Prediction and information: inference of nonlinear dependence using supervised learning

Charles Zheng and Yuval Benjamini

November 6, 2016

Abstract

Neuroscientists have a variety of tools for quantifying multivariate dependence: mutual information, linear correlation-based statistics, Fisher information, and more recently, measures of performance on supervised learning tasks such as classification. We argue that both mutual information and classification accuracy capture intuitive properties of an "information metric" for a channel, and we proceed to develop a general axiomatic characterization of information metrics for channels consisting of a pair of input and output random variables. The key axioms of an information metric are that (i) it is a scalar functional of the joint distribution of the input-output pair, (ii) it is zero for independent variables, and positive for dependent variables, (iii) it is monotonic with respect to inclusion of additional output variables, and (iv) it is invariant to bijective transformations of either the input or output. We show how prediction tasks can be used to define a general class of information metrics which includes mutual information, as well as a novel information metric, *average Bayes accuracy*, which can be considered an "idealization" of classification accuracy. Furthermore, we consider the possibility of developing a general theory of statistical inference for this class of information metrics. Concretely, we derive a lower confidence bound for average Bayes accuracy as well as a novel lower confidence bound for mutual information.

1 Introduction

Historically, neuroscience has largely taken a reductionist approach to understanding the nervous system, proceeding by defining elements and subelements of the nervous system (e.g. neurons), and investigating relationship between two different elements, or the response of an element to external stimulation: say, the response of a neuron's average firing rate to skin temperature. At one level of abstraction, neuroscientists might seek to characterize the functional relationship between elements, but at a higher level of abstraction, it may be sufficient to report scalar measures of dependence. Since neural dynamics are generally both stochastic and nonlinear, it was a natural choice for early neuroscientists to adopt Shannon's *mutual information* as a quantitative measure of dependence. But as new technologies enabled the recording of neural data at larger scales and resolution, the traditional reductionist goals of neuroscience were supplemented by increasingly ambitious attempts within neuroscience to understand the dynamics of neural ensembles, and by efforts originating within psychology and medicine to link the structure and function of the entire human brain to behavior or disease. The larger scope of the data and the questions being asked of the data created an increasing demand for multivariate statistical methods for analyzing neural data of increasingly high dimension. Due to the complexity, variety, and practical difficulties of multivariate statistical analysis of the brain, alternative measures of multivariate dependence such as linear-based correlational statistics, or Fisher information, started to gain traction. For the most part, alternative measures of dependence sacrifice flexibility for a gain in practical convenience: linear-based statistics such as canonical correlation or correlation coefficients fail to capture nonlinear dependencies, and Fisher information requires strong parametric assumptions. Therefore, it was of considerable interest when Haxby (2001) introduced the usage of *supervised learning* (classification tasks) for the purpose of quantifying stimulus information in task fMRI scans. Since then, an entire subfield of neuroimaging, multivariate pattern analysis (MVPA) has been established dedicated to quantifying multivariate information in the brain, and both mutual information and classification accuracy are used by practitioners within the field. Judging from the language used by the practioners themselves, it is intuitively clear to them how classification accuracies can be used to quantify information in brain scans. However, a more thorough examination of the practice raises many questions with regards to the use of classification accuracy as a metric of information: this is one motivation for the current work. But taking a step back, it would seem valuable at this historical juncture to examine the intuitive properties of "information" as a measure of multivariate dependence, and not only consider whether classification accuracy can be considered or used to derive a new information metric, but whether other such metrics might also exist, and whether a unified theory can be developed to account for all of them. This is the larger purpose of the current work, and towards that end we not only propose a general class of information metrics which unifies both information-theoretic and supervised-learning-based approaches, but with an eye toward practical applications, we also examine the question of inferring these quantities from data. An initial result in this direction is the derivation of nonparametric lower confidence bounds for average Bayes accuracy (a novel information metric closely related to classification accuracy,) and an inequality between average Bayes accuracy and mutual information, which, combined with the preceding result, yields a novel lower confidence bound for mutual information.

While Shannon's theory of information was motivated by the problem of designing communications system, the applicability of mutual information was quickly recognized by neuroscientists. Only four years after Shannon's seminal paper in information theory (1948), McKay and McCullough (1952) inaugurated the application of mutual information to neuroscience. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ have joint density $p(\boldsymbol{x}, \boldsymbol{y})$, then the mutual information is defined as

$$I(\boldsymbol{X}; \boldsymbol{Y}) = \int p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} dxdy.$$

where $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are the marginal densities. Since then, mutual information has enjoyed a celebrated position in both experimental and theoretical neuroscience. Experimentally, mutual information has been used to detect strong dependencies between stimulus features and features derived from neural recordings, which can be used to draw conclusions about the kinds of stimuli that a neural subsystem is designed to detect, or to distinguish between signal and noise in the neural output. Theoretically, the assumption that neural systems maximize mutual information between salient features of the stimulus and neural output has allowed scientists to predict neural codes from signal processing models: for instance, the center-surround structure of human retinal neurons matches theoretical constructions for the optimal filter based on correlations found in natural images [cite].

However, the dominance of mutual information in neuroscience started to wane as new experimental approaches shifted the focus from pairwise comparisons to questions of *population coding*. The extreme reductionist approach fails to account for the complex ways in which neurons cooperatively encode complex information, and to get a richer picture of neural dynamics, it becomes necessary to consider multivariate measures of dependence. Also, the adoption of "high-throughput" recording technologies such as EEG and fMRI naturally led in the direction of considering the system-level dynamics of the whole brain. More importantly, the technology for studying the relationships between brain structure, function, and behavior have enabled a wider population of investigators (psychologists and neurologists) to quantitatively examine brain activity, but with motivations which tend to be more holistic and instrumental (e.g. finding neural correlates of mental disorders) when compared to the reductionist orientation of classical neuroscience. For both reasons, the demand for multivariate statistical techniques in neuroscience has increased dramatically in recent years. While the theoretical properties of mutual information extend gracefully to the multivariate setting, the difficulty of estimating mutual information increases exponentially in the dimension in the absence of strong modeling assumptions [cite]. Partially for this reason, alternative measures of multivariate dependence started to enjoy increasing usage in studies of population coding or in systems-level investigations of the brain: these include measures of linear dependence (canonical correlation and multivariate R^2) and Fisher information. However, while both correlation-based statistics and Fisher information may be easier to estimate than mutual information in high-dimensional settings, they are both less flexible in terms of capturing nonlinear relationships. Correlation-based statistics can only capture linear dependence, and Fisher information requires the assumption of a parametric model.

Machine learning algorithms showed a way forward: a seminal work by Haxby (2001) proposed to quantify the information in multiple channels by measuring how well the stimulus can be identified from the brain responses, in what is known as "multivariate pattern analysis" (MVPA). To demonstrate that a particular brain region responds to a certain type of sensory information, one employs supervised learning to build a classifier that classifies the stimulus class from the brain activation in that region. Classifiers that achieve above-chance classification accuracy indicate that information from the stimulus is represented in the brain region. In principle, one could just as well test the statistical hypothesis that the Fisher information or mutual information between the stimulus and the activation patterns is nonzero. But in practice, the machine learning approach enjoys several advantages: First, it is invariant to the parametric representation of the stimulus space, and is opportunistic in the parameterization of the response space. This is an important quality for naturalistic stimulus-spaces, such as faces or natural images. Second, it scales better with the dimensionality of both the stimulus space and the responses space, because a slimmer discriminative model can be used rather than a fully generative model.