluate a set of models and select the best one, that is the most predictive. For each cost model $R(\cdot|X)$ and a given dataset, the optimal solutions are provided by the set of empirical minimizers

$$\mathcal{C}^{\perp}(X) = \arg \min_{c \in \mathcal{C}(X)} R(c|X). \tag{1}$$

Since costs are evaluated as a function of the data, we must take into account the variability with respect to X. In order to perform this step, we consider the minimal case in which two datasets (each of size n) are available to the modeler. The extension to settings with a larger number of sample sets is straightforward and exhibits analogous results. We assume that two datasets X_1 and X_2 are drawn independently from the same distribution. Since the hypothesis class might also depend on the dataset, we need a way to map solutions from $\mathcal{C}(X^1)$ to $\mathcal{C}(X^2)$. Transferring solutions between instances is a necessary requirement to evaluate the generalization properties from training to test data. For that, we introduce the mapping function $\psi : \mathcal{C}(X^1) \to \mathcal{C}(X^2)$.

By mapping the solutions from one dataset to another, ψ allows the modeler to map solutions across instances (for instance, by mapping to the nearest neighbor). For every subset of solutions $A \subseteq \mathcal{C}(X_1)$, we denote the mapped subset as

$$\psi \circ A = \{\psi(a),\ a \in A\} \subseteq \mathcal{C}(X_2). \tag{2}$$

In case of noise, the set of mapped empirical minimizers do not necessarily coincide with the solutions induced by the second dataset. The intersection $\psi \circ \mathcal{C}^{\perp}(X_1) \cap \mathcal{C}^{\perp}(X_2)$ might be small or even empty. In fact, fluctuations in the data might induce perturbations in the empirical minimizers, which will tend to diverge from each other as the noise level increases. Instead of taking the two sets of empirical minimizers (to avoid inconsistency due to instability), we consider larger sets of solutions. These sets are called approximation sets and are defined as a function of a parameter γ so that

$$\mathcal{C}_{\gamma}(X_i) = \{c \in \mathcal{C}(X_i) :\ R(c|X_i) \le R_{\perp}(X_i) + \gamma\} \tag{3}$$

for $i = 1, 2$. These sets are γ-close to the solution costs $R^{\perp}(X_i) := R(c_i^{\perp}|X_i)$ of the respective empirical minimizers $c_i^{\perp} \in \mathcal{C}^{\perp}(X_i)$, $i = 1, 2$. At this point, the question is which γ should we select? For $\gamma = 0$ we get only the empirical minimizers. If γ is too small, the results are unstable. For too large γ, the selection tends to include all the entire hypothesis class (thus yielding uninformative results). The communication analogy is introduced to address this question. It is based on the sender-receiver scenario in which distinguishing individual solutions based on data corresponds to transmitting messages over a noisy channel. The communication capacity reflects the ability to discriminate solutions through the applied transformations. Ultimately, the success of the communication depends on noise level and coding strategy.

The communication process for a certain γ is described by the following procedures:

- *Coding:*
  1. Sender and receiver agree on R and share X_1.
  2. They both calculate the γ-approximation sets.
  3. The sender generates a set of transformations $\Sigma = \{\sigma : \mathcal{X} \to \mathcal{X}\}$ which define a set of training optimization problems $R(\cdot|\sigma \circ X_1)$ and their respective γ-approximation sets.
  4. The sender sends Σ to the receiver which calculates the approximation sets for each transformation.

- *Transmission:*
  1. The sender is a stationary source: it selects a transformation σ_s as message without directly revealing it to the receiver.
  2. The transformation σ_s is applied by the sender to X_2.
  3. The transformed dataset $\sigma_s \circ X_2$ is sent to the receiver.
  4. The receiver has to reconstruct the transformation σ_s from the approximation set of $\sigma \circ X_2$ without directly knowing X_2 and σ_s.

Each transformation σ_s generated by the sender is estimated by the receiver through the decoding rule

$$\widehat{\sigma} = \arg \max_{\sigma \in \Sigma} \left| \psi \circ \mathcal{C}_{\gamma}(\sigma \circ X_1) \cap \mathcal{C}_{\gamma}(\sigma_s \circ X_2) \right|. \tag{4}$$

Decoding is possible because, in contrast to σ_s and X_2, $\sigma_s \circ X_2$ is known to the receiver. It can be used to calculate the approximation sets used to uniquely identify σ_s. The aim is the following: achieving optimal communication (which is reliable and informative). Approximation sets define codebook vectors; while large γ correspond to small sets of distinct vectors for coding, small γ might correspond to higher error rates for decoding.

Communication errors are due to wrong decoding, that is when $\widehat{\sigma} \neq \sigma_s$. The probability of a communication error is hence given by

$$P(\widehat{\sigma} \neq \sigma_s|\sigma_s) = P\left(\max_{\sigma_j \in \Sigma \setminus \{\sigma_s\}} |\Delta\mathcal{C}_{\gamma}^{j}| \ge |\Delta\mathcal{C}_{\gamma}^{s}| \middle| \sigma_s \right), \tag{5}$$

where, for all $\sigma_j \in \Sigma$,

$$\Delta\mathcal{C}_{\gamma}^{j} = \psi \circ \mathcal{C}_{\gamma}(\sigma_j \circ X_1) \cap \mathcal{C}_{\gamma}(\sigma_s \circ X_2) \tag{6}$$

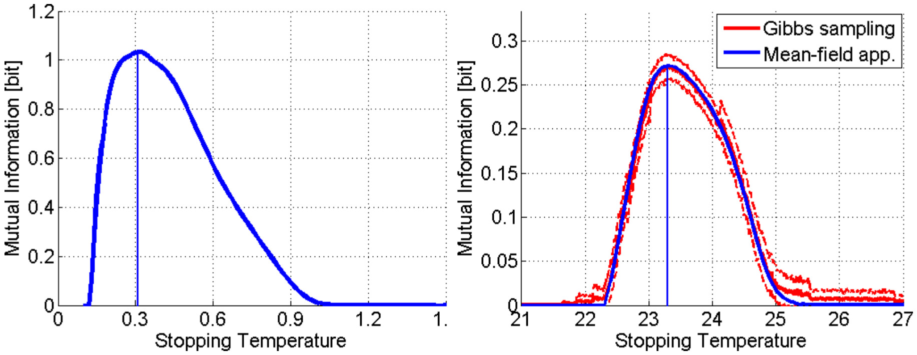

Figure 1. Comparison of the informativeness of pairwise clustering (left) and correlation clustering (right) in terms of AC for gene expression data. The former is approximately four times more informative than the latter. For correlation clustering, the mutual information is estimated by mean-field approximation and Gibbs sampling for comparison.

denotes the intersection between the j-th approximation set and that of the test set.

The direct evaluation of the error probability can be bounded through the union bound as follows:

$$P(\widehat{\sigma} \neq \sigma_s | \sigma_s) \leq \sum_{\sigma_j \in \Sigma \backslash \{\sigma_s\}} P\left(|\Delta \mathcal{C}_\gamma^j| \geq |\Delta \mathcal{C}_\gamma^s| \middle| \sigma_s \right), \tag{7}$$

Furthermore, one has that

$$P(\widehat{\sigma} \neq \sigma_s) \leq (|\Sigma| - 1) \exp\left(-n \mathcal{I}_\gamma(\sigma_s, \widehat{\sigma})\right), \tag{8}$$

where $\mathcal{I}_\gamma(\sigma_j, \widehat{\sigma})$ is the mutual information

$$\mathcal{I}_\gamma(\sigma_s, \widehat{\sigma}) = \frac{1}{n} \log \left(\frac{|\Sigma| \, |\Delta \mathcal{C}_\gamma^s|}{|\mathcal{C}_\gamma(X_1)| \, |\mathcal{C}_\gamma(X_2)|} \right). \tag{9}$$

The optimal γ is found solving

$$\gamma^* = \arg \max_{\gamma \in [0, \infty)} \mathcal{I}_\gamma(\sigma_s, \widehat{\sigma}). \tag{10}$$

This procedure provides to the modeler:

- a set of γ-optimal solutions, as well as
- a measure of the informativeness of the selected approximation set for the model R: the Approximation Capacity (AC) $\mathcal{I}_\gamma^*(\sigma_s, \widehat{\sigma})$.

This selection criterion enables the comparison of different models R for the cost of selecting solutions c given training and test.

3. APPLICATIONS AND RESULTS

Recently, ASC has been applied to perform model selection in clustering [12], yielding results consistent with BIC in the analysis of biological data. In clustering, Σ corresponds to the set of permutations of cluster labels. It is worth noting that in the case of clustering the cardinality of the hypothesis class grows exponentially with the sample size. This is because solutions are defined as label assignments in this application.

Experimental results in the context of gene expression analysis show that pairwise clustering [13] yields superior amounts of reliable information in comparison to correlation clustering [14]. Relational clustering problems are often defined with respect to an attributed graph $(\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. The vertices have to be clustered into groups $\mathcal{G}_u := \{i \ : \ c(i) = u\}, 1 \leq u \leq K$ where c is the cluster solution which assigns label u to the i-th sample. The set of edges between elements of group $\mathcal{G}_u$ and $\mathcal{G}_v$ is denoted by $\mathcal{E}_{uv} := \{(i,j) \ : \ c(i) = u \wedge c(j) = v\}$.

In both cases, the datasets consisted of matrices of pairwise similarities X. The pairwise clustering cost model is defined as

$$R_{\text{pc}}(c, X) = -\frac{1}{2} \sum_{k=1}^{K} |\mathcal{G}_k| \sum_{(i,j) \in \mathcal{E}_{kk}} \frac{X_{ij}}{|\mathcal{E}_{kk}|}, \tag{11}$$

where X_{ij} denotes the similarity between object i and j. The correlation clustering model is

$$R^{\text{cc}}(c, X) = \frac{1}{2} \sum_{1 \leq u \leq K} \sum_{(i,j) \in \mathcal{E}_{uu}} (|X_{ij}| - X_{ij}) + \frac{1}{2} \sum_{1 \leq u \leq K} \sum_{1 \leq v < u} \sum_{(i,j) \in \mathcal{E}_{uv}} (|X_{ij}| + X_{ij}).$$

Figure 1 shows the application to gene expression data with temporal structure (expression level time points for 12 consecutive months) [15]. The feature vector is splitted into two and the similarity matrices are constructed by taking the Pearson correlation coefficients for each pair of genes (295 differentially expressed genes). This dataset has been selected because it is one of the many cases in which the choice of a cost is challenging. The figure compares the AC of the two models, showing the advantage of pairwise clustering over correlation clustering. The result means that under identical noise effects, pairwise clustering discovers a more predictive structure than correlation clustering. ASC validates pairwise clustering ($\max_\beta \mathcal{I}_\beta = 1.03$, where β is the inverse computational temperature) as approximately 3.5 times more informative than correlation clustering ($\max_\beta \mathcal{I}_\beta = 0.272$). At the optimal resolution (temperature), 7 clusters are discovered by pairwise

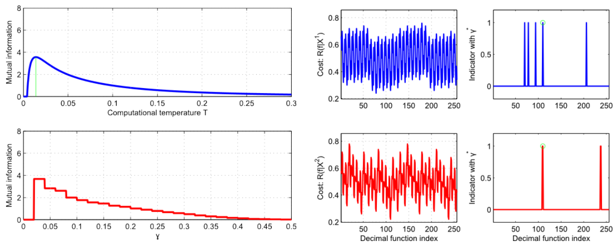

Figure 2. Calculation of mutual information and approximation sets for the Boolean case. On the left, the mutual information is calculated exactly and with the Boltzmann approximation (left top and bottom, respectively). The green line identifies the optimal computational temperature (no normalization). On the right, the model is evaluated for the two split datasets over the hypothesis class (decimal indexing of the Boolean outputs). The green dot indicates the membership of the data generator.

clustering (in contrast to the 2 clusters identified by correlation clustering). The number of clusters in pairwise clustering is also consistent with that obtained with BIC (with number of parameters calculated as the ratio between the trace and the largest eigenvalue of the similarity matrix).

To learn logical propositions we define the hypothesis class of Boolean functions of d literals. We consider both the supervised and the unsupervised case. In contrast to clustering, Σ is given by the set of distinguishable bitwise flips of the data (in input for the unsupervised case, and in both input and output in the supervised case). The set of transformations is therefore given by a set of local $\neg$ (NOT) operators applicable to the available sample components. Hence, in the unsupervised case the cardinality of the set of perturbations is smaller or equal to that of the hypothesis class:

$$|\Sigma| \leq |\mathcal{C}(X)| = 2^{2^d}. \tag{12}$$

The goal is the identification of predictive formulas which generalize the available binary observations. Figure 2 compares the exact solution and Boltzmann approximation with a dataset generated by the 110-th Boolean function with $d = 3$ subject to uniform sampling of the input. The bit flipping probability is 1/8 both for input and for output.

4. ACKNOWLEDGMENTS

We thank Marcus Hutter, Cheng Soon Ong, Ludwig Busse, Brian McWilliams and David Balduzzi for insightful discussions and helpful comments. The authors are financed with grants from the Swiss SystemsX.ch initiative (projects YeastX and LiverX), evaluated by the Swiss National Science Foundation.

5. REFERENCES

[1] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, pp. 465–471, 1978.

[2] A.N. Kolmogorov, “Complexity of algorithms and objective definition of randomness,” in *Talk at Moscow Math. Soc. Meet. (transl. from Russian by L. A. Levin)*, Moscow, Apr. 16 1974.

[3] G.E. Schwarz, “Estimating the dimension of a model,” *Ann. of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[4] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[5] C.S. Wallace and D.M. Boulton, “An information measure for classification,” *Computer Journal*, vol. 11, no. 2, pp. 185–194, 1968.

[6] R. Solomonoff, “A formal theory of inductive inference, part i,” *Information and Control*, vol. 7, no. 1, pp. 1–22, 1964.

[7] R. Solomonoff, “A formal theory of inductive inference, part ii,” *Information and Control*, vol. 7, no. 2, pp. 224–254, 1964.

[8] L. Valiant, “A theory of the learnable,” *Comm. of the ACM*, vol. 27, pp. 1134–1142, 1984.

[9] Yevgeny Seldin and Naftali Tishby, “Pac-bayesian analysis of co-clustering and beyond,” *J. Mach. Learn. Res.*, vol. 11, pp. 3595–3646, 2010.

[10] J. M. Buhmann, “Information theoretic model validation for clustering,” in *Proc. of IEEE Int. Symp. on Information Theory 2010*, 2010, pp. 1398–1402.

[11] D. Pál S. Ben-David, U. von Luxburg, “A sober look at clustering stability,” in *Springer Verlag LNAI 4005 Proc. of COLT*, 2006, pp. 5–19.

[12] M. Haghir Chehreghani, A.G. Busetto, and J.M. Buhmann, “Information theoretic model validation for spectral clustering,” in *J. of Mach. Learn. Res. Proc. of AISTATS 2012*, 2012, pp. 495–503.

[13] T. Hofmann and J.M. Buhmann, “Pairwise data clustering by deterministic annealing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19(1), pp. 1–14, 1997.

[14] S. Chawla N. Bansal, A. Blum, “Correlation clustering,” *Machine Learning*, vol. 56, pp. 89–113, 2004.

[15] F. Mignone H. Boussetta A. Viarengo F. Dondero M. Banni, A. Negri, “Gene expression rhythms in the mussel Mytilus galloprovincialis (lam.) across an annual cycle.,” *PLoS ONE*, vol. 6(5), pp. e18904, 2011.