Information theoretic model selection in clustering

Joachim M. Buhmann
Department of Computer Science
ETH Zurich, 8092 Zurich
jbuhmann@inf.ethz.ch

Abstract

Model selection in clustering requires (i) to specify a clustering principle and (ii) to decide an appropriate number of clusters depending on the noise level in the data. We advocate an information theoretic perspective where the uncertainty in the data set induces an uncertainty in the solution space of clusterings. A clustering model, which can tolerate a higher level of noise in the data than competing models, is considered to be superior provided that the clustering solution is equally informative. This tradeoff between informativeness and robustness is used as a model selection criterion. The request that solutions should generalize from one data set to an equally probable second data set gives rise to a new notion of structure induced information.

1 Clustering: Science or Art ?

Data clustering or data partitioning has emerged as the workhorse of *exploratory data analysis*. This unsupervised learning methodology comprises a set of data analysis techniques which group data into clusters by either optimizing a quality criterion or by directly employing a clustering algorithm. The zoo of models range from centroid based algorithms like k-means or k-medoids, spectral graph methods like Normalized Cut, Average Cut or Pairwise Clustering to linkage inspired grouping principles like Average Linkage or Path-based Clustering.

In this talk I will argue for a shift of viewpoint away from the problem *"What is the 'right' clustering model?"* to the question *"How can clustering models algorithmically be validated?"*. This conceptual shift roots in the assumption that ultimately, the data should vote for their prefered model type and model complexity[3]. Therefore, algorithms which are endowed with the ability to validate clustering concepts can maneuver through the space of clustering models and, dependent on the training and validation data sets, they can select a model with maximal information content and optimal robustness. While the design of clustering models based on prior knowledge of the data source might be considered as *Art*, the systematic search through the space of clustering models by cluster validation based on information theoretic principles defines an algorithmic strategy of a scientific program.

Information theoretic model validation uses empirical risk approximation (ERA) [2] to quantize the hypothesis class of clusterings. ERA employs an hypothetical communication framework where sets of approximate clustering solutions for the training and for the test data are used as a communication code. Approximations of the empirical minimizer with model averaging favors stability of clusterings. Information theoretic model validation is formulated in the context of risk approximation, although it only requires a clustering method which can output a set of clusterings without necessarily minimizing a risk function. Furthermore, it is well known that stability based model selection [4] yields highly satisfactory results in applications although the theoretical foundation of this model selection strategy is still controversely debated [1]. Cluster inference based on approximations is motivated by the uncertainty in data which induce uncertainty in the solution space. Clusterings are considered to be similar if they are statistically indistinguishable due to data noise. The request that

solutions should generalize from one data set to an equally probable second data set gives rise to a new notion of structure induced information.

2 Statistical learning of clustering

Given are a **set of objects** $\mathcal{O} = \{o_1, \ldots, o_n\}$ and measurements $\mathbf{X} \in \mathcal{X}$ to characterize these objects. $\mathcal{X}$ denotes the measurement space. Such measurements might be vectors $\mathbf{x}_i \in \mathbb{R}^d, 1 \leq i \leq n$ in a d-dimensional space or relations $\mathbf{D} = (D_{ij}) \in \mathbf{R}^{n \cdot n}$ which describe the (dis)-similarity between object o_i and o_j. More complicated data structures than vectors or relations, e.g., three-way data or graphs, are used in various applications. In the following, we use the generic notation $\mathbf{X}$ for measurements. Data denote the relation $\mathcal{O} \times \mathcal{X}$ of object-measurement relations.

The **hypothesis class** for a clustering problem is defined by the set of assignments of data to groups, i.e., $\mathcal{C} = \{c : \mathcal{O} \times \mathcal{X} \to \{1, \ldots, k\}\}$. $\mathcal{C}(\mathbf{X})$ is a set of functions which map objects to cluster indices. For n objects we can distinguish $O(k^n)$ such functions. Special clustering models might require additional parameters θ which characterize a cluster like the centroids in k-means clustering. The hypothesis class is then the product space of possible assignments and possible parameter values.

Pattern analysis in data clustering requires to quantify the quality of such hypotheses, e.g., in vector quantization we use the k-means cost function and we use the nearest centroid assignment rule. For the subsequent discussion on empirical risk approximation we assume that a cost function $R(c, \theta; \mathbf{X})$ is given which measures how well a particular clustering with assignments $c(o, \mathbf{X})$ and cluster parameters θ groups the objects. A suitable metric on the space of hypotheses might be chosen based on such a cost function R.

3 Why information theory for clustering?

To formulate the statistical learning question we have to consider the following problem: Quite often the measurement space $\mathcal{X}$ has a much higher "dimension" than the solution space. Consider for example the problem of spectral clustering with k groups based on dissimilarities $\mathbf{D}$: The measurements are elements of $\mathbb{R}^{n(n-1)/2}$ for real valued, symmetric weights with vanishing self-dissimilarity, but we can only distinguish less than or equal to k^n different clusterings. Any approach which relies on estimating the probability distribution of the data ultimately will fail since we require far too many observations than needed to identify one hypothesis or a set of hypotheses, i.e., one clustering or a set of clusterings.

Using an information theoretic perspective, we might ask the question how the uncertainty in the observations limit the resolution in the hypothesis class. How different can two hypotheses be so that they are still statistically indistinguishable given a cost function $R(c, \theta; \mathbf{X})$? The core question of statistical learning, "How well does a learning solution generalize?", is intimately related to the problem of distinguishing hypotheses.

Shannon's information theory provides a framework to study such questions of how many bit strings can be reliably distinguished in the presence of noise and, therefore, can be used as a code for communication. This study is based on the idea that approximation sets of clustering cost functions can be used as a reliable code. The capacity of such a coding scheme then answers the question how sensitive a particular cost function is to noise. "Good" models exhibit high robustness to noise and at the same time, they are highly informative due to a large hypothesis class. "Poor" models might be sensitive to noise (overfitting) or might be very restrictive with a small hypothesis class (underfitting).

To identify the correlates of *code vector* and *code book* in classical information theory, we define the set $\mathcal{C}_\gamma$ of hypotheses which are γ-optimal w.r.t. the minimum cost solution $c^\perp(\mathbf{X}) = \arg\min_{c,\theta} R(c, \theta; \mathbf{X})$, i.e.,

$$\mathcal{C}_\gamma(\mathbf{X}) = \{c \ : \ R(c, \theta; \mathbf{X}) \leq R(c^\perp, \theta^\perp; \mathbf{X}) + \gamma\}. \tag{1}$$

To test how well a γ-optimal solution generalizes to a new data set, we assume two samples of a problem to be given, i.e., we have measurements $\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \sim \Pr(\mathbf{X})$ for training and testing.

These two measurements $X^{(1,2)}$ define two optimization problems $R(c, \theta; \mathbf{X}^{(\mathbf{1},\mathbf{2})})$. For both measurements we can determine γ-optimal approximations $\mathcal{C}_\gamma^{(1,2)} := \mathcal{C}_\gamma(\mathbf{X}^{(1,2)})$. To measure stability of a solution, we require that the intersection between both sets is as large as its intersection with the complement $\overline{\mathcal{C}}_\gamma(\mathbf{X}^{(1)})$. If this condition is met then the noise in the data will not affect the property to be γ-optimal to the optimum. γ-optimality can be considered as a similarity criterion based on the cost function $R(c, \theta; \mathbf{X})$.

4 Coding by approximation

The informativeness-robustness tradeoff is expressed by the condition that the approximation precision γ should be as small as possible and the intersection between the two approximation sets should be as large as their union. This condition corresponds to Shannon's random coding argument that the received bit string should be jointly typical with the codeword which has been selected by the sender. The error of this communication process vanishes for asymptotically large bit strings provided we do not exceed the capacity of the communication channel.

In our setting, where we use approximation sets for coding, we have to generate $2^{n\rho}$ different code problems with respective approximation sets so that a zero error condition can be used to determine the optimal model. $n\rho$ defines a coding rate which should be maximized. Furthermore, such a procedure will allow us to measure the number of bits relative to the hypothesis class which we have selected for our clustering problem.

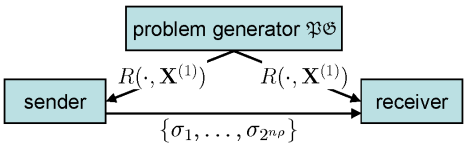

Figure 1: Generation of a set of $2^{n\rho}$ code problems for communication by e.g. permuting the object indices.

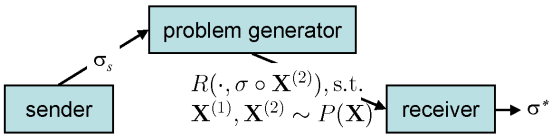

Figure 2: Communication process: the sender selects transformation σ_s and the receiver estimates $\sigma^\star$.

As depicted in fig. 1, the sender follows the following procedure to define the set of code problems: (i) the problem generator send the data set $\mathbf{X}^{(1)}$ to sender and receiver; (ii) the sender permutes the object indices of the data in such a way that the new optimal solution $c^\perp(o, \mathbf{X})$ is transformed to $\sigma_j \circ c^\perp(o, \mathbf{X})$. In total, there exist $2^{nH(p_\alpha)}$ with $H(p_\alpha) = -\sum_{1 \leq \nu \leq n} np_\nu \log p_\nu$ many transformations σ. This set of transformations is shared with the receiver which establishes a code.

In the **communication process** (fig. 2), the sender selects the transformation σ_s and send this transformation to the problem generator $\mathfrak{PG}$ which generates a second data set $\mathbf{X}^{(2)} \sim \Pr(\mathbf{X})$. This *test* data set is drawn from the same probability distribution as the training data set $\mathbf{X}^{(1)}$. The $\mathfrak{PG}$ then applies the transformation σ_s to the test data and send the transformed data $\tilde{\mathbf{X}} = \sigma_s \circ \mathbf{X}^{(2)}$ to the receiver.

The receiver now faces the question which transformation $\sigma_j,\ 1 \leq j \leq 2^{n\rho}$ has been selected by the sender. If he is able to estimate the transformation selected by the sender, then he has received $n\rho$ bits in this communication. To compute the intersection between the approximation set of the test problem and the approximation set of one of the code problems, solutions of the test problem have to be mapped to the hypothesis class $\mathcal{C}(\mathbf{X}^{(1)})$. This mapping is denoted by $\phi\ :\ \mathcal{C}(\mathbf{X}^{(2)}) \to \mathcal{C}(\mathbf{X}^{(1)})$.

For **decoding**, the receiver intersects the approximation set $\phi \circ \mathcal{C}_\gamma(\tilde{\mathbf{X}})$ with all approximation sets in the codebook $\{\mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}), 1 \leq j \leq 2^{n\rho}\}$. Under the condition that a sufficiently large overlap exists, the receiver declares the transformation $\sigma^\star$

$$\sigma^\star = \arg\max_{\sigma} \left| \mathcal{C}_\gamma(\sigma \circ \mathbf{X}^{(1)}) \cap \phi\left(\mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})\right) \right| \quad (2)$$
$$\text{if} \quad \frac{|\mathcal{C}_\gamma(\sigma^* \circ \mathbf{X}^{(1)}) \cap \phi\left(\mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})\right)|}{|\mathcal{C}_\gamma(\sigma^* \circ \mathbf{X}^{(1)})|} \geq 1 - \epsilon$$

as being selected by the sender. $\sigma^\star$ is the received message which has been transmitted by an approximate optimization protocol using the problem generator as a channel.

5 Error Analysis of this Code

To analyze the error of this communication protocol, we introduce the following events

$$\mathcal{E}_j = \mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}) \cap \left(\phi \circ \mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})\right). \quad (3)$$

The event $\forall j \neq s,\ \mathcal{E}_s > \mathcal{E}_j$ corresponds to correct communication with $\sigma^\star = \sigma_s$.

Two types of errors can occur using this communication protocol:

1. The approximation set $\phi \circ \mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})$ does not substantially intersect with the "correct", sender selected approximation set $\mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(1)})$, i.e.,
$$\overline{\mathcal{E}_s} := \mathcal{C}(\mathbf{X}^{(1)}) \setminus \mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(1)}) \cap \left(\phi \circ \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(2)})\right) \quad (4)$$
2. The approximation set $\phi \circ \mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})$ substantially intersects with an "incorrect" approximation set $\mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}),\ j \neq s$, i.e., the event $\mathcal{E}_j,\ j \neq s$ occurs.

The conditional error of communication

$$P(\text{error}|\sigma_s) = P\left(\bigvee_{j=1}^{2^{n\rho}} \mathcal{E}_j > \mathcal{E}_s \middle| \sigma_s\right) \quad (5)$$

will determine the capacity of our communication channel, which we name approximation capacity. The communication rate $n\rho$ should not exceed the mutual information

$$\mathcal{I}(\mathcal{C}_\gamma(X^{(1)}), \phi \circ \mathcal{C}_\gamma(X^{(2)})). \quad (6)$$

Different clustering models $R(., X)$ can be ranked according to their approximation capacity where good model will demonstrate a high approximation capacity. They are robust to the noise in the data and, therefore, allow us to select a small γ with a correspondingly large $n\rho$.

Acknowledgement: This work has been partially supported by the DFG-SNF research cluster FOR916 and by the FP7 EU project SIMBAD.

References

[1] Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In G. Lugosi and H.U. Simon, editors, *Learning Theory, Proceedings of 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA*, volume 4005 of *LNAI*, pages 5–19. Springer-Verlag Berlin Heidelberg, 2006.

[2] Joachim M. Buhmann. Empirical risk approximation: An induction principle for unsupervised learning. Technical Report IAI-TR-98-3, Department of Computer Science III / University of Bonn, 1998.

[3] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008.

[4] Tilman Lange, Mikio Braun, Volker Roth, and Joachim M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323, June 2004.