ON MUTUAL INFORMATION AND IDEAL FEATURE CLASSIFICATION

PATRICK BARDSLEY

ABSTRACT. We concern ourselves with quantifying the extent a multidimensional set of feature vectors $\mathbf{x}_i \in \mathbb{R}^N$ can be accurately classified. Foremost in our study is the concept of mutual information [6, 2, 1, 5]. We recall essential definitions and interpretations from information theory. Then, leveraging these interpretations, we derive information-content metrics designed to quantify the class label information present in a given feature design and thus its 'classifiability'. Formal derivations of these information theory estimators and metrics are given in the appendix.

CONTENTS

1. INTRODUCTION AND NOTATIONAL CONVENTIONS

A recurring problem in industrial and academic applications is that of classification. This problem is the act of assigning one of several discrete categories or class labels to a given realization of data, termed a feature. Mathematically, we assume a feature is simply a vector $\mathbf{f} \in \mathbb{R}^N$, and thus a classifier is the mapping $C : \mathbb{R}^N \to \mathcal{L}$ that takes a feature $\mathbf{f}$ to a label:

$$C(\mathbf{f}) = \widehat{\ell},$$

where $\widehat{\ell} \in \mathcal{L}$ and $\mathcal{L}$ denotes the *discrete* set of possible class labels

$$\mathcal{L} := \{\ell_1, \ldots, \ell_L\}.$$

A natural question to ask, and is the one we concern ourselves with in the remainder of this document, is to what extent *any* classifier (i.e., any mapping $C : \mathbb{R}^N \to \mathcal{L}$) can accurately classify a given feature based on the feature design (i.e., how the feature was extracted and generated from raw data). In other words, how much label information is contained in the feature design? We address this question in the following sections using well-known quantities in information theory (see, e.g., [6, 2, 1, 5]).

Throughout this document, we assume the availability of a dataset $\mathcal{D}$, consisting of M features with N-dimensions:

$$\mathcal{D} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M\}$$

where

$$\mathbf{f}_i \in \mathbb{R}^N \quad \text{for } i = 1, \ldots, M.$$

For each feature vector $\mathbf{f}_i, i = 1, \ldots, M$, we denote its associated *true* class label as $\ell_i \in \mathcal{L}$. Estimates or *predictions* of the label will be denoted as $\widehat{\ell_i}$.

The remainder of this document is organized as follows. In §2 we recall essential definitions and interpretations from information theory, chief among them being that of mutual information. Then in §3 we leverage these interpretations to form feature design metrics. These metrics are intended to quantify the class label information content held within a given feature design, and the extent we expect the design to perform well in classification tasks. We give full, albeit formal, (re-)derivations of estimators for entropy and mutual information in Appendix A.

2. Information theory essentials

The concept of "information" is given a precise mathematical definition in the vast literature of information theory (see, e.g., [6, 2, 1, 5]). Essentially, the theory is built on the concept of self-information of a given event which is defined as

$$\mathrm{SI}(A) = -\log(P(A)), \tag{1}$$

for an event A. This definition follows from requiring the "information" of an event to meet several axioms:

- Information (i.e., SI) depends only on the probability of an event
- Events with $P = 1$ convey no information (i.e., $P(A) = 1 \implies \mathrm{SI}(A) = 0$)
- Less-probable events are more surprising to encounter and thus convey more information (i.e., SI must be monotonic decreasing)
- The self-information of two independent events must equal the sum of the self-information of the two separate events (i.e., $P(A \cap B) = P(A)P(B) \implies \mathrm{SI}(A \cap B) = \mathrm{SI}(A) + \mathrm{SI}(B)$)

From (1), it is then possible to define the important quantity of *entropy* for a given *discrete* random variable Y:

$$\begin{aligned} H(Y) &= \mathbb{E}(\mathrm{SI}(Y)) \\ &= -\sum_y P(Y = y) \log P(Y = y) dy, \end{aligned} \tag{2}$$

where the sum runs over all possible values of Y. For our purposes, the base of the logarithm here is relatively unimportant as it simply gives a scale change, but is often base-2 to give information contents in 'bits'. From (1) and (2) we see entropy is the average self-information of the random variable Y and is often interpreted as the amount of uncertainty of Y. In other words, it is the amount of additional information that must be specified to completely identify the value of Y. Moreover, if Y is in fact deterministic such that $Y \equiv y_*$ and thus $P(Y = y) = \delta_{yy_*}$, we obtain

$$\begin{aligned} H(Y) &= -\sum_y P(Y = y) \log P(Y = y) \\ &= -\sum_{y \neq y_*} P(Y = y) \log P(Y = y) - P(Y = y_*) \log P(Y = y_*) \\ &= 0, \end{aligned}$$

where by convention we assume $0 \log 0 = 0$. That is, no additional information needs to be specified to determine the value of Y as it always assumes the value y_*.

Mutual information between two random variables X and Y can in turn defined as

$$I(X; Y) = H(Y) - H(Y|X), \tag{3}$$

where, $H(Y|X)$ denotes the conditional entropy of Y given X:

$$\begin{aligned} H(Y|X) &= \mathbb{E}_x(\mathbb{E}_y(\mathrm{SI}(Y|X = x))) \\ &= -\sum_x P(X = x) \sum_y P(Y = y|X = x) \log P(Y = y|X = x). \end{aligned}$$

From (3) we see $I(X; Y)$ quantifies the *reduction in uncertainty* or *gain of information* of Y given the value of X. Furthermore, by dividing (3) by the entropy of Y, we can obtain the *normalized mutual information*:

$$\mathrm{NMI}(X; Y) := I(X; Y)H(Y)^{-1} = 1 - H(Y|X)H(Y)^{-1}, \tag{4}$$

This quantity satisfies

- $\mathrm{NMI}(X; Y) = 0$ if $H(Y|X) = H(Y) \iff X$ and Y are independent,
- $\mathrm{NMI}(X; Y) = 1$ if $H(Y|X) = 0$ (i.e., the uncertainty of Y is 0 if the value of X is known).

Thus, $\mathrm{NMI}(X; Y)$ can be interpreted as the %-gain of information about Y given a measured value of X.

Our above (re-)formulation of entropy and mutual information is precise for *discrete* random variables. However, a rigorous analogue for *continuous* random variables is also possible by essentially replacing the sums over probability sample spaces with integrals. That is, assuming sufficient regularity of the distribution of Y such that it admits a probability density $\rho(y)$, we can define the differential entropy as

$$H(Y) = -\int \rho(y) \log \rho(y) dy \tag{5}$$

where the integral is taken over the support of $\rho(y)$. The continuous analogue of $I(X; Y)$ is then

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y|X) \\ &= -\int \rho(y) \log \rho(y) dy + \int \rho(x) \int \rho(y|x) \log \rho(y|x) dy dx \end{aligned} \tag{6}$$

where $\rho(y|x)$ denotes the conditional density function

$$\rho(y|x) = \frac{\rho(x, y)}{\rho(x)}.$$

While differential entropy is not exactly analogous to discrete entropy (see Remark 1), mutual information retains its interpretation as a reduction of uncertainty or gain in information of one variable given another. In the remainder of this document, we will make use of both discrete and differential entropy formulas depending on the type of random variable in question.

Remark 1 (Differential entropy and limiting density of discrete points)**.** *In the original formulation of differential entropy [6], Claude Shannon assumed incorrectly that* (5) *was the continuous analogue of the discrete entropy* (2)*. As later shown by Jaynes [3], the correct analogue is obtained via the limit*

$$H(Y) = \lim_{k\to\infty}\left(log(k) - \int \rho(y)\log\frac{\rho(y)}{q(y)}dy\right) \tag{7}$$

where $q(y)$ is termed the 'invariant measure' and is the limiting density of a set of discrete points:

$$q(y) = \lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}\delta(y-y_i).$$

For our purposes, we can consider $q(y)$ to be the uniform probability density on the support of $\rho(y)$. Thus, (7) *is the (negative) Kullback-Leibler divergence (i.e., relative entropy) of $\rho(y)$ from $q(y)$, plus an infinite offset. To avoid this infinite offset problem for continuous random variables, entropy is typically treated and interpreted* only *in a relative sense (i.e., relative to a base or reference measure $q(y)$). However, in the convenient case of mutual information, the term giving rise to the infinite offset in fact cancels due to the form of $I(X;Y) = H(Y) - H(Y|X)$.*

3. Information content and feature design

We turn now to estimating the information content contained in a given feature design. In particular, we provide two metrics to quantify how well a feature design could potentially perform in an ideal classifier. That is, our estimates are *independent* of any particular classifier training or architecture allowing us to evaluate the feature design itself. In §3.1, we estimate class label information spread across dimensions with each dimension treated independently. Then in §3.2 we give a more general metric which simply estimates the mutual information between a feature design and the class labels.

3.1. **Information concentration.**

Estimating mutual information is relatively straightforward in low dimensions (e.g., $N = 1$). A density estimator (see, e.g., [5]) can be constructed for the underlying probability distributions, and then numerical approximations of the integrals defined in (3) can be utilized. Thus, given a dataset $\mathcal{D}$, we can readily estimate the mutual information of the class label ℓ and the j-th feature dimension $f_j := (\mathbf{f})_j$ as

$$\begin{aligned} I(f_j;\ell) &= H(\ell) - H(\ell|f_j) \\ &= \sum_{\ell'\in\mathcal{L}}\widehat{\rho}_{\ell'}\left(\int \widehat{\rho}(f_j|\ell=\ell')\log\frac{\widehat{\rho}(f_j|\ell=\ell')\widehat{\rho}_{\ell'}}{\widehat{\rho}(f_j)}df_j - \log\widehat{\rho}_{\ell'}\right), \end{aligned} \tag{8}$$

where the $\widehat{\cdot}$ denotes probability density/mass estimators computed from the dataset $\mathcal{D}$. The normalized mutual information is then easily computed as

$$\mathrm{NMI}(f_j;\ell) = I(f_j;\ell)H(\ell)^{-1}.$$

To estimate the spread of information across the various dimensions for the dataset $\mathcal{D}$ (i.e, the information concentration), we consider the ratio of minimal information to maximal information, i.e.,

$$\mathrm{IC}(\mathcal{D}) = \frac{\min_j I(f_j;\ell)}{\max_j I(f_j;\ell)}. \tag{9}$$

This quantity satisfies the following:

- $\mathrm{IC}(\mathcal{D}) \in [0,1]$,
- $\mathrm{IC}(\mathcal{D}) \approx 0 \implies$ at least one dimension $j = 1, \ldots, N$ contains insignificant information of the class label,
- $\mathrm{IC}(\mathcal{D}) \approx 1 \implies$ all feature dimensions contain essentially the same amount of class label information.

Therefore, $\mathrm{IC}(\mathcal{D})$ quantifies the spread across dimensions of class label information for a feature design, with values close to 1 giving uniform spread and values close to 0 indicating degenerate or superfluous feature dimensions. We explicitly note, values of $\mathrm{IC}(\mathcal{D}) \approx 1$ do not imply significant amounts of class label information, simply that all dimensions contain roughly the same amount of information. Therefore, this metric should be considered in combination with $\max_j \mathrm{NMI}(f_j;\ell)$.

3.2. **Mutual information.** We now aim to compute the normalized mutual information between the full feature vector $\mathbf{f} = [f_1, \ldots, f_M]^T$ and associated class labels. In an information-theoretic sense, this is similar to computing the correlation between the feature vectors and the class labels in that a value of 1 indicates perfect knowledge (i.e., perfect correlation) of the feature with its label, and a value of 0 indicates statistical independence[1] between features and labels. Said differently, we aim to quantify the information gained about the class label when given knowledge of the full feature vector.

The mutual information between the feature vector $\mathbf{f}$ and the label ℓ can be expressed analytically as

$$I(\mathbf{f};\ell) = \sum_{\ell' \in \mathcal{L}} \rho_{\ell'} \left(\int \rho(\mathbf{f}|\ell = \ell') \log \frac{\rho(\mathbf{f}|\ell = \ell')\rho_{\ell'}}{\rho(\mathbf{f})} d\mathbf{f} - \log \rho_{\ell'} \right),$$

This formula is clearly restrictive for large dimension N, as naively computing it analogous to (8) requires density estimates for $\rho(\mathbf{f}|\ell)$ and $\rho(\mathbf{f})$, which is computationally prohibitive even for $N \geq 3$. However, a method developed by Kraskov et. al [4] provides a computationally efficient way to estimate $I(\mathbf{f};\ell)$ without the need for these density estimators. We make minor modifications/specifications here to adapt this method to our particular semi-discrete probability space consisting of the continuous-valued feature vectors $\mathbf{f}$ and discrete-valued class labels ℓ.

[1] Vanishing mutual information is in fact a stronger statement than vanishing correlation as $I(X;Y) = 0 \iff X$ and Y are statistically independent, while $\mathrm{Corr}(X,Y) = 0 \;\not\Rightarrow\; X$ and Y are independent.

First, we re-express the mutual information (10) as

$$
\begin{aligned}
I(\mathbf{f};\ell) &= \sum_{\ell'\in\mathcal{L}} \int \rho(\mathbf{f}|\ell=\ell')\rho_{\ell'} \log \frac{\rho(\mathbf{f}|\ell=\ell')\rho_{\ell'}}{\rho(\mathbf{f})} d\mathbf{f} - \rho_{\ell'} \log \rho_{\ell'} \\
&= \sum_{\ell'\in\mathcal{L}} \Big(\int \rho(\mathbf{f}|\ell=\ell')\rho_{\ell'} \log \rho(\mathbf{f}|\ell=\ell')\rho_{\ell'} d\mathbf{f} - \int \rho(\mathbf{f}|\ell=\ell')\rho_{\ell'} \log \rho(\mathbf{f}) d\mathbf{f} \\
&\qquad - \rho_{\ell'} \log \rho_{\ell'} \Big) \\
&= -\int \rho(\mathbf{f}) \log \rho(\mathbf{f}) d\mathbf{f} \\
&\quad - \sum_{\ell'\in\mathcal{L}} \rho_{\ell'} \left(\log \rho_{\ell'} - \int \rho(\mathbf{f}|\ell=\ell') \log \rho(\mathbf{f}|\ell=\ell')\rho_{\ell'} d\mathbf{f} \right) \\
&= -\int \rho(\mathbf{f}) \log \rho(\mathbf{f}) d\mathbf{f} + \sum_{\ell'\in\mathcal{L}} \rho_{\ell'} \int \rho(\mathbf{f}|\ell=\ell') \log \rho(\mathbf{f}|\ell=\ell') d\mathbf{f} \\
&= H(\mathbf{f}) - \sum_{\ell'\in\mathcal{L}} \rho_{\ell'} H(\mathbf{f}|\ell=\ell').
\end{aligned}
\tag{10}
$$

Next, we rely on the KSG-method [4] to estimate the two differential entropy terms $H(\mathbf{f})$ and $H(\mathbf{f}|\ell=\ell')$ as follows. For each feature vector $\mathbf{f}_i \in \mathcal{D}$ we define

$$
d(\mathbf{f}_i;k) = \|\mathbf{f}_i - \mathbf{f}_{k_\ell(i)}\|, \tag{11}
$$

where $\mathbf{f}_{k_\ell(i)}$ denotes the k-th nearest neighbor of $\mathbf{f}_i$ with the *same class label* ℓ. Thus, $d(\mathbf{f}_i;k)$ is the distance from $\mathbf{f}_i$ to its k-th nearest same-label-neighbor. Next, we define

$$
n(\mathbf{f}_i) = \sum_{i'\neq i}^{M} \mathcal{I}\big(\|\mathbf{f}_i - \mathbf{f}_{i'}\| < d(\mathbf{f}_i;k)\big), \tag{12}
$$

as the number of features, regardless of their class label, that are strictly closer to $\mathbf{f}_i$ than its k-th nearest same-label-neighbor. As shown in [4] (see also Appendix A), the differential entropy terms can then be approximated via the estimators

$$
\begin{aligned}
\widehat{H}(\mathbf{f}) &= -\frac{1}{M}\sum_{i=1}^{M} \psi(n(\mathbf{f}_i)+1) + \psi(M) + \log|\mathcal{B}_N| + \frac{N}{M}\sum_{i=1}^{M} \log d(\mathbf{f}_i;k), \\
\widehat{H}(\mathbf{f}|\ell=\ell') &= -\psi(k) + \psi(M_{\ell'}) + \log|\mathcal{B}_N| + \frac{N}{M_{\ell'}}\sum_{i=1}^{M} \log d(\mathbf{f}_i;k)\mathcal{I}(\ell_i=\ell'),
\end{aligned}
\tag{13}
$$

where $M_{\ell'}$ is the number of vectors with class label ℓ', $|\mathcal{B}_N|$ is the volume of the N-dimensional unit ball, and ψ is the digamma function:

$$
\psi(x) = \frac{d}{dx} \ln \Gamma(x)
$$

with

$$
\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} dy.
$$

Combining (10) with (13) we have the estimator

(14)

$$\begin{aligned}\widehat{I}(\mathbf{f};\ell) &= -\frac{1}{M}\sum_{i=1}^{M}\psi(n(\mathbf{f}_i)+1)+\psi(M)+\log|\mathcal{B}_N|+\frac{N}{M}\sum_{i=1}^{M}\log d(\mathbf{f}_i;k)\\ &\qquad -\sum_{\ell'\in\mathcal{L}}\widehat{\rho}_{\ell'}\left(\psi(M_{\ell'})-\psi(k)+\log|\mathcal{B}_N|+\frac{N}{M_{\ell'}}\sum_{i=1}^{M}\log d(\mathbf{f}_i;k)\mathcal{I}(\ell_i=\ell')\right)\\ &= \psi(M)+\psi(k)-\sum_{\ell'\in\mathcal{L}}\widehat{\rho}_{\ell'}\psi(M_{\ell'})-\frac{1}{M}\sum_{i=1}^{M}\psi(n(\mathbf{f}_i)+1),\end{aligned}$$

where $\widehat{\rho}_{\ell'} = M_{\ell'}/M$. Here we have used that $\sum_{\ell'\in\mathcal{L}}\widehat{\rho}_{\ell'} = 1$ and $\widehat{\rho}_{\ell'}N/M_{\ell'} = N/M$ for all $\ell' \in \mathcal{L}$. We note the selection of the k-th nearest same-label-neighbor and subsequent definition of $n(\mathbf{f}_i)$, was precisely to achieve the cancellation of the $\log d(\mathbf{f}_i;k)$ terms in the above formula; these are essentially the specializations we made in modifying the method in [4] to our semi-discrete purpose (see Appendix A). Lastly, we note that the parameter k is a free parameter, though it may affect the bias of the estimate[2].

In a slight abuse of notation, we define the metric

$$\mathrm{NMI}(\mathcal{D}) := \frac{I(f;\ell)}{H(\ell)},$$

where $I(\mathbf{f};\ell)$ is estimated using (14) for the dataset $\mathcal{D}$. This metric satisfies the following properties:

- $\mathrm{NMI}(\mathcal{D}) \approx 0 \implies$ no substantial information gain about class label ℓ from knowledge of the feature vector $\mathbf{f}$,
- $\mathrm{NMI}(\mathcal{D}) \approx 1 \implies$ perfect knowledge of class label ℓ from knowledge of feature vector $\mathbf{f}$ is possible if an ideal (i.e., perfect) classifier is used,
- $\mathrm{NMI}(\mathcal{D})$ can be interpreted as the %-gain of information about a class label ℓ given knowledge of the feature vector $\mathbf{f}$.

Therefore, $\mathrm{NMI}(\mathcal{D})$ quantifies the extent to which a feature design is classifiable with values near 1 (resp. 0) indicating the feature design is an ideal (resp. poor) choice for classification. If a good feature design is chosen for a given dataset $\mathcal{D}$, as indicated by $\mathrm{NMI}(\mathcal{D}) \approx 1$, it can then be inferred that any loss of performance is due entirely to classifier architecture or training.

Appendix A. Computing mutual information in $\mathbb{R}^N$

Here we give a formal derivation of the mutual information estimate we use in (14). This estimate in turn relies on a differential entropy estimate whose formal derivation was first obtained by Kraskov et al. [4]. As far as this author is aware, these formal derivations have not yet been made rigorous and appeal more to intuition. Nonetheless, in §A.1 we recall the derivation from [4] of an estimate for the differential entropy. Then in §A.2 we make minor adaptations to account for our semi-discrete context, and estimate the mutual information between feature vectors $\mathbf{f}$ and class labels ℓ.

[2] To our knowledge, no rigorous study of this estimator has been undertaken, only formal arguments

A.1. **Kraskov-Stögbauer-Grassberger (KSG)-estimate for differential entropy.** Assume, as before, the vectors $\mathbf{f}_i$ for $i = 1, \dots, M$ are i.i.d. according to some unknown distribution, but which admits a density function $\rho(\mathbf{f})$. Then, if we had an estimator $\widehat{\log \rho}(\mathbf{f}_i)$, the differential entropy could simply be found by leveraging Monte-Carlo integration:

$$\widehat{H}(\mathbf{f}) = -\frac{1}{M}\sum_{i=1}^{M} \widehat{\log \rho}(\mathbf{f}_i). \tag{15}$$

Thus, we first proceed to find obtain an estimator $\widehat{\log \rho}(\mathbf{f}_i)$.

For the moment, we fix a single vector $\mathbf{f}_i$ and denote the distances to the other vectors as

$$\epsilon(\mathbf{f}_i; k) = \|\mathbf{f}_i - \mathbf{f}_{k(i)}\|,$$

where $\mathbf{f}_{k(i)}$ is the k-th nearest neighbor to $\mathbf{f}_i$. Thus we have the ordering

$$\epsilon(\mathbf{f}_i; 1) \leq \epsilon(\mathbf{f}_i; 2) \leq \cdots \leq \epsilon(\mathbf{f}_i; M-1).$$

Consider now the probability that $\epsilon(\mathbf{f}_i; k) \in [\mu, \mu + \delta\mu]$ for some distance $\mu \in \mathbb{R}^+$, and denote this probability $P_k(\mu)$. Assuming no other neighbors fall in this same μ-shell, we can use the multinomial formula to find

$$P_k(\mu) = \binom{M-1}{k-1, 1, M-k-1} p_i^{k-1} \frac{dp_i}{d\mu}(1 - p_i)^{M-k-1},$$

where we have used $p_i(\mu)$ as the ρ-measure of the μ-sphere centered at $\mathbf{f}_i$:

$$p_i(\mu) = \int_{\|\mathbf{f}' - \mathbf{f}_i\| \leq \mu} \rho(\mathbf{f}') d\mathbf{f}'.$$

A straightforward computation[3] now shows

$$\begin{aligned} E(\log p_i) &= \int_0^\infty \log(p_i(\mu)) P_k(\mu) d\mu \\ &= \binom{M-1}{k-1, 1, M-k-1} \int_0^\infty \log(p_i(\mu)) p_i(\mu)^{k-1} \frac{dp_i}{d\mu}(\mu)(1 - p_i(\mu))^{M-k-1} d\mu \\ &= \binom{M-1}{k-1, 1, M-k-1} \int_0^1 \log(p) p^{k-1}(1-p)^{M-k-1} dp \\ &= \psi(k) - \psi(M). \end{aligned}$$

If we now approximate $\rho(\mathbf{f}) \equiv \rho(\mathbf{f}_i)$ for all $\mathbf{f} \in \{\mathbf{f}' : \|\mathbf{f}' - \mathbf{f}_i\| \leq \mu\}$ (i.e., assume $\rho(\mathbf{f})$ is constant in the entire μ-ball), we easily obtain the approximation

$$p_i(\mu) \approx |\mathcal{B}_N| \mu^N \rho(\mathbf{f}_i),$$

where $|\mathcal{B}_N| = \pi^{N/2}/\Gamma(N/2+1)$ is the volume of the unit N-sphere. This then leads us to the estimator:

$$\begin{aligned} \widehat{\log \rho}(\mathbf{f}_i) &\approx E(\log p_i - N \log \mu - \log |\mathcal{B}_N|) \\ &= \psi(k) - \psi(M) - N \log \epsilon(\mathbf{f}_i; k) - \log |\mathcal{B}_N|, \end{aligned} \tag{16}$$

where we recall $\epsilon(\mathbf{f}_i; k)$ is distance from $\mathbf{f}_i$ to its k-th nearest neighbor.

[3] As of writing this document, we are unable to verify the final equality $\int \log(p) p^{k-1}(1-p)^{M-k-1} dp \propto \psi(k) - \psi(M)$ by direct computation. Presently, we assume the correctness of the original authors Kraskov et al. [4] and will verify this formula at a later date.

Finally, combining (15) and (16), we have

$$\widehat{H}(\mathbf{f}) \approx \psi(M) - \psi(k) + \frac{N}{M}\sum_{i=1}^{M} \log \epsilon(\mathbf{f}_i; k) + \log |\mathcal{B}_N|, \tag{17}$$

where k is a free parameter.

A.2. **k-th nearest same-label-neighbor.** Having estimated the differential entropy with (17), we now make minor modifications to the mutual information estimate found in [4] to account for our semi-discrete context (recall our feature vectors $\mathbf{f}_i$ are assumed continuously valued while the class labels ℓ_i are discrete). Essentially, these modifications are to choose the appropriate value of k in the differential entropy formulas to achieve proper cancellation of terms in the mutual information estimate

$$\widehat{I}(\mathbf{f}; \ell) = \widehat{H}(\mathbf{f}) - \sum_{\ell' \in \mathcal{L}} \widehat{\rho}_{\ell'} \widehat{H}(\mathbf{f} | \ell = \ell').$$

We note that in estimating the conditional entropy terms $\widehat{H}(\mathbf{f}|\ell = \ell')$ using (17), we are simply estimating the differential entropy but using *only* vectors $\mathbf{f}_i$ whose corresponding label $\ell_i = \ell'$. Thus, we can express this estimate as

$$\begin{aligned}\widehat{H}(\mathbf{f}|\ell = \ell') &\approx \psi(M_{\ell'}) - \psi(k) + \log |\mathcal{B}_N| + \frac{N}{M_{\ell'}}\sum_{i=1}^{M} \log \epsilon(\mathbf{f}_i; k) \mathcal{I}(\ell_i = \ell') \\ &= \psi(M_{\ell'}) - \psi(k) + \log |\mathcal{B}_N| + \frac{N}{M_{\ell'}}\sum_{i=1}^{M} \log d(\mathbf{f}_i; k) \mathcal{I}(\ell_i = \ell')\end{aligned}$$

where $M_{\ell'} = \sum_{i=1}^{M} \mathcal{I}(\ell_i = \ell')$ is the number of data with label $\ell_i = \ell'$ and $d(\mathbf{f}_i; k)$, defined in (11), is the distance to the k-th nearest *same-label-neighbor.* This replacement of $\epsilon(\mathbf{f}_i; k)$ (i.e., the distance to the k-th nearest neighbor) with $d(\mathbf{f}_i; k)$ (i.e., the distance to the k-th nearest same-label-neighbor) is the critical observation we make in adjusting our mutual information estimate.

Since the estimate (16) holds for any value of the free parameter k, we are free to choose this value (and its associated ϵ value) for each data point $i = 1, \ldots, M$, independently. Thus, we may choose $k(i) = n(\mathbf{f}_i) + 1$ where $n(\mathbf{f}_i)$, defined in (12), is the number of features strictly strictly closer to $\mathbf{f}_i$ than its k-th nearest same-label-neighbor. In doing so, we then have $\epsilon(\mathbf{f}_i; k) = d(\mathbf{f}_i; k)$ and our estimate of the unconditional entropy becomes

$$\widehat{H}(\mathbf{f}) \approx \psi(M) - \frac{1}{M}\sum_{i=1}^{M} \psi(n(\mathbf{f}_i) + 1) + \frac{N}{M}\sum_{i=1}^{M} \log d(\mathbf{f}_i; k) + \log |\mathcal{B}_N|.$$

This choice of $k(i)$ and $\epsilon(\mathbf{f}_i;k)$ was done precisely to achieve cancellation of the $\log d(\mathbf{f}_i;k)$ terms in our final expression of the mutual information estimate:

$$\begin{aligned}\widehat{I}(\mathbf{f};\ell) &= \widehat{H}(\mathbf{f}) - \sum_{\ell'\in\mathcal{L}} \widehat{\rho}_{\ell'} \widehat{H}(\mathbf{f}|\ell=\ell') \\ &= \psi(M) - \frac{1}{M}\sum_{i=1}^{M}\psi(n(\mathbf{f}_i)+1) + \frac{N}{M}\sum_{i=1}^{M}\log d(\mathbf{f}_i;k) + \log|\mathcal{B}_N| \\ &\quad - \sum_{\ell'\in\mathcal{L}} \widehat{\rho}_{\ell'}\left(\psi(M_{\ell'}) - \psi(k) + \log|\mathcal{B}_N| + \frac{N}{M_{\ell'}}\sum_{i=1}^{M}\log d(\mathbf{f}_i;k)\mathcal{I}(\ell_i=\ell')\right) \\ &= \psi(M) + \psi(k) - \sum_{\ell'\in\mathcal{L}} \widehat{\rho}_{\ell'}\psi(M_{\ell'}) - \frac{1}{M}\sum_{i=1}^{M}\psi(n(\mathbf{f}_i)+1),\end{aligned}$$

which follows immediately from $\sum_{\ell'\in\mathcal{L}}\widehat{\rho}_{\ell'} = 1$ and $\widehat{\rho}_{\ell'}N/M_{\ell'} = N/M$ for all $\ell' \in \mathcal{L}$.

References

[1] R. M. Gray. *Entropy and information theory.* Springer Science & Business Media, 2011.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, 2009.

[3] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[4] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[5] K. P. Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[6] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.