Prediction, information, and inference: with application to neuroimaging

Charles Zheng and Yuval Benjamini

November 7, 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 coefficient" for a channel, and we proceed to develop a general axiomatic characterization of information coefficients for channels consisting of a pair of input and output random variables. Arguably, the key properties of an information coefficient are that (i) it is a scalar measure of multivariate dependence, and (ii) that it satisfies a *stochastic data-processing inequality*: any coefficient with such properties can be used for model selection. We show how prediction tasks can be used to define a general class of information coefficients which includes mutual information, as well as a novel information coefficient, *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 coefficients. 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 coefficient 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 coefficient, but whether other such coefficients 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 coefficients 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 coefficient 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.

1.1 Organization

The rest of the paper is organized as follows. Section 2 plays the role of a "background" section that gives the basics of mutual information and supervised learning as they are used in neuroscience, as well as practical issues related to estimation. In section 3, we present an axiomatic characterization of information coefficients, and introduce a general class of information coefficients which satisfies our axioms. We define a new information coefficient belonging to this class, average Bayes accuracy, and we also show how mutual information can be considered an "extended member" of the class. In section 4 we develop the basic theory of what kinds of inferences about our information coefficients are possible discuss the kinds of experimental designs and supervised learning pipelines which are needed to enable such inference. Concretely, we develop a lower confidence bound for average Bayes accuracy. In section 5 we outline a comparative theory for different coefficients within our framework: how are the different information coefficients related? We discuss the calculus of variations as a possible general technique for establishing inequalities between different information coefficients, and in particular we derive a "randomized Fano's inequality": a lower bound for mutual information as a function of average Bayes accuracy. Combined with our lower confidence bound for average Bayes accuracy, this yields a novel lower confidence bound for mutual information. We provide a practical data analysis example in section 6. A discussion section includes future directions and loose ends are treated, and most of the technical proofs and lemmas are found in the appendix.

2 Background

2.1 Mutual information and its usage

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 $p(\boldsymbol{x}, \boldsymbol{y})$ with respect to the product measure $\mu_x \times \mu_y$, then the mutual information is defined as

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

where $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are the marginal densities with respect to μ_x and μ_y[1]. 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].

The mutual information measures the information "capacity" of a channel consisting of an input $\boldsymbol{X}$ and an output $\boldsymbol{Y}$, and satisfies a number of important properties.

1. The channel input $\boldsymbol{X}$ and output $\boldsymbol{Y}$ can be random vectors of arbitrary dimension, and the mutual information remains a scalar functional of the joint distribution P of $(\boldsymbol{X}, \boldsymbol{Y})$.

[1] Note that the mutual information is invariant with respect to change-of-measure.

2. When $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = 0$; otherwise, $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) > 0$.

3. The data-processing inequality: for any vector-valued function $\vec{f}$ of the output space,
$$\mathrm{I}(\boldsymbol{X};\vec{f}(\boldsymbol{Y})) \leq \mathrm{I}(\boldsymbol{X};\boldsymbol{Y}).$$

4. Symmetry: $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{I}(\boldsymbol{Y};\boldsymbol{X})$.

5. Independent additivity: if $(\boldsymbol{X}_1,\boldsymbol{Y}_1)$ is independent of $(\boldsymbol{X}_2,\boldsymbol{Y}_2)$, then
$$\mathrm{I}((\boldsymbol{X}_1,\boldsymbol{Y}_1);(\boldsymbol{X}_2,\boldsymbol{Y}_2)) = \mathrm{I}(\boldsymbol{X}_1;\boldsymbol{Y}_1) + \mathrm{I}(\boldsymbol{X}_2;\boldsymbol{Y}_2).$$

Three additional consequences result from the data-processing inequality:

- *Stochastic data-processing inequality* If $\vec{f}$ is a stochastic function independent of both $\boldsymbol{X}$ and $\boldsymbol{Y}$, then
$$\mathrm{I}(\boldsymbol{X};\vec{f}(\boldsymbol{Y})) \leq \mathrm{I}(\boldsymbol{X};\boldsymbol{Y}).$$

  This can be shown as follows: any stochastic function $\vec{f}(\boldsymbol{Y})$ can be expressed as a deterministic function $\vec{g}(\boldsymbol{Y},W)$, where W is a random variable independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$. By independent additivity,
$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{I}(\boldsymbol{X};(\boldsymbol{Y},W)).$$

  Then, by the data-processing inequality,
$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{I}(\boldsymbol{X};(\boldsymbol{Y},W)) \geq \mathrm{I}(\boldsymbol{X};\vec{g}(\boldsymbol{Y},W)) = \mathrm{I}(\boldsymbol{X};\vec{f}(\boldsymbol{Y})).$$

- *Invariance under bijections.* If $\vec{f}$ has an inverse $\vec{f}^{-1}$, then
$$\mathrm{I}(\boldsymbol{X};\vec{f}(\boldsymbol{Y})) \leq \mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{I}(\boldsymbol{X};\vec{f}^{-1}(\vec{f}(\boldsymbol{Y}))) \leq \mathrm{I}(\boldsymbol{X};\vec{f}(\boldsymbol{Y})),$$

  therefore, $\mathrm{I}(\boldsymbol{X};\vec{f}(\boldsymbol{Y})) = \mathrm{I}(\boldsymbol{X};\boldsymbol{Y})$.

- *Monotonicity with respect to inclusion of outputs.* Suppose we have an output ensemble $(\boldsymbol{Y}_1,\boldsymbol{Y}_2)$. Then the individual component $\boldsymbol{Y}_1$ can be obtained as a projection of the ensemble. By the data-processing inequality, we therefore have
$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}_1) \leq \mathrm{I}(\boldsymbol{X};(\boldsymbol{Y}_1,\boldsymbol{Y}_2)).$$

  Intuitively, if we observe both $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, this can only *increase* the information we have about $\boldsymbol{X}$ compared to the case where we only observe $\boldsymbol{Y}_1$ by itself.

And it is the property of *invariance under bijections*, inclusive of non-linear bijections, which qualifies mutual information as a *non-linear measure of dependence.* Linear measures, such as Pearson correlation, are not invariant under bijections.

Besides the formal definition, there are a number of well-known alternative characterizations of mutual information in terms of other information-theoretic quantities: the *entropy* H:

$$\mathrm{H}_{\mu}(\boldsymbol{X}) = -\int p(\boldsymbol{X}) \log p(\boldsymbol{X}) d\mu(\boldsymbol{X}),$$

and the *conditional entropy*:

$$\mathrm{H}_{\mu}(\boldsymbol{X}|\boldsymbol{Y}) = -\int p(\boldsymbol{Y}) d\mu_y(\boldsymbol{Y}) \int p(\boldsymbol{X}|\boldsymbol{Y}) \log p(\boldsymbol{X}|\boldsymbol{Y}) d\mu_x(\boldsymbol{X}).$$

Some care needs to be taken with entropy and conditional entropy since they are not invariant with respect to change-of-measure: hence the use of the subscript in the notation H_{μ}. In particular, there is a difference between *discrete entropy* (for counting measure) and *differential entropy* (for Lesbegue measure.) Intutively, entropy measures an observer's uncertainty of the random variable $\boldsymbol{X}$, supposing the observer has no prior information other than the distribution of $\boldsymbol{X}$. Conditional entropy measures the *expected uncertainty* of $\boldsymbol{X}$ supposing the observer observes $\boldsymbol{Y}$.

However, regardless of the base measure, the following identities hold:

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{H}_{\mu_x \times \mu_y}((\boldsymbol{X},\boldsymbol{Y})) - \mathrm{H}_{\mu_x}(\boldsymbol{X}) - \mathrm{H}_{\mu_y}(\boldsymbol{Y}).$$

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{H}_{\mu}(\boldsymbol{Y}) - \mathrm{H}_{\mu}(\boldsymbol{Y}|\boldsymbol{X}). \tag{1}$$

The second identity (1) is noteworthy as being practically important for estimation of mutual information. Since the entropies in question only depend on the marginal and conditional distributions of $\boldsymbol{Y}$, the problem of estimating $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y})$ can be reduced from a $\dim(\boldsymbol{X}) + \dim(\boldsymbol{Y})$-dimensional nonparametric estimation problem to a $\dim(\boldsymbol{Y})$-dimensional problem: hence this identity is a basis of several methods of estimation used in neuroscience, such as Gastpar (2014).

However, by symmetry, we also have the flipped identity

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{H}_{\mu}(\boldsymbol{X}) - \mathrm{H}_{\mu}(\boldsymbol{X}|\boldsymbol{Y}). \tag{2}$$

In neuroscience studies, where $\boldsymbol{X}$ is the controlled stimulus, and $\boldsymbol{Y}$ is the neural activity, the two mirror pairs (1) and (2) have different interpretations. Rather than providing a basis for practical estimation, (2) provides an *interpretation* of the mutual information. Loosely speaking, $\mathrm{H}_\mu(\boldsymbol{X})$ is the uncertainty of $\boldsymbol{X}$ before having observed $\boldsymbol{Y}$, and $\mathrm{H}_\mu(\boldsymbol{X}|\boldsymbol{Y})$ is the uncertainty of $\boldsymbol{X}$ after having observed $\boldsymbol{Y}$, hence $\mathrm{H}_\mu(\boldsymbol{X}) - \mathrm{H}_\mu(\boldsymbol{X}|\boldsymbol{Y})$ is how much the observation of $\boldsymbol{Y}$ has *reduced* the uncertainty of $\boldsymbol{X}$. Stated in words,

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \text{average reduction of uncertainty about } \boldsymbol{X} \text{ upon observing } \boldsymbol{Y}.$$

We list these properties of mutual information in preparation for section 3.1, where we prepare a "minimal" set of properties for an information coefficient, and consider how much of the functionality of the mutual information would be preserved by an alternative information coefficient satisfying only those minimal properties.

But what, exactly, is the functionality of mutual information in neuroscience? How it is used in practice? A nice summary of the applications of mutual information is provided in the introduction of Gastpar (2014). Taking their list as a starting point, we briefly overview the main use-cases of mutual information, and illustrate each with a representative example.

- Example: Comparison of decoders in Nelken. Property (i) is important to enable model comparison. Property (iii) is needed because relationships may be nonlinear.
- Example: Redundancy in population code of retina. Property (i)-(iii) and (v) are needed to obtain a meaningful measure of redundancy.
- In general, symmetry not important, but additivity is desirable for measures of redundancy. Property (ii) can usually be enforced since any measure needs to have a unique "minimum" value for the case of independence.

2.2 Supervised learning

- Supervised learning task is defined using a prediction task.
- 1. A predictive model is learned using training data

- 2. The performance of the model on the prediction task is estimated using independent test data
- Classical examples of prediction tasks: regression and classification
- Third example: identification
- Definition of Bayes prediction model
- General definition of supervised learning task
- SL performance can be a scalar
- SL can be used to test for independence
- Bayes performance satisfies data-processing inequality
- How SL is interpreted in MVPA as information
- Section 3, we'll see how Bayes performance can be legitimately considered a measure of information

2.3 Connections

- Information theory and decoding. Fano's inequality
- Quiroga's method
- MVPA people use them interchangeably

3 Information, uncertainty and prediction

In the previous section, we examined how mutual information and supervised learning are used in neuroscience. By analyzing the use-cases of each method side-by-side, and by reviewing the known connections between the two methods, we hoped to suggest the notion that both mutual information and supervised learning are being employed as means of quantifying the same underlying concept. In this section, we propose a reification of this underlying concept of "information", and propose the *minimal* properties needed for a functional to be considered an *information coefficient.* We argue that these minimal properties are sufficient to support many of the existing use-cases of mutual information and supervised learning in neuroscience.

3.1 Axiomatic characterization of information

We claim that neuroscientists use both mutual information and supervised learning to quantifying a common concept of "information." Furthermore, we claim that neuroscientists largely share a set of common intuitions about information.

- *Intuition 1: Information is a measure of dependence.* If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are statistically independent, then $\boldsymbol{X}$ gives no information about $\boldsymbol{Y}$, and vice-versa.

This intuition is employed when researchers test the null hypothesis of chance accuracy for classification. If the null is accepted, the researcher concludes that there is no information in the predictors about the response.

- *Intuition 2a: Monotonicity with respect to inclusion of outputs.* If $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are ensembles of neurons (or individual neurons), then the combined ensemble $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ has equal or more information about $\boldsymbol{X}$ than either component by itself.

- *Intuition 2b: Noise adds no information.* If, in the previous example, $\boldsymbol{Y}_2$ is independent of $\boldsymbol{X}$, then $\boldsymbol{Y}_2$ adds no information to the ensemble.

The non-informativity of noise is vital for the purpose of *localizing* information within fMRI voxels (as is done in searchlight analysis.) Since noise voxels fail to improve classification performance (and indeed, sometimes harm empirical performance,) the optimal searchlight radius will concentrate on clusters of signal voxels, and minimize the inclusion of noise voxels.

- *Intuition 3:* Information can be used as a basis of model selection. Among multiple encoding/decoding models, a more accurate model should tend to have greater information relative to less accurate models.

Compared to the first two, this third intuition is somewhat less obvious, but nevertheless appears as an important use-case for mutual information, as seen in the application of mutual information to choose encoding models.

Given the first two intuitions, we find that an essential property of information is that it can be said to 'increase'–i.e., there exists at least a partial ordering on information. Furthermore, there should exist a minimal element in this ordering, which is the information between independent variables–that is, 'no information.' However, this is fully consonant with either information being quantified as a scalar quantity, or as a positive-definite matrix. Indeed, Fisher information could be taken as an example of a matrix-valued information coefficient. However, the problem with matrix-valued coefficients is that channels may be incomparable within the partial ordering. Thus, if we consider the third intuition an important property of an information coefficient, then it should be scalar to enable model comparison.

We find that all of the preceding intuitions follow from the following axioms.

Axioms of information

Let $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y})$ denote an *information coefficient.* Then,

1. $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y})$ is a scalar functional of the joint distribution P of $(\boldsymbol{X},\boldsymbol{Y})$.

2. When $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) = 0$; otherwise, $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) \geq 0$.

3. The stochastic data-processing inequality. If $\vec{f}$ is stochastic vector-valued function of the output space indepedent of $(\boldsymbol{X},\boldsymbol{Y})$, then

$$\mathcal{I}(\boldsymbol{X};\vec{f}(\boldsymbol{Y})) \leq \mathcal{I}(\boldsymbol{X};\boldsymbol{Y}).$$

Since this minimal list of properties is a subset of the properties of mutual information, it is clear that mutual information satisfies the axioms. However, classification accuracy does not satisfy the axioms–we will see in section 3.3 how to define a proper information coefficient based on supervised learning.

Now let us check that any information coefficient $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y})$ necessarily satisfies the intuitions. Intuition 1 is satisfied by property (ii). We gave a general argument in 2.1 how Intuition 2a follows from the data-processing inequality. Intuition 2b follows from the stochastic data-processing inequality, taking $\vec{f}(\boldsymbol{Y}_1) = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$.

Now, in order to justify intuition 3, we need to formalize the notion of "model selection." A complete discussion of model selection falls outside the scope of the paper, so we limit the discussions to the important special case of selecting an encoding model.

Let $\boldsymbol{X}$ be some environmental stimulus, and

$$\boldsymbol{Y} = \vec{f}(g(\boldsymbol{X}))$$

where $\vec{f}$ is a stochastic function independent of $\boldsymbol{X}$, and g is an *encoding function*, which is known to lie in some class of functions $\mathcal{G}$. The goal of model selection is estimate g.

Given multiple competing models $(\hat{S}_1, \hat{g}_1), \ldots, (\hat{S}_k, \hat{g}_k)$, we obtain a lower confidence bound on the maximum score,

$$M = \max_{i=1}^{k} \mathcal{I}(\boldsymbol{Y}; \hat{g}_i(\boldsymbol{X})).$$

We can then test each individual model for the hypothesis $H_i : \mathcal{I}(\boldsymbol{Y}; \hat{g}_i(\boldsymbol{X})) < M$ at level α/k. All of the models which are rejected are considered 'candidate models.' While a common next step is to select a single model from the set of candidates (e.g. using a measure of complexity), this is not essential to our discussion. For now, we are only concerned that the model selection procedure should at the very least, *not reject* the *correct* model g in the circumstance that g is included in the initial set of candidate models. A necessary condition for this criterion is that the correct model g maximizes the information:

$$\mathcal{I}(\boldsymbol{Y}; g(\boldsymbol{X})) = \sup_{g \in \mathcal{G}} \mathcal{I}(\boldsymbol{Y}; g(\boldsymbol{X})). \tag{3}$$

Indeed, the criterion (3) is a consequence of the stochastic data-processing inequality. First observe that letting $V = g(\boldsymbol{X})$, we have $\boldsymbol{X}$ and $\boldsymbol{Y}$ conditionally independent given V. Therefore, we can write

$$\boldsymbol{X} = \vec{h}(V)$$

where $\vec{h}$ is a stochastic function independent of V and $\boldsymbol{Y}$. It follows from the stochastic data-processing inequality that

$$\mathcal{I}(\boldsymbol{Y}; \boldsymbol{X}) = \mathcal{I}(\boldsymbol{Y}; \vec{h}(V)) = \mathcal{I}(\boldsymbol{Y}; V) = \mathcal{I}(\boldsymbol{Y}; g(\boldsymbol{X})).$$

Now consider an alternative encoding model $\hat{g}$. From the stochastic data-processing inequality,

$$\mathcal{I}(\boldsymbol{Y}; \boldsymbol{X}) \geq \mathcal{I}(\boldsymbol{Y}; \hat{g}(\boldsymbol{X})).$$

Therefore,

$$\mathcal{I}(\boldsymbol{Y}; g(\boldsymbol{X})) \geq \mathcal{I}(\boldsymbol{Y}; \hat{g}(\boldsymbol{X}))$$

as needed.

3.2 General characterization of supervised learning

As the central theme of the paper is the link between supervised learning and information, we now give an generalized definition of supervised learning to complement our axiomatic characterization of information, in preparation for the synthesis of the two in section 3.3.

A *supervised learning problem* is given by a *prediction task*, and a *sampling scheme* for training and test data.

Let us first define the notion of a sampling scheme. Given a joint density $p(\boldsymbol{x}, \boldsymbol{y})$ with respect to $\mu_x \times \mu_y$, define the marginal densities $p(\boldsymbol{x})$, $p(\boldsymbol{y})$ and conditional densities $p(\boldsymbol{y}|\boldsymbol{x})$, $p(\boldsymbol{x}|\boldsymbol{y})$. A *sample* is defined as a vector of observations $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_X}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_Y})$. The joint distribution of the sample is given by a colored graph G, called the *sampling scheme.* The graph satisfies the following properties:

- G is bipartite with vertex sets V_X and V_Y, directed, and acyclic.

- V_X has n_X vertices, labeled $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_X}$. All vertices in V_X are colored red.

- V_Y has n_Y vertices, labeled $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_Y}$. All vertices in V_X are colored blue.

Let $R_X = \{i\}$ be the set of indices corresponding to vertices in V_X without parents, and define R_Y analagously. Let $E_{XY} = \{(i,j)\}$ denote the set of directed edges from V_X to V_Y, and define E_{YX} analogously. Then, the distribution of the sample is given by the density

$$p_{samp}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_X}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_Y}) = \prod_{i \in V_X} p(\boldsymbol{x}_i) \prod_{j \in V_Y} p(\boldsymbol{y}_j) \prod_{(a,b) \in E_{XY}} p(\boldsymbol{y}_i|\boldsymbol{x}_i) \prod_{(c,d) \in E_{YX}} p(\boldsymbol{x}_i|\boldsymbol{y}_i).$$

with respect to the product measure $\mu_x^{n_X} \times \mu_y^{n_Y}$. And in practice, one can sample from p_{samp} as follows:

1. For all $i \in R_X$, sample $\boldsymbol{x}_i$ from $p(\boldsymbol{x}_i) d\mu_x(\boldsymbol{x}_i)$.

2. For all $i \in R_Y$, sample $\boldsymbol{y}_i$ from $p(\boldsymbol{y}_i) d\mu_y(\boldsymbol{y}_i)$.

3. Iterate the following until all components have been sampled:

    - Define S_X (S_Y) to be the set of all vertices in V_X (V_Y) that have already been sampled.

- For all edges (i, j) in E_{XY} such that $i \in S_X$ and $j \notin S_Y$, sample $\boldsymbol{y}_j$ from $p(\boldsymbol{y}_j|\boldsymbol{x}_i)d\mu_x(\boldsymbol{x}_i)$.
- For all edges (i, j) in E_{YX} such that $i \in S_Y$ and $j \notin S_X$, sample $\boldsymbol{x}_j$ from $p(\boldsymbol{x}_j|\boldsymbol{y}_i)d\mu_y(\boldsymbol{y}_i)$.

Two common examples of sampling schemes are as follows.

- *Pair-sampling.* The sample consists of i.i.d. pairs $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ drawn from the joint distribution. The sampling scheme G is a graph where the only edges are from $\boldsymbol{X}_i$ to $\boldsymbol{Y}_i$ for $i = 1, \ldots, n$.

- *Repeated measures.* From each $\boldsymbol{X}_i$, one draws r conditionally independent responses $\boldsymbol{Y}_i^1, \ldots, \boldsymbol{Y}_i^r$. The sampling scheme G is a graph where the only edges are from $\boldsymbol{X}_i$ to $\boldsymbol{Y}_i^j$ for $i = 1, \ldots, n$, $j = 1, \ldots, r$.

The *prediction task* involves a *prediction sampling scheme*, given by a graph G_{task}, a pair of functions $\vec{u}, \vec{v}$, and a *cost function* C, and an *auxillary* random variable W which is independent of the sample. Let p_{task} denote the density specified by G_{task}, and let S_{task} denote the support of the p_{task}. The *predictors* of the prediction task, $\boldsymbol{U}$ is given by $\vec{u}$, a function of components of $\boldsymbol{S}_{task}$ and possibly of W, and the *prediction target* $\boldsymbol{V}$, is given by $\vec{v}$, which is a function of components of $\boldsymbol{S}_{task}$ and possibly of W. The cost function $C(\boldsymbol{v}, \hat{\boldsymbol{v}})$ specifies the *cost* incurred when the true target is $\boldsymbol{v}$ and the prediction is $\hat{\boldsymbol{v}}$.

A *predictive model* is a function which predicts the target $\boldsymbol{V}$ as a function of the predictors $\boldsymbol{U}$. The prediction task and joint density $p(\boldsymbol{x}, \boldsymbol{y})$ jointly define a *risk function* for any predictive model $\vec{f}$:

$$\text{Risk}(\vec{f}) = \mathbf{E}[C(\boldsymbol{V}, \hat{f}(\boldsymbol{V}))].$$

The *Bayes risk* is the infimal risk over all predictive models:

$$\text{R}_{Bayes}[p(\boldsymbol{x}, \boldsymbol{y})] = \inf_{\vec{f}} \text{Risk}(\vec{f}).$$

In anticipation of the next section, we write R_{Bayes} as a functional of the joint density.

The training data is given by a sampling scheme G_{train}, and the testing data is given by a sampling scheme G_{test}. A *learning algorithm* outputs a predictive model $\vec{f}$ given the training sample $\boldsymbol{S}_{train}$ as input.

The analysis culminates in an unbiased estimate of the risk of the learned model, $\text{Risk}(\vec{f})$. Define a *rooted subgraph* H of a graph G as a subgraph where all parentless vertices in H are also parentless in G. Then, let $G_1, \ldots, G_m$ be the rooted subgraphs of G_{test} that are color-isomorphic to G_{task} (i.e. the graph isomorphism preserves vertex color.) Let $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_m$ be the corresponding samples obtained from the subgraphs $G_1, \ldots, G_m$, and let $(\boldsymbol{u}_i, \boldsymbol{v}_i) = (\vec{u}(\boldsymbol{S}_i, W), \vec{v}(\boldsymbol{S}_i, W))$ for $i = 1, \ldots, m$. An unbiased estimate of $\text{Risk}(\vec{f})$ is given by

$$\widehat{\text{Risk}(\vec{f})} = \frac{1}{m}\sum_{i=1}^m \mathbf{E}[C(\vec{v}_i, \vec{f}(\vec{u}_i))|\boldsymbol{S}_{test}]$$

where the expectation is taken over the distribution of W.

Now we revisit the examples of supervised learning given in section 2.2.

- *Regression.* The task, training and test sampling schemes are all pair-sampling, with $n = 1$ for the task sampling scheme.

  $$\vec{u}(\boldsymbol{x}) = \boldsymbol{x}$$

  and

  $$\vec{v}(\boldsymbol{y}) = \boldsymbol{y}.$$

  The cost function for squared-error loss is

  $$C(\boldsymbol{y}, \hat{\boldsymbol{y}}) = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2.$$

- *Classification.* Here, $\boldsymbol{X}$ is the class label and $\boldsymbol{Y}$ is the feature vector. The sampling schemes are the same as in regression. The predictor and response are flipped:

  $$\vec{u}(\boldsymbol{x}) = \boldsymbol{x},$$

  $$\vec{v}(\boldsymbol{y}) = \boldsymbol{y}.$$

  And since $\boldsymbol{X}$ is discrete, it becomes reasonable to adopt the zero-one loss

  $$C(\boldsymbol{x}, \hat{\boldsymbol{x}}) = I\{\boldsymbol{x} \neq \hat{\boldsymbol{x}}\}.$$

- *Identification.* Let k be the number of classes in the test set. The sampling schemes are pair-sampling, with k pairs for the task sampling scheme. Let W be an independent uniform variate drawn from $\{1, \ldots, k\}$. We have

  $$\vec{u}((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_k, \boldsymbol{y}_k), W) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{y}_W).$$

  $$\vec{v}(W) = W.$$

  The cost function C is the zero-one loss.

We give one more example of a supervised learning task, due to its connection to information theory.

- *M, k-decoding for a random codebook.* X, Y are discrete random variables. The sampling schemes are pair-sampling, with Mk pairs for the task sampling scheme. Let W be an independent uniform variate drawn from $\{1, \ldots, M\}$. We have

$$\vec{u}((x_1, y_1), \ldots, (x_{Mk}, y_{Mk}), W) = (x_1, \ldots, x_{Mk}, y_{(W-1)k+1}, \ldots, y_{(W-1)k}).$$

$$\vec{v}(W) = W.$$

The cost function C is the zero-one loss.

The prediction task is the classical example of *decoding from a random codebook* seen in information theory. An information channel is given by $p(y|x)$ and a source distribution is given by $p(x)$. First, one forms a codebook consisting of M random codewords $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M\}$, where each codeword is a length-M vector

$$\boldsymbol{X}_i = (X_{(i-1)k+1}, \ldots, X_{(i-1)k}).$$

The codebook is shared with both the sender and the reciever. Next, the sender chooses a codeword at random to send over the channel, $\boldsymbol{X}_W$. The reciever observes the message

$$\boldsymbol{Y} = (Y_{(W-1)k+1}, \ldots, Y_{(W-1)k}).$$

Then, the decoding task is for the reciever to guess the index of the message, W, based on $\boldsymbol{Y}$. It is a consequence of a famous result in information theory, *the channel-coding theorem*, that taking $k \to \infty$ and $M = \exp[k\iota]$, then

$$\lim_{k \to \infty} R_{Bayes} = \begin{cases} 0 & \text{if } \iota \leq \mathrm{I}(X;Y), \\ 1 & \text{otherwise.} \end{cases}$$

Later, we make use of this result to relate mutual information to our general class of information coefficients.

3.3 A general class of information coefficients

Now that we have the axioms of information and a general definition of supervised learning, we can deliver the central concept in the paper: a general class of information coefficients based on supervised learning.

First, recall the intuitive characterization of information as "reduction of uncertainty." If we measure uncertainty using Shannon entropy $\mathrm{H}(\boldsymbol{X})$, then the resulting information coefficient is mutual information $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y})$. Therefore, to generalize this definition, we use a *general prediction task* to define uncertainty.

Let $\boldsymbol{S}$ be a sample drawn from sampling scheme p_{task}, $C : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a cost function, and let $\vec{v}$ be a function, with $\boldsymbol{V} = \vec{v}(\boldsymbol{S})$. Define the *unconditional uncertainty* of the task to be

$$\inf_{\hat{\boldsymbol{v}}} \mathbf{E}[C(\boldsymbol{V}, \hat{\boldsymbol{v}})].$$

Note that this is the Bayes risk for the prediction problem in the case that $\vec{u}$ is a trivial (constant) function. How much is the uncertainty reduced upon seeing $\boldsymbol{U}$? The *conditional uncertainty* is

$$\inf_{\vec{f}} \mathbf{E}[C(\boldsymbol{V}, \vec{f}(\boldsymbol{U}))].$$

Therefore, the resulting definition for a putative information coefficient is

$$\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) = \inf_{\hat{\boldsymbol{v}}} \mathbf{E}[C(\boldsymbol{V}, \hat{\boldsymbol{v}})] - \inf_{\vec{f}} \mathbf{E}[C(\boldsymbol{V}, \vec{f}(\boldsymbol{U}))].$$

However, some additional conditions on the prediction task are needed to ensure that $\mathcal{I}$ is indeed an information coefficient. Firstly, it is clear that $\mathcal{I}$ is a functional of $p(\boldsymbol{x}, \boldsymbol{y})$, and that it must be non-negative. However, we need to find conditions which ensure that $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) = 0$ in the case of independence, and to ensure the stochastic data-processing inequality.

We say that *V, U-orthogonality* holds for $(\vec{u}, \vec{v})$ if $\boldsymbol{X} \perp \boldsymbol{Y}$ implies $\vec{u}(\boldsymbol{S}) \perp \vec{v}(\boldsymbol{S})$ for all marginal distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. This is clearly met in the regression and classification case. The property holds in the identification case as well, since if $\boldsymbol{X} \perp \boldsymbol{Y}$ then $\boldsymbol{y}_W$ is independent of W.

We say that a function $\vec{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is *faithful* with respect to the arguments $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j\}$ if for any function $\vec{h}(\boldsymbol{x})$, one can find a function $\vec{g}$ such that

$$\vec{f}(\vec{h}(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_j), \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p) = \vec{g}(\vec{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)).$$

A sufficient condition for the data-processing inequality is that:

- *V-compatibility:* $\vec{v}$ not have any arguments in V_Y, and also

- *U-compatibility:* $\vec{u}$ is faithful with respect to its arguments in V_Y.

To show the data-processing inequality, consider the joint distribution of $(\boldsymbol{S}, b\tilde{S})$, where $\boldsymbol{S}$ is drawn from the task sampling scheme with the joint distribution P of $(\boldsymbol{X}, \boldsymbol{Y})$, and $\tilde{\boldsymbol{S}}$ is obtained by replacing $\boldsymbol{Y}_i$ in $\boldsymbol{S}$ with $\vec{h}(\boldsymbol{Y}_i)$. Letting W be an auxillary variable, obtain $\boldsymbol{V}$ by applying $\vec{v}$ to $(\boldsymbol{S}, W)$, obtain $b\tilde{V}$ by applying $\vec{v}$ to $(\tilde{\boldsymbol{S}}, W)$, and define $\boldsymbol{U}$ and $\tilde{\boldsymbol{U}}$ analogously. We have

$$\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) = \inf_{\hat{\boldsymbol{v}}} \mathbf{E}[C(\boldsymbol{V}, \hat{\boldsymbol{v}})] - \inf_{\vec{f}} \mathbf{E}[C(\boldsymbol{V}, \vec{f}(\boldsymbol{U}))],$$

$$\mathcal{I}(\boldsymbol{X};\vec{h}(\boldsymbol{Y})) = \inf_{\hat{\boldsymbol{v}}} \mathbf{E}[C(\tilde{\boldsymbol{V}}, \hat{\boldsymbol{v}})] - \inf_{\vec{f}} \mathbf{E}[C(\tilde{\boldsymbol{V}}, \vec{f}(\tilde{\boldsymbol{U}}))].$$

where all expectations are with respect to P. Because $\vec{v}$ has no arguments in V_Y, we have

$$\boldsymbol{V} = \tilde{\boldsymbol{V}}.$$

Therefore,

$$\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) - \mathcal{I}(\boldsymbol{X};\vec{h}(\boldsymbol{Y})) = \inf_{\vec{f}} \mathbf{E}[C(\boldsymbol{V}, \vec{f}(\tilde{\boldsymbol{U}}))] - \inf_{\vec{f}} \mathbf{E}[C(\boldsymbol{V}, \vec{f}(\boldsymbol{U}))].$$

However, due to faithfulness, $\tilde{\boldsymbol{U}}$ is a function of $\boldsymbol{U}$. Therefore, for any $\vec{f}$, we can find $\vec{g}$ such that

$$\vec{f}(\tilde{\boldsymbol{U}})) = \vec{g}(\boldsymbol{U}).$$

This implies that

$$\inf_{\vec{f}} \mathbf{E}[C(\boldsymbol{V}, \vec{f}(\boldsymbol{U}))] \leq \inf_{\vec{f}} \mathbf{E}[C(\boldsymbol{V}, \vec{f}(\tilde{\boldsymbol{U}}))].$$

and hence,

$$\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) \geq \mathcal{I}(\boldsymbol{X};\vec{h}(\boldsymbol{Y})).$$

We summarize our findings in the following theorem.

Theorem 3.1 *Let $(G_{task}, \vec{v}, \vec{u}, C)$ define a prediction task. Then, if $(\vec{v}, \vec{u})$ satisfy V,U-orthogonality, V-compatibility and U-compatibility, then*

$$\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) = \inf_{\hat{\boldsymbol{v}}} \boldsymbol{E}[C(\boldsymbol{V}, \hat{\boldsymbol{v}})] - \inf_{\vec{f}} \boldsymbol{E}[C(\boldsymbol{V}, \vec{f}(\boldsymbol{U}))]$$

defines an information coefficient, where $(\boldsymbol{V}, \boldsymbol{U})$ are the targets and predictions obtained from a task sample drawn using the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$.

$\square$.

We call such information coefficients *prediction-based* information coefficients.

Both classification and identification satisfy the three sufficient conditions of V,U-orthogonality and V/U-compatibility. Thus, classification and identification yield information coefficients of the form $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y})$. Regression satisfies V,U-orthogonality but violates V/U-compatibility. But, if we switch the place of $\boldsymbol{X}$ and $\boldsymbol{Y}$, then V/U compatibility holds. Therefore, the Bayes risk of regression can be used to obtain an information coefficient of the form $\mathcal{I}(\boldsymbol{Y};\boldsymbol{X})$.

We will take a closer look at some of these new information coefficients later on. However, first we show that *mutual information* is also a member of the same class.

3.4 Mutual information as a prediction-based information coefficient

What is prediction task yields mutual information as the associated information coefficient? It suffices to find a prediction task where the Shannon entropy is the Bayes risk. There is such a prediction task: *distribution estimation under logarithmic loss.*

Take a discrete random variable X with probability mass function $p(x)$. The prediction task is to estimate the mass function p. The loss is evaluated by drawing a new observation $X \sim p(x)$:

$$L(X, \hat{p}) = -\log \hat{p}(X).$$

Defining the risk as the expected loss, then it is a well-known result in information theory and statistics that taking $\hat{p} = p$ minimizes the risk, and the value of the minimal risk is the discrete entropy $\mathrm{H}_{discrete}(X)$. And, if one is allowed to observe Y before making the prediction, then the risk becomes $\mathrm{H}_{discrete}(X|Y)$. So the distribution estimation problem is exactly what we need.

Using the language of our framework for supervised learning,

- The sampling scheme for the task is to draw a single pair (X, Y).

- $\vec{u}(y) = y$.

- $\vec{v}(x) = \delta_x$, a vector with the value 1 at x and 0 in other entries.
- Prediction loss is a measure of uncertainty. Mean-square loss corresponds to variance.
- Consider two prediction problems: in one case you are given no side information, and in the second case you are given $\boldsymbol{Y}$. The difference in risk between the two problems gives a measure of uncertainty reduction.
- In regression, this gives "variance explained" as a measure of information.
- In classification, we get a normalized form of Bayes accuracy.
- In randomized classification, we get (normalized) average Bayes accuracy.
- We can prove that any prediction task yields an information measure i.e. satisfying the three axioms.
- If the no-information risk is known in advance, we say the information measure is estimable.
- Mutual information can be characterized this way, but it is not estimable.
- Mutual information can be derived as a limit of linear combinations of normalized ABA, due to channel coding theorem. Still not estimable due to limit property.

4 Statistical inference

- Only lower bounds are possible, because...

5 Connections

6 Applications

7 Discussion