Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Relation between classification accuracy and mutual information in equally weighted classification tasks

Master thesis
in Zusammenarbeit

Arbeitsbereich Theory of Machine Learning
Prof. Dr. U. v. Luxburg

Fachbereich Informatik (Wilhelm-Schickard-Institut)
Mathematisch-Naturwissenschaftliche Fakultät
Universität Tübingen

und

Arbeitsbereich Knowledge Technology, WTM
Prof. Dr. S. Wermter

Department Informatik
MIN-Fakultät
Universität Hamburg

vorgelegt an der Universität Hamburg von
Sascha Meyen
am
2.12.2016

Gutachter: Prof. Dr. U. v. Luxburg
Prof. Dr. S. Wermter

Sascha Meyen
Matrikelnummer: 6204566
Augustenburger Ufer 7
22049 Hamburg

Abstract

In classification tasks where an observation is used to predict a label, we investigate the relation between the two measures – *classification accuracy* and *mutual information*. Assuming equally weighted labels we prove that the two measures impose lower and upper bounds on each other because they capture related but different properties of the observation. Processing the observation may discard these properties lowering the mutual information while the classification accuracy is unaffected. Because the effect of such aggregations was neglected, a flawed conclusion has been drawn in psychological research about conscious vs. unconscious processing. Additionally, we demonstrate in which scenarios it is particularly important to pay attention to the mutual information. Altogether, the mutual information complements the classification accuracy and we recommend to consider it as a quality measure in classification tasks. Open questions remain about how to deal with the inherently biased estimators for mutual information and what effect mutual information has on boosting.

Zusammenfassung

Bei Klassifikationsaufgaben werden Beobachtungen zur Vorhersage ihrer Klassenzugehörigkeit verwendet. Dabei kann die *Vorhersagegenauigkeit* oder die *gemeinsame Information* als Maß für die Güte der Klassifikation verwendet werden. Unter der Annahme gleichverteilter Klassenzugehörigkeiten untersuchen wir den Zusammenhang dieser beiden Maße. Beide Maße geben einander untere und obere Schranken, da sie verwandte aber unterschiedliche Eigenschaften der Beobachtung messen. Das Zusammenfassung von Beobachtungen kann diese Eigenschaften verwerfen und damit die gemeinsame Information verringert, ohne dabei dabei die Vorhersagegenauigkeit zu reduzieren. Weil die Effekte solcher Zusammenfassungen nicht beachtet wurden, kam es zu einer fehlerhaften Interpretation in der psychologischen Forschung zu bewusster vs. unbewusster Verarbeitung. Zusätzlich zeigen wir, in welchen Szenarien es besonders wichtig ist, die gemeinsame Information als Maß zu beachten. Insgesamt ergänzt die gemeinsame Information die Vorhersagegenauigkeit als Gütekriterium in Klassifikationsaufgaben. Eine offene Forschungsfrage bleibt, wie mit inheränt verzerrten Schätzern für die gemeinsame Information umgegangen werden soll und welchen Effekt die gemeinsame Information auf Boosting hat.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In classification tasks an observation X is used to predict a label Y. Researchers typically evaluate such tasks by measuring the classification accuracy. This measure indicates the probability of making a correct prediction. However, this approach alone neglects relevant aspects of the classification task (Congalton, 1991, Provost et al., 1997, Baldi et al., 2000). To complement measuring classification accuracy researchers can apply information theory (IT) and measure the mutual information. The mutual information indicates how much uncertainty about Y is reduced by observing X. In this thesis we investigate the relation between the two measures, *classification accuracy* and *mutual information*. We extend the previous research on this topic by Fisher et al. (2009) and Hu (2014), who only considered binary labels, to non-binary labels Y.

Consider the two observations X_a and X_b as indicators for a binary label Y in table 1.1. The first observation is either $X_a = -1$ or $X_a = 1$ and it indicates the correct label with a probability of 80%, see table 1.1(a). The second observation is sometimes informative and sometimes non-informative. An informative observation of $X_b = -1$ or $X_b = 1$ always allows for a correct prediction, but we do not always get such an informative observation. The non-informative observations $X_b = 0$ only allows to guess the correct label at chance level, see table 1.1(b). The classification accuracy of both observations is at 80%.

		X_a	
		-1	1
Y	-1	0.4	0.1
	1	0.1	0.4

(a)

		X_b		
		-1	0	1
Y	-1	0.3	0.2	0.0
	1	0.0	0.2	0.3

(b)

Table 1.1: Two contingency tables with the same classification accuracy $acc(Y|X_a) = acc(Y|X_b) = 0.8$ but with different mutual information $I(Y;X_a) = 0.278bit \neq I(Y;X_b) = 0.6bit$.

Nevertheless, we argue that X_b is strictly better than X_a. Intuitively, X_b tells us how accurate our prediction will be. Predictions based on informative observations

are always correct. Predictions based on non-informative observations are known to be a random guessing (conditional accuracies are 100% or 50%). On the other hand the first observation X_a does not allow to distinguish informative vs. non-informative observations (conditional accuracies are 80%). We can not distinguish between these two observations using the classification accuracy but by measuring the mutual information for both observations, $I(Y; X_a) = 0.278bit$ vs. $I(Y; X_b) = 0.6bit$. Then, the superiority of X_b over X_a is revealed because it carries more information about the label Y.

We recommend that researchers pay attention to the mutual information as a measure complementing the classification accuracy. To derive this conclusion we will first give the necessary definitions in chapter 2. With this we will show that:

1. A fixed classification accuracy defines lower and upper bounds on the mutual information and vice versa because both measures capture different properties of the observation (section 3.1).

2. The mutual information can decrease when observations are further processed indicating that desirable properties are lost even though the classification accuracy stays unchanged (section 3.2).

3. Some psychological researchers draw flawed conclusions from comparing different observations because they do not acknowledge that processed observation may inherently convey less information (chapter 4).

4. It is particularly important to pay attention to the mutual information in "risky" scenarios where it is favorable to make no prediction instead of an unreliably prediction (chapter 5).

Chapter 2

Definitions and assumptions

In this thesis we investigate the relationship between classification accuracy and mutual information. First, we provide the relevant definitions for classification accuracy, conditional accuracy and mutual information. Then, we state four assumptions to define the setting we are in.

2.1 Classification accuracy and conditional accuracy

In a classification task an observation X is used to predict a label Y. The probability of a correct prediction is the classification accuracy. The highest classification accuracy is achieved by predicting the most probable label $Y = y$ given an observation $X = x$. This optimal prediction strategy defines the Bayes classifier C^*.

Definition 1 (**Bayes classifier**). *The Bayes classifier C^* predicts for a given observation $X = x$ the most probable label $Y = y$.*

$$\begin{aligned} C^*(X = x) &= \operatorname*{argmax}_{y \in \Omega_Y} P(Y = y | X = x) \\ &= \{y | P(Y = y | X = x) = \max_{y^*} P(Y = y^* | X = x)\} \end{aligned}$$

We denote the support of Y by Ω_Y. For technical reasons we define argmax *to yield a set of the most probable labels so that C^* is well defined even in cases with multiple optimal labels.*

In this paper we want to compare different observations. That is why we will only consider the Bayesian classifier. Otherwise the comparison would not only depend on the observation itself but also on the strategy to use it. But this would unnecessarily complicate the theoretical analyses. Therefore, whenever we mention the classification accuracy we refer to the highest possible classification accuracy achieved by the Bayes classifier.

Definition 2 (**Classification accuracy of the Bayes classifier**). *The classification accuracy of the Bayes classifier using X to predict Y is*

$$acc(Y|X) = \sum_{x \in \Omega_X} P(X = x) \max_y P(Y = y|X = x)$$

A relevant concept in the comparison between classification accuracy and mutual information is the conditional accuracy. The conditional accuracy is defined by the probability of correct prediction for one particular observation $X = x$.

Definition 3 (**Conditional Accuracy**). *The conditional accuracy $acc(Y|X = x)$ is the classification accuracy of the Bayes classifier for a given observation $X = x$.*

$$acc(Y|X = x) = \max_y P(Y = y|X = x)$$

The conditional accuracies determine the classification accuracy. By definition the classification accuracy is a weighted mean of the conditional accuracies.

$$acc(Y|X) = \sum_{x \in \Omega_X} P(X = x) acc(Y|X = x)$$

We will shorten terms such as $acc(Y|X = x)$ to $acc(Y|x)$. We will consistently use upper case letters such as X for random variables and lower case letters such as x for realizations. The same goes for $P(X = x)$ which will be shortened to $P(x)$.

2.2 Mutual information

Mutual information is defined in the framework of information theory (IT). This framework models signal transmission. Consider a signal sent through a channel, e.g. a binary value -1 or 1 that is sent through a fibre optic cable. The channel may not convey the signal flawlessly but add noise N to the signal S. Then, the noisy signal, or response, $R = S + N$ is received at the other end of the channel. From now on, in order to stay within this framework, we will refer to observations X as responses R and to labels Y as signals S.

Other than for the classification accuracy, the perspective of IT is not exclusively concerned about the proportion of correctly predicted signals. In contrast, IT states that there is an amount of uncertainty about S which is reduced when the response R is observed. The amount of uncertainty that is reduced defines the mutual information. Thereby, the mutual information relies on the definition of uncertainty, or entropy, which was given by Shannon (1948, 2001).

Definition 4 (**Entropy**). *The entropy H of a discrete random variable S is (Shannon, 2001, p. 19):*

$$H(S) = \sum_{s \in \Omega_S} P(S = s) \log_2 \frac{1}{P(S = s)} \text{ [bit]}$$

where Ω_S is the finite support of the discrete random variable S. If $P(s) = 0$ then $P(s) \log_2 1/P(s) \equiv 0$ because $\lim_{p \to 0} p \log 1/p = 0$.

If the signal S is *binary* then the entropy of S has a particular shape denoted by the binary entropy function H_2. H_2 only depends on the probability $p = P(S = 1)$ because p sufficiently specifies the probability distribution of S.

Definition 5 (**Binary entropy function** H_2). *Let S be a binary random variable with $p = P(S = 1)$. Then, the entropy $H(S)$ can denoted as the binary entropy function $H_2(p)$:*

$$\begin{aligned} H(S) &= P(S = 1) \log_2 \frac{1}{P(S = 1)} + P(S = 0) \log_2 \frac{1}{P(S = 0)} \\ &= p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p} =: H_2(p) \end{aligned}$$

$H_2(p)$ is shown in figure 2.1. For example if $p = 0$ or $p = 1$ then there is no inherent uncertainty about S because it is always the same, $H_2(1) = 0bit$. The maximum uncertainty is reached for $p = 0.5$, $H_2(0.5) = 1bit$.

When the response R is known the entropy of S may be reduced. The remaining entropy of S given R can be expressed by the conditional entropy.

Definition 6 (**Conditional entropy**). *The conditional entropy H of S given R is (MacKay, 2003, p. 138):*

$$H(S|R) = \sum_{r \in \Omega_R} P(R = r) \left[\sum_{s \in \Omega_S} P(S = s|R = r) \log_2 \frac{1}{P(S = s|R = r)} \right] \text{ [bit]}$$

Additionally, the conditional entropy H of S given $R = r$ is:

$$H(S|R = r) = \sum_{s \in \Omega_S} P(S = s|R = r) \log_2 \frac{1}{P(S = s|R = r)} \text{ [bit]}$$

The mutual information is the difference between the entropy and the conditional entropy, which is the uncertainty about S minus the uncertainty about S after R is observed. This represents the amount of uncertainty that is reduced about S given R.

Definition 7 (**Mutual information**). *The mutual information between S and R is (MacKay, 2003, p. 139):*

$$I(S; R) = H(S) - H(S|R)$$

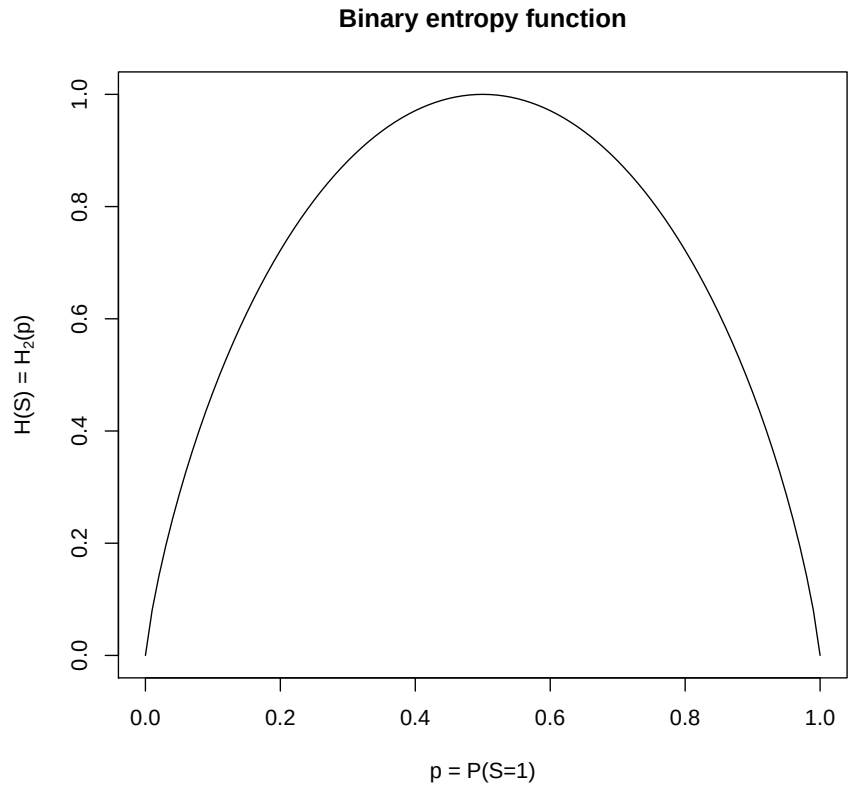


Figure 2.1: A binary signal S has an entropy equal to the binary entropy function H_2 which depends only on $p = P(S = 1)$.

We want to provide some intuitive examples. If $S = R$ then observing R reduces all uncertainty about S, $H(S|R) = 0$. Thus the conditional entropy of S given R is $I(S;R) = H(S) - H(S|R) = H(S) - 0 = H(S)$. In other words all the uncertainty about S is reduced after we observe R. In contrast, if S and R are independent, $S \perp\!\!\!\perp R$, then $P(S = s|R = r) = P(S = s)$ and the mutual information is $I(S;R) = H(S) - H(S|R) = H(S) - H(S) = 0bit$. In other words no uncertainty about S is reduced. $S = R$ and $S \perp\!\!\!\perp R$ are the two extreme cases and mutual information measures any intermediate case. The visual interpretation of the relation between entropy and mutual information in figure 2.2 was given by MacKay (2003, p. 140). From this visualization the symmetry of $I(S;R)$ becomes evident (MacKay, 2003, p. 143).

$$I(S;R) = H(S) - H(S|R) = I(R;S) = H(R) - H(R|S)$$

In contrast, the classification accuracy is not symmetric. Furthermore, the mutual information is non-negative, $I(S;R) \geq 0$ (MacKay, 2003, p. 143).

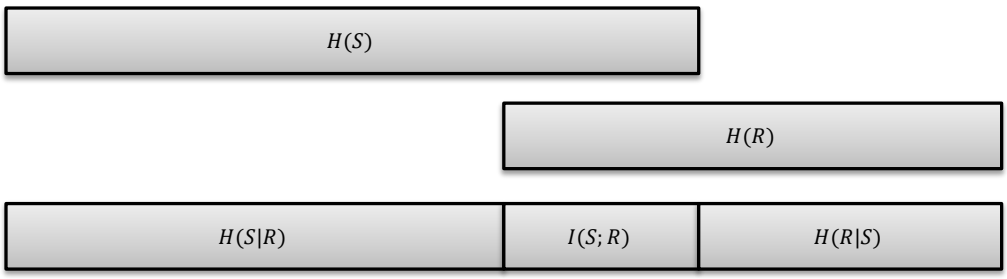


Figure 2.2: The relation between entropies $H(S)$, $H(R)$, conditional entropies $H(S|R)$, $H(R|S)$ and the mutual information $I(R;S)$

2.3 Assumptions

We will introduce four assumptions, two on the signal and two on the noise. These assumptions will be restrictive but we will relax them in section 3.4.

We will assume the signal S to be binary and equally weighted.

Assumption 1 (**binary signal**). $S \in \Omega_S = \{-1, 1\}$

Assumption 2 (**equal weights**). $P(S = s) = \frac{1}{|\Omega_S|}$

From assumption 1 follows for the conditional accuracies that $\forall r : 0.5 \leq acc(S|r)$ because the Bayes classifier predicts the signal with the largest probability. This implies $0.5 \leq acc(S|R)$ for the classification accuracy. Instead of writing assumption 2 as $P(S = s) = 0.5$, we have chosen this general form because we will later use it with non-binary signals, $|\Omega_S| > 2$.

We will further assume the noise N to be discrete and independent of signal S.

Assumption 3 (**discrete noise**). $N \in \Omega_N$ *and* Ω_N *is countable*

Assumption 4 (**independent noise**). $P(N|S) = P(N)$

With assumption 3 we restrict the setting to discrete random variables because there are conceptual difficulties with IT and continuous random variables. In principle, the definitions we introduced can be generalized replacing the sums by integrals. Then, we would replace the probability distributions by density distributions. But density distributions can have values greater than 1 so that the entropy can assume negative values and it is not clear how to interpret a negative uncertainty. Because of such difficulties we will first discuss the discrete case and then generalize to the continuous case in section 3.4.

Given assumptions 1-4 we are interested in a relation between the classification accuracies $acc(R|S)$ and $acc(S|R)$ and the mutual information $I(R;S)$.

2.4 Example

As an example for this setting consider S and N as given by assumptions 1-4 and the following probability distribution:

$$P(N = n) = \begin{cases} 0.25, & \text{if } n = -2 \\ 0.5, & \text{if } n = 0 \\ 0.25, & \text{if } n = 2. \end{cases}$$

Signal S and noise N will lead to a contingency table between S and $R = S+N$ as shown in table 2.1.

		R			
		-3	-1	1	3
S	-1	0.125	0.250	0.125	0.000
	1	0.000	0.125	0.250	0.125
$acc(S\|R = r)$		1	$0.66\bar{6}$	$0.66\bar{6}$	1
$1 - H(S\|R = r)$		1	0.082	0.082	1

Table 2.1: Contingency table between S and R. The classification accuracies are $acc(S|R) = 0.75$ and $acc(R|S) = 0.5$. The conditional accuracies are $acc(S|R = -3) = acc(S|R = 3) = 1$ and $acc(S|R = -1) = acc(S|R = 1) = 2/3$. The mutual information is $I(R; S) \approx 0.311$.

The classification accuracies of the Bayes classifiers are $acc(S|R) = 0.75$ when predicting S by R and $acc(R|S) = 0.5$ when predicting R by S (non-symmetric).

The conditional accuracies are $acc(S|R = -3) = acc(S|R = 3) = 1$ and $acc(S|R = -1) = acc(S|R = 1) = 2/3$. In other words, with an informative response $R \in \{-3, 3\}$ the signal S can be predicted with perfect accuracy. On the other hand with a less-informative response $R \in \{-1, 1\}$ the conditional accuracy is only at $66.\bar{6}\%$.

The mutual information is $I(R; S) = I(S; R) = 0.311$ (symmetric). The difference between the informative responses $R \in \{-3, 3\}$ and the less-informative responses $R \in \{-1, 1\}$ can also be expressed in terms of conditional entropy. At first, the uncertainty is $H(S) = 1bit$. An informative response reduces all uncertainty, $H(S|R \in \{-3, 3\}) = 0bit$ whereas a less-informative response reduces the uncertainty only by a small margin, $H(S|R \in \{-1, 1\}) = 918bit$.

Chapter 3

Main results

Our main results concern the question: What is the relation between classification accuracy and mutual information? The two measures are only loosely interrelated because they capture different properties of the classification task. Their relation is mediated through conditional accuracies and we derive lower and upper bounds for the mutual information given the classification accuracy, and vice versa. (section 3.1).

Additionally, we ask: What happens when responses are further processed by aggregations? Our result is that aggregating responses may decrease the mutual information even when the classification accuracy is unchanged. This means that an observation can loose desirable properties without changes in the classification accuracy (section 3.2).

The proofs can be found in the subsequent section 3.3. We implemented R scripts to represent our results and experiment with them (R Core Team, 2016). The scripts are made publicly accessible via Open Science Framework and can be found at `http://osf.io/zru7b/`.

3.1 Bounds

One would assume a tight relation between the accuracy of predicting a signal S with an observation R (classification accuracy $acc(S|R)$) and the reduction in uncertainty about S given R (mutual information $I(S;R)$). However, the relationship between these two measures is non-trivial and, in fact, loose (Wang and Hu, 2009).

Proposition 1 (**Mutual information is not a function of classification accuracies**). *Given assumptions 1-4 there is no function f so that $I(S;R) = f(acc(R|S), acc(S|R))$.*

For example table 3.1 shows two contingency tables with the same classification accuracies but different mutual information. Thus, there is no tight relation between the mutual information and classification accuracies. However, there is a tight relation between the mutual information and the conditional accuracies.

		R_1					
		-5	-3	-1	1	3	5
S	1	0.050	0.125	0.150	0.125	0.050	0.000
	1	0.000	0.050	0.125	0.150	0.125	0.050

(a)

		R_2					
		-5	-3	-1	1	3	5
S	1	0.075	0.100	0.150	0.100	0.075	0.000
	1	0.000	0.075	0.100	0.150	0.100	0.075

(b)

Table 3.1: Contingency table between S and R_1 in (a) and between S and R_2 in (b). The classification accuracies are equal, $acc(R_1|S) = acc(R_2|S) = .3$ and $acc(S|R_1) = acc(S|R_2) = 0.65$, but the mutual information differs, $I(S; R_1) \approx 0.151 \neq I(S; R_2) = .17$.

Proposition 2 (**Mutual information is a function of conditional accuracies**)**.** *Let assumptions 1 - 4 be true. Then, the mutual information is the weighted mean of a function of the conditional accuracies* $acc(S|r)$*.*

$$\begin{aligned} I(S;R) &= \sum_{r\in\Omega_R} P_R(r)\Big(H(S) - H(S|r)\Big) \\ &= \sum_{r\in\Omega_R} P_R(r)\Big(1 - H_2(acc(S|r))\Big)[bit] \end{aligned}$$

Following this proposition the mutual information corresponds to a weighted mean of a function of the conditional accuracies. In contrast, the classification accuracy is also a weighted mean but of the conditional accuracies themselves. Consider the example from section 2.4 again in table 3.2. This proposition states that the mutual information can be calculated in the following way, whereas the classification accuracy can be calculated in a similar manner:

$$\begin{aligned} I(S;R) &= 0.125\cdot 1 + 0.375\cdot 0.082 + 0.375\cdot 0.082 + 0.125\cdot 1 && \approx 0.311 bit \\ acc(S|R) &= 0.125\cdot 1 + 0.375\cdot 0.667 + 0.375\cdot 0.667 + 0.125\cdot 1 && = .75 \end{aligned}$$

This perspective on the relation between classification accuracy and mutual information is illustrated in the schematic in figure 3.1.

		R			
		-3	-1	1	3
S	-1	0.125	0.250	0.125	0.000
	1	0.000	0.125	0.250	0.125
$1 - H_2(acc(S\|r))$		1	0.082	0.082	1
$acc(S\|r)$		1	0.667	0.667	1
$P_R(r)$		.125	0.375	0.375	0.125

Table 3.2: Contingency table between S and R. The informative responses $R \in \{-3, 3\}$ have conditional accuracies of $acc(S|R = -3) = acc(S|R = 3) = 1$, reduce $1bit$ of uncertainty about S and appear with probability $P_R(r) = 0.125$, each. The mutual information is a weighted mean with weightings $P_R(r)$ of a function $g(x) = 1 - H_2(x)$ of the conditional accuracies $acc(S|r)$.

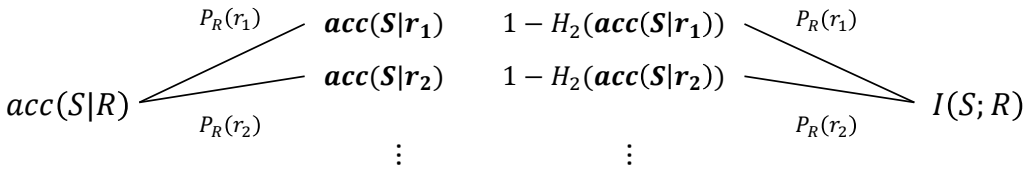


Figure 3.1: From proposition 2 follows that the measures of classification accuracy $acc(S|R)$ and mutual information $I(S;R)$ are the weighted mean of two different functions on the conditional accuracy. The classification accuracy uses the identity function $acc(S|r)$ and the mutual information uses the reduction of uncertainty $1 - H_2(acc(S|r))$ per response.

Up to now we have shown that the mutual information is not tightly related to classification accuracy but to the conditional accuracies. However, there is a loose relation between both measures through the conditional accuracies. This means that a fixed classification accuracy $acc(S|R)$ restricts the conditional accuracies which in turn restrict the range of the mutual information $I(S;R)$. As a result there are lower and upper bounds for $I(S;R)$ given $acc(S|R)$. These bounds have been reported before by Fano and Hawkins (1961) and Kovalevsky (1967).

Proposition 3 (**Bounds on the mutual information given the classification accuracy**). *Let assumptions 1 - 4 be true. For a given classification accuracy* $acc(S|R)$ *the mutual information* $I(S;R)$ *has lower and upper bounds:*

$$1 - H_2(acc(S|R)) \leq I(S;R) \leq 2acc(S|R) - 1$$

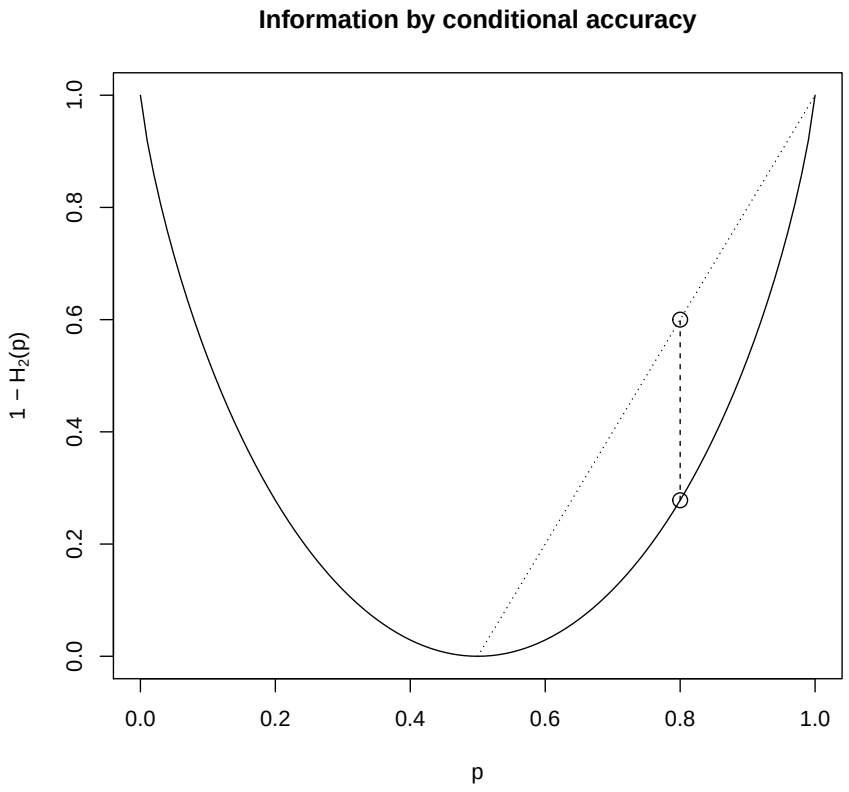


Figure 3.2: The so-called binary symmetric channel capacity $1 - H_2(p)$ is strictly convex because H_2 is strictly concave. Jensen's inequality for strictly convex functions states that means of function values are strictly larger than function values of means. For example if the function values (and weights) are 0.5 (0.4) and 1 (0.6) then the mean of function values (dashed line) is $0.4 \cdot [1 - H_2(0.5)] + 0.6 \cdot [1 - H_2(1)]$ compared to the function value of the mean (solid curve) $1 - H_2(0.4 \cdot 0.5 + 0.6 \cdot 1)$. The terms resolve to $0.4 \cdot 0 + 0.6 \cdot 1 = 0.6$ and $1 - H_2(0.8) \approx 0.278$.

The proof for this proposition relies on Jensen's inequality (Jensen, 1906) as visualized in figure 3.2. In short, the argument of our proof is that $g(acc(S|r)) = 1 - H_2(acc(S|r))$ is a strictly convex function for which Jensen's inequality implies that the lower bound is reached when all conditional accuracies are on average, $\forall r : acc(S|r) = acc(S|R)$. The upper bound occurs when all conditional accuracies are either 0.5 or 1, $\forall r : acc(S|R = r) \in \{0.5, 1\}$.

Conversely, it immediately follows that there are bounds on the classification accuracy for a given mutual information.

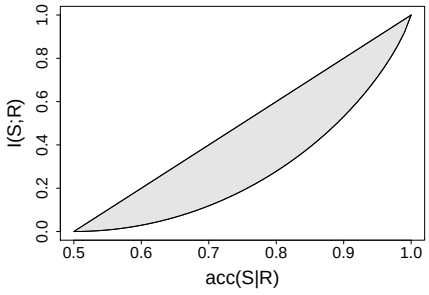
Proposition 4 (**Bounds on the classification accuracy given the mutual information in the binary case**). *Let assumptions 1 - 4 be true. For a given mutual information $I(S;R)$ the classification accuracy $acc(S|R)$ has lower and upper bounds:*

$$\frac{I(S;R)+1}{2} \leq acc(S|R) \leq H_2^{-1}(1-I(S;R))$$

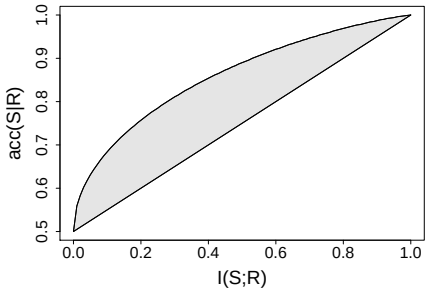
With $H_2^{-1}(x)$ being the inverse of $H_2(x)$ given that $x \geq 0.5$ so that $H_2^{-1}(x) \geq 0.5$ as well.

The last two propositions state that there are information theoretic worst and best cases for a given classification accuracy as well as minimal and maximal classification accuracy for a given mutual information. The bounds given by these two proofs are shown in figure 3.3. For example consider a classification accuracy of $acc(S|R) = 0.8$. For that classification accuracy the mutual information is at least $I(S;R) = 0.278bit$ and at best $I(S;R) = 0.6bit$. The worst case occurs when all conditional accuracies are $\forall r \in \Omega_R : acc(S|r) = 0.8$. In the best case 40% of the conditional accuracies are 0.5 and 60% of them are 1. These are the examples from the introduction.

Notice that the examples from the introduction do not fulfill assumption 4 (independent noise). In section 3.4.4 we will show that this assumption is not necessary to our results and that the bounds are tight for both, the independent and the dependent case.



(a) Lower and upper bounds on the mutual information $I(S;R)$ given the classification accuracy $acc(S|R)$ in the binary case.



(b) Lower and upper bounds on the classification accuracy $acc(S|R)$ given the mutual information $I(S;R)$ in the binary case.

Figure 3.3: The bounded relationship between $acc(S|R)$ and $I(S;R)$ according to propositions 3 and 4.

Up to now we have only considered binary signals and the previous literature has already provided these bounds. Our main contribution is the generalization to

the case with M signal values. We chose to provide the proposition and proof for the binary case because we have shaped them in a way that immediately allows for the generalization. Thus, the proof for the non-binary case has the same structure as the proof for the binary case. That is why we considered it a good starting point. Our generalization yields the following bounds for the non-binary case with M signal values.

Assumption 5 (M **signal values**). $S \in \{1, 2, ..., M\}$

M signal values implies that the conditional accuracies are larger than $1/M$, $\forall r : 1/M \leq acc(S|r)$ because the Bayes classifier predicts the signal with the largest probability. This implies $1/M \leq acc(S|R)$.

Proposition 5 (**Bounds on the mutual information given the classification accuracy**). *Let assumptions 5 (M signal values) and 2 - 4 be true. For a given classification accuracy $acc(S|R)$ the mutual information $I(S;R)$ has lower and upper bounds:*

$$\log(M) - H_2(acc(S|R)) - (1 - acc(S|R)) \log(M-1) \leq I(S;R)$$

$$I(S;R) \leq \log(M) \left(1 - M \left(\frac{1 - acc(R|S)}{M-1}\right)\right)$$

((The last line has changed on January 6, 2022 after I was made away that there was a mistake in the old version, which read: $I(S;R) \leq \frac{M}{M-1}(acc(S|R)-1)$. Thank you, CR!))

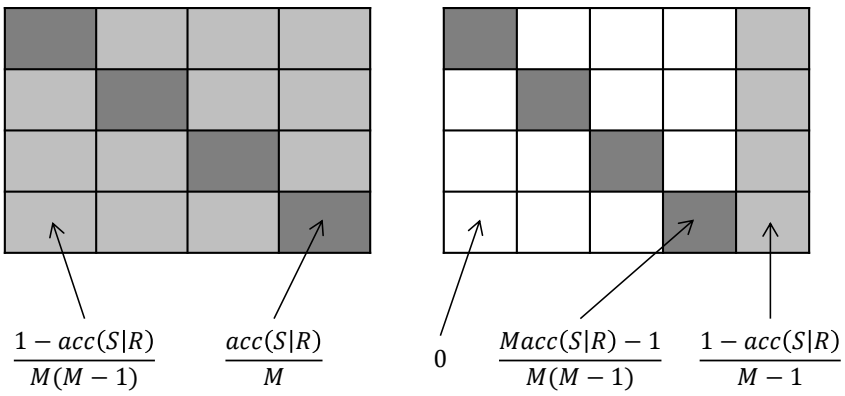


Figure 3.4: The left table sketches a contingency table with minimal mutual information, the right table sketches the maximal mutual information. The minimal mutual information occurs when all responses are on average. In contrast, the maximal mutual information is obtained when there are only perfectly accurate and non-informative responses. The pattern is similar to the binary case.

The lowest possible mutual information occurs when all conditional accuracies are on average, $\forall r : acc(S|R = r) = acc(S|R)$. The highest possible mutual information occurs when all conditional accuracies are either $1/M$ or 1, $\forall r : acc(S|R = r) \in \{1/M, 1\}$. This is the same principle as in the binary case with $M = 2$. Figure 3.4 shows a schematic example of the shape of the minimal vs. maximal mutual information observations with $M = 4$. The bounds for different signal support sizes M are illustrated in figure 3.5. We refused on showing the inverted bounds for the non-binary case because they can be obtained by inverting our result and because they are not relevant here.

From these propositions and proofs we understand that the mutual information captures multiple aspects of the response next to the classification accuracy, which is only one of these aspects. In the binary case the mutual information also captures the variation of conditional accuracies as another aspect. If all conditional accuracies are on average (no variation) then there is minimal mutual information. If all conditional accuracies are 0.5 or 1 (maximal variation) then the mutual information is maximal. The variation of conditional accuracies is *aspect 1*. In chapter 5 we will show how this aspects become important in practical examples.

Similarly, in the non-binary case the mutual information captures a second aspect. This aspect is revealed when we had to optimize the distribution over the remaining response values in the proof of proposition 5. For example, consider a response r with a marginal probability of $P_R(r) = 0.1$. In the binary case a conditional accuracy of $acc(S|R = r) = 0.8$ determines the probability distribution, $P(s_1, r) = 0.08$ and $P(s_2, r) = 0.02$. In the non-binary case with $M = 3$ the probability $P(s_1, r) = 0.08$ is still determined. But the probabilities for the remaining signal values s_2 and s_3 are not restricted. They could be $P(s_2, r_1) = 0.01$ and $P(s_3, r_1) = 0.01$ (minimal mutual information); or they could be $P(s_2, r_2) = 0.02$ and $P(s_3, r_2) = 0$ (maximal mutual information). If we observe r_2 and our initial prediction s_1 is wrong then we at least know, that the true signal is s_2. With r_1 there is no such guarantee. Thus, r_2 conveys more information about the signal compared to r_1. This property, the entropy over the remaining response values, is *aspect 2*. Aspect 2 may be important for example in medical settings where a test outcome (response) indicates different diseases (signal). If the first diagnose was wrong then it is relevant what other diseases are likely.

Because the mutual information additionally captures these two aspects it can complement the classification accuracy. We encourage researchers to think about how relevant these aspects are in their field and to take the mutual information into account accordingly.

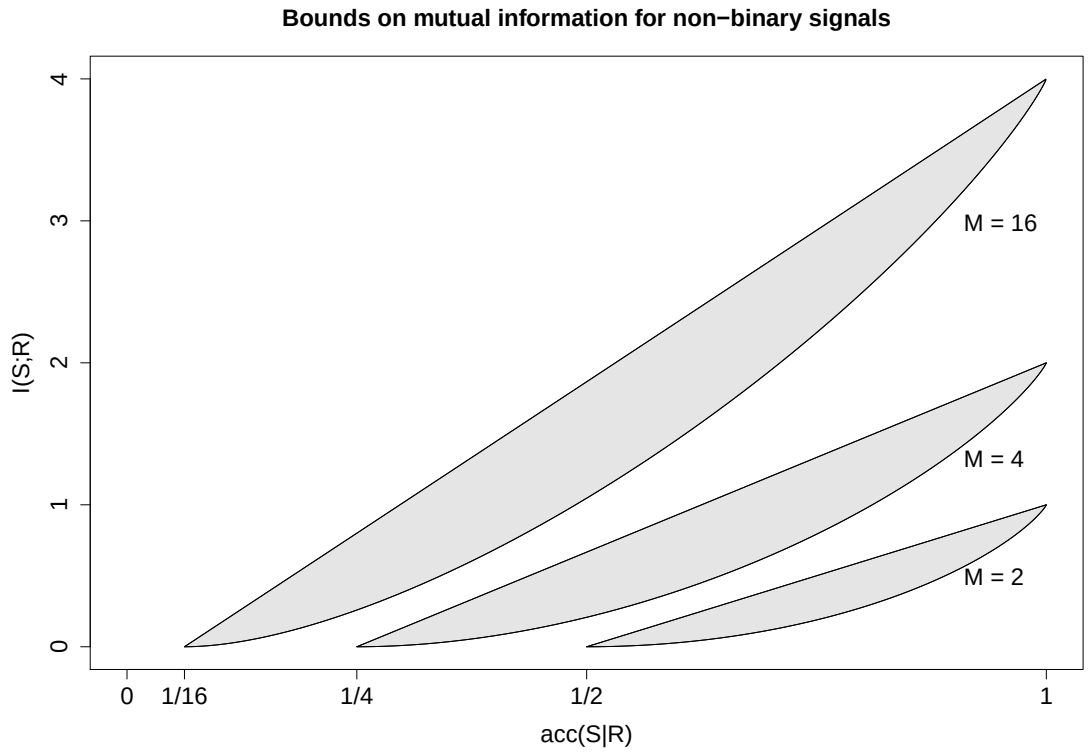


Figure 3.5: The bounds on $I(S;R)$ given $acc(S|R)$ according to proposition 5 for different signal support sizes M.

3.2 Aggregation of responses

Sometimes it is convenient to further process the response. For example, we may want to simplify it by binning, or aggregating, all those response values that predict the same signal value. But such aggregations may reduce the mutual information so that we end up with a worse observation even though the classification accuracy stays unchanged. In this section we will define aggregations as a way to process the response and show when they decrease the mutual information.

First, we define simple aggregation functions that aggregate two response values.

Definition 8 (**Simple aggregation function**). *A simple aggregation function f is the identical function on R except for two responses r_1 and r_2, $\forall r \in \Omega_R \backslash \{r_1, r_2\}$: $f(r) = r$. The two responses are mapped – aggregated – onto the same new value $r' \notin \Omega_R$, $f(r_1) = f(r_2) = r'$.*

Now, we extend the definition from simple aggregation functions to (complex) aggregation functions. An aggregation function is a possibly infinite composition of simple aggregation functions.

Definition 9 (**Aggregation function**). *An aggregation function f is a composition of (possibly infinite) simple aggregation functions $h_1, h_2, ...$*

$$f = h_1 \circ h_2 \circ ...$$
$$f(R) = h_1(h_2(...R...))$$

The properties of simple aggregation functions translate to aggregation functions by induction. Therefore, we will only show when simple aggregation functions decrease the classification accuracy vs. the mutual information. In general, an aggregation can only keep the mutual information constant or decrease it as given by the data processing inequality (Cover and Thomas, 2006, p. 34). But when does the mutual information strictly decrease? And when does the classification accuracy change?

Proposition 6 (**Aggregating responses decreases classification accuracy and mutual information**). *Let assumptions 5 (M signal values) and 2 - 4 be true. Consider a simple aggregation function f that aggregates r_1 and r_2.*

(a) If $C^*(R = r_1) \cap C^*(R = r_2) \neq \varnothing$ then $acc(S|R) = acc(S|f(R))$.

(b) If $C^*(R = r_1) \cap C^*(R = r_2) = \varnothing$ then $acc(S|R) > acc(S|f(R))$.

(c) If $P(S|R = r_1) = P(S|R = r_2)$ then $I(S;R) = I(S;f(R))$.

(d) If $P(S|R = r_1) \neq P(S|R = r_2)$ then $I(S;R) > I(S;f(R))$.

An aggregation keeps the classification accuracy constant if the aggregated response values indicate the same signal value (a), otherwise the classification accuracy strictly decreases (b). Of course, such an aggregation is always to avoid. On the other hand, an aggregation keeps the mutual information constant if and only if the aggregated responses have identical conditional probabilities (c), otherwise the mutual information strictly decreases (d). See table 3.3 for an illustrative example with $M = 2$.

This means that if we treat responses with different conditional accuracies as equal (aggregate them) then we decrease the mutual information. This happens even if they indicate the same signal value. We loose the desirable properties which we have called aspect 1 and 2. But as we will show in chapters 4 and 5 these properties indicated by mutual information are sometimes important and they will be lost upon such aggregations. In conclusion, when responses are processed further we recommend to pay close attention to the mutual information.

		$f_a(R)$		
		(-2,-1)	1	2
S	-1	0.30	0.20	0.00
	1	0.15	0.25	0.10

(a) $acc(S|R) = 0.65,\ I(S;R) \approx 0.141bit$

		$f_b(R)$		
		-2	(-1,1)	2
S	-1	0.10	0.40	0.00
	1	0.05	0.35	0.10

(b) $acc(S|R) = 0.6,\ I(S;R) \approx 0.115bit$

		$f_c(R)$		
		(-2,-1)	1	2
S	-1	0.30	0.20	0.00
	1	0.15	0.25	0.10

(c) $acc(S|R) = 0.65,\ I(S;R) \approx 0.141bit$

		$f_d(R)$		
		-2	-1	(1,2)
S	-1	0.10	0.20	0.20
	1	0.05	0.10	0.35

(d) $acc(S|R) = 0.65,\ I(S;R) \approx 0.067bit$

		R			
		-2	-1	1	2
S	-1	0.10	0.20	0.20	0.00
	1	0.05	0.10	0.25	0.10

(e) $acc(S|R) = 0.65,\quad I(S;R) \approx 0.141bit$

Table 3.3: Contingency tables (a) - (d) are aggregations of (e) exemplifying cases (a) - (d) from proposition 6. Starting with contingency table (e) the classification accuracy is $acc(S|R) = 0.65$ and the mutual information is $I(S;R) = 0.141bit$. (a) If responses -2 and -1 are aggregated the classification accuracy stays unchanged, $acc(S|f_c(R)) = acc(S|R) = 0.65$, because both responses indicate the same signal, e.g. $C^*(-2) \cap C^*(-1) = \{-1\}$. But (b) if responses -1 and 1 are aggregated the accuracy decreases, $acc(S|f_b(R)) = 0.6$. In (c) the same aggregation as (a) is used and this does not decrease the mutual information $I(S;f_c(R)) = 0.141bit$ because the conditional probabilities are equal (a ratio of 2:1 for both responses). (d) Aggregating responses 1 and 2 does decrease the mutual information to $I(S;f_d(R)) \approx 0.067$ while the classification accuracy stays unchanged. This is because highly informative cases $R = 2$ and less-informative cases $R = 1$ are mixed.

3.3 Proofs

In this section we present the proofs for the propositions from the previous sections.

Proof of proposition 1 (Mutual information is not a function of classification accuracies)

We make a proof by contradiction. Assume that the opposite of this proposition was true, that there is a function f with $I(S;R) = f(acc(R|S), acc(S|R))$. If this is the case then every signal-response pair with the same classification accuracies must have the same mutual information.

$$acc(R_1|S_1) = acc(R_2|S_2) \wedge acc(S_1|R_1) = acc(S_2|R_2) \implies I(R_1;S_1) = I(R_2;S_2)$$

Now consider following noise distributions, N_1 and N_2, with the same support $\Omega_{N_1} = \Omega_{N_2} = \{-4, -2, 0, 2, 4\}$ as a counter example.

$$P(N_1 = n) = \begin{cases} 0.1, & \text{if } n = -4 \\ 0.25, & \text{if } n = -2 \\ 0.3, & \text{if } n = 0 \\ 0.25, & \text{if } n = 2 \\ 0.1, & \text{if } n = 4 \end{cases} \qquad P(N_2 = n) = \begin{cases} 0.15, & \text{if } n = -4 \\ 0.2, & \text{if } n = -2 \\ 0.3, & \text{if } n = 0 \\ 0.2, & \text{if } n = 2 \\ 0.15, & \text{if } n = 4 \end{cases}$$

Noise N_1 and N_2 will lead to the responses $R_1 = S + N_1$ and $R_2 = S + N_2$ as shown in table 3.1. Remember, that the noise S is fully specified by the assumptions. The accuracies of both examples are equal, $acc(R_1|S) = acc(R_2|S) = .3$ and $acc(S|R_1) = acc(S|R_2) = 0.65$. But the mutual information is not equal, $I(R_1;S) \approx 0.151 \neq I(R_2;S) = .17$. This is the desired contradiction. ■

Proof of proposition 2 (Mutual information is a function of conditional accuracies)

The first step is to use the definition of the mutual information and to reformulate it to be the weighted sum of the reduction in uncertainty per response. We denote marginal probabilities as $P_S(s) = P(S = s)$ and $P_R(r) = P(R = r)$.

$$\begin{aligned} I(S;R) &= H(S) - H(S|R) \\ &= \sum_{s \in \Omega_S} P_S(s) \log \frac{1}{P_S(s)} - \sum_{r \in \Omega_R} P_R(r) \sum_{s \in \Omega_S} P(s|r) \log \frac{1}{P(s|r)} \\ &= \sum_{r \in \Omega_R} P_R(r) \Big[\sum_{s \in \Omega_S} P_S(s) \log \frac{1}{P_S(s)} - \sum_{s \in \Omega_S} P(s|r) \log \frac{1}{P(s|r)} \Big] \end{aligned}$$

$$= \sum_{r \in \Omega_R} P_R(r) \Big[H(S) - H(S|r) \Big]$$

This is the general case. Additionally, the signal S is binary and equally weighted by assumption. Thus, $H(S) = H_2(0.5) = 1bit$ and every response can be considered as a binary channel with an accuracy equal to the conditional accuracy of that response, $acc(S|r)$. Therefore, the conditional entropy can be replaced by the binary entropy function, $H_2(acc(S|r))$.

$$\begin{aligned}
I(S;R) &= \sum_{r \in \Omega_R} P_R(r) \Big[1bit - \sum_{s \in \Omega_S} P(s|r) \log \frac{1}{P(s|r)} \Big] \\
&= \sum_{r \in \Omega_R} P_R(r) \Big[1bit - P(S=-1|r) \log \frac{1}{P(S=-1|r)} \\
&\qquad\qquad - P(S=1|r) \log \frac{1}{P(S=1|r)} \Big] \\
&= \sum_{r \in \Omega_R} P_R(r) \Big[1bit - acc(S|r) \log \frac{1}{acc(S|r)} \\
&\qquad\qquad - (1 - acc(S|r)) \log \frac{1}{(1 - acc(S|r))} \Big] \\
&= \sum_{r \in \Omega_R} P_R(r) \Big[1bit - H_2(acc(S|r)) \Big]
\end{aligned}$$

■

Proof of proposition 3 (Bounds on the mutual information given the classification accuracy)

The lowest and highest mutual information for a given classification accuracy are expressed by two optimization problems corresponding to the lower and upper bounds:

$$\begin{aligned}
&(1) \min_{S,R} I(S;R) \text{ s.t. } acc(S|R) \text{ const.} \\
&(2) \max_{S,R} I(S;R) \text{ s.t. } acc(S|R) \text{ const.}
\end{aligned}$$

In order to solve (1) and (2) it is important to note that the classification accuracy is a weighted mean of the conditional accuracies:

$$acc(S|R) = \sum_{r \in \Omega_R} P_R(r) \max_s P(s|r) = \sum_{r \in \Omega_R} P_R(r) acc(S|r)$$

(1) For the mutual information proposition 2 states that it is a weighted mean of the reduction of uncertainty per response:

$$I(S;R) = \sum_{r \in \Omega_R} P_R(r) \Big(1 - H_2(acc(S|r)) \Big)$$

We denote the conditional accuracies $x_i = acc(S|r_i)$, the marginal probabilities $w_i = P(r_i)$ and the reduction in uncertainty as $g(x) = 1 - H_2(x)$. Then, the first optimization problem is:

$$\begin{aligned} &\min_{x_1, w_1, x_2, w_2, \ldots} \sum_i w_i g(x_i) \\ &\text{s.t. } \sum_i w_i x_i \text{ const.}, \sum_i w_i = 1,\ 0 \leq w_i, \leq 1 \text{ and } 0.5 \leq x_i \leq 1 \end{aligned}$$

$H_2(x)$ is a strictly concave function and thus $g(x) = 1 - H_2(x)$ is strictly convex. Therefore, we can apply Jensen's inequality in its general form:

$$g\left(\sum_i w_i x_i \right) \leq \sum_i w_i g(x_i)$$

By resubstitution of $\sum_i w_i x_i = acc(S|R)$ we get the minimal goal value as $g\left(\sum_i w_i x_i\right) = 1 - H_2(acc(S|R)) \leq I(S;R)$. According to Jensen's inequality, the equality holds if and only if x_i is constant. Since x_i denotes the conditional accuracies, this means that the lower bound for the mutual information is reached, when all conditional accuracies are equal. Then, the conditional accuracies are equal to the classification accuracy because it is their weighted mean, $\forall r_i \in \Omega_R : acc(S|r_i) = acc(S|R)$.

(2) We denote the optimization problem as in (1).

$$\begin{aligned} &\max_{x_1, w_1, x_2, w_2, \ldots} \sum_i w_i g(x_i) \\ &\text{s.t. } \sum_i w_i x_i \text{ const.}, \sum_i w_i = 1,\ 0 \leq w_i, \leq 1 \text{ and } 0.5 \leq x_i \leq 1 \end{aligned}$$

We will show that the mutual information is only maximized when the conditional accuracies of all response values are either 0.5 or 1. If there was a

conditional accuracy $0.5 < x_i = acc(S|r_i) < 1$ with $w_i = P_R(r_i)$ then we could replace x_i and w_i by:

$$x_i \to \begin{cases} x_i^{'} = 0.5 \\ x_i^{''} = 1 \end{cases} \qquad \text{and} \qquad w_i \to \begin{cases} w_i^{'} = 2w_i(1 - x_i) \\ w_i^{''} = 2w_i(x_i - 0.5) \end{cases}$$

The substitution will not violate the constraints because if the constraints did hold before, they will do so after the substitution as well. The weights add up and can not be negative because $x_i > 0.5$: $w_i = w_i^{'} + w_i^{''}$ and $0 \leq w_i^{'}, w_i^{''} \leq 1$. The conditional accuracies obviously satisfy the constraint that $0.5 \leq x_i^{'}, x_i^{''} \leq 1$. The classification accuracy stays constant:

$$\begin{aligned} w_i^{'}x_i^{'} + w_i^{''}x_i^{''} &= 2w_i(1 - x_i) \cdot 0.5 + 2w_i(x_i - 0.5) \cdot 1 \\ &= w_i - w_i x_i + 2w_i x_i - w_i \\ &= w_i x_i \end{aligned}$$

Thus, the constraints are not violated. But the goal function is strictly larger as stated by Jensen's inequality for strictly convex functions. As a result all conditional accuracies $acc(S|r_i) = x_i$ must be 0.5 or 1. They can not be further apart because conditional accuracies must be within $[0.5; 1]$ for binary S.

Now, if all response values have a conditional accuracy of 0.5 or 1 then there is only one ratio between these conditional accuracies that satisfies the constraint that $acc(S|R) = \sum_i w_i x_i$ is constant. Responses with conditional accuracy $acc(S|r_i) = x_i = 0.5$ must appear with probability $2(1 - acc(S|R))$ and responses with $acc(S|r_i) = x_i = 1$ must appear with probability $2(acc(S|R) - 0.5)$ so that the classification accuracy $acc(S|R)$ is constant:

$$\begin{aligned} \sum_i w_i x_i &= \sum_{i \text{ with } x_i = 0.5} w_i \cdot 0.5 + \sum_{i \text{ with } x_i = 1} w_i \cdot 1 \\ &= 2(1 - acc(S|R)) \cdot 0.5 + 2(acc(S|R) - 0.5) \cdot 1 \\ &= 1 - acc(S|R) + 2acc(S|R) - 1 \\ &= acc(S|R) \end{aligned}$$

Knowing the conditional accuracies, 0.5 and 1, and their respective weightings, $2(1 - acc(S|R))$ and $2(acc(S|R) - 0.5)$, we can now apply proposition 2 to calculate the upper bound on the mutual information:

$$I(S; R) = \sum_{r \in \Omega_R} P_R(r) \Big(1 - H_2(acc(S|r)) \Big)$$

$$
\begin{aligned}
&\leq \sum_{r \in \Omega_R \text{ with } acc(S|r)=0.5} P_R(r)\Big(1 - H_2(0.5))\Big) \\
&\quad + \sum_{r \in \Omega_R \text{ with } acc(S|r)=1} P_R(r)\Big(1 - H_2(1)\Big) \\
&= \sum_{r \in \Omega_R \text{ with } acc(S|r)=0.5} P_R(r)\Big(1 - 1\Big) \\
&\quad + \sum_{r \in \Omega_R \text{ with } acc(S|r)=1} P_R(r)\Big(1 - 0\Big) \\
&= \sum_{r \in \Omega_R \text{ with } acc(S|r)=1} P_R(r) \\
&= 2(acc(S|R) - 0.5) \\
&= 2acc(S|R) - 1
\end{aligned}
$$

■

Proof of proposition 4 (Bounds on the classification accuracy given the mutual information in the binary case)

Given proposition 3 the bounds are simply inverted. The upper bound on $I(S;R)$ given $acc(S|R)$ becomes the lower bound on $acc(S|R)$ given $I(S;R)$.

$$
\begin{aligned}
I(S;R) &\leq 2acc(S|R) - 1 \\
\frac{I(S;R)+1}{2} &\leq acc(S|R)
\end{aligned}
$$

Consistently, the lower bound on $I(S;R)$ becomes the upper bound on $acc(S|R)$.

$$
\begin{aligned}
1 - H_2(acc(S|R)) &\leq I(S;R) \\
1 - I(S;R) &\leq H_2(acc(S|R)) \\
H_2^{-1}(1 - I(S;R)) &\geq acc(S|R)
\end{aligned}
$$

The inequality sign switches because we take the inverse of $H_2(acc(S|R))$ which is strictly increasing in the range of $0.5 \leq acc(S|R) \leq 1$. Hence, H_2^{-1} is strictly decreasing and applying this function inverts the inequality sign. ■

Proof of proposition 5 (Bounds on the mutual information given the classification accuracy)

Similar to proposition 3 we have two optimization problems and the proofs follow the same structure.

$$(1)\ \min_{S,R} I(S;R) \text{ s.t. } acc(S|R) \text{ const.}$$
$$(2)\ \max_{S,R} I(S;R) \text{ s.t. } acc(S|R) \text{ const.}$$

(1) From proposition 2 we know a suitable form of the mutual information:

$$I(S;R) = \sum_{r\in\Omega_R} P_R(r)\Big(1 - H(S|r)\Big)$$

But unlike in proposition 3 we can not use the binary entropy function $H_2(S|r)$ because the signal is not binary. Instead we have to derive the general form of the optimal $H(S|r)$. First, note that $H(S|r) = \sum_s P(s|r)\log(1/P(s|r))$ is strictly concave because $x\log(1/x)$ is strictly concave (MacKay, 2003, p. 35). Now, because we want to minimize $I(S;R)$, we want to maximize $H(S|r)$. At the same time, the constraint is $acc(S|R) = \sum_r P_R(r)acc(S|r)$. Thus, we want to maximize $H(S|r)$ subject to constant $acc(S|r)$. With a constant $acc(S|r)$ there must be one signal value s^* for which the conditional probability is $P(s^*|r) = acc(S|r)$. Then, the maximal entropy $H(S|r)$ is reached when all other $M-1$ signal values have a uniform distribution because of Jensen's inequality for strictly concave functions. As a result, the maximal conditional entropy given a conditional accuracy $acc(S|r)$ is:

$$\begin{aligned}
H(S|r) &\leq acc(S|r)\log\left(\frac{1}{acc(S|r)}\right) \\
&\quad + (M-1)\frac{(1-acc(S|r))}{M-1}\log\left(\frac{M-1}{(1-acc(S|r))}\right) \\
&= acc(S|r)\log\left(\frac{1}{P_R(r)acc(S|r)}\right) + (1-acc(S|r))\log\left(\frac{1}{(1-acc(S|r))}\right) \\
&\quad + (1-acc(S|r))\log(M-1) \\
&= H_2(acc(S|r)) + (1-acc(S|r))\log(M-1)
\end{aligned}$$

We denote the conditional accuracies $x_i = acc(S|r_i)$, the marginal probabilities $w_i = P(r_i)$ and the reduction of uncertainty as
$g(x) = \log(M) - [H_2(x) + (1-x)\log(M-1)]$. With this we are back to

what we have done in the proof of proposition 3. The optimization problem now is:

$$\min_{x_1,w_1,x_2,w_2,\ldots} \sum_i w_i g(x_i)$$
$$\text{s.t. } \sum_i w_i x_i \text{ const.}, \sum_i w_i = 1,\ 0 \leq w_i, \leq 1 \text{ and } 1/M \leq x_i \leq 1$$

Now, $g(x)$ is strictly convex because $H(S)$ is constant and $H(S|r)$ is, as we noted, strictly concave. We can apply Jensen's inequality again and get that all conditional accuracies must be equal, $\forall r_i \in \Omega_R : acc(S|r_i) = acc(S|R)$. This leads to the minimal goal value.

$$g\left(\sum_i w_i x_i\right) \leq I(S;R)$$
$$\log(M) - H_2(acc(S|R)) - (1 - acc(S|R))\log(M-1) \leq I(S;R)$$

Note, that we do not run into a local optimum by choosing the distribution of conditional accuracies based on our choice of $H(S|r)$. As we have shown, whatever the form of $H(S|r)$ it must be strictly concave. Therefore, our choice of the conditional accuracies is optimal and thus our (local) choice of $H(S|r)$ is also globally optimal.

(2) We denote the problem as in (1).

$$\max_{x_1,w_1,x_2,w_2,\ldots} \sum_i w_i g(x_i)$$
$$\text{s.t. } \sum_i w_i x_i \text{ const.}, \sum_i w_i = 1,\ 0 \leq w_i, \leq 1 \text{ and } 1/M \leq x_i \leq 1$$

We will show that the mutual information is only maximized when the conditional accuracies of all response values are either $1/M$ or 1. Because if there was a response r_i with conditional accuracy $1/M < x_i = acc(S|r_i) < 1$ with $w_i = P_R(r_i)$ then we could replace x_i and w_i by:

$$x_i \to \begin{cases} x_i^{'} = 1/M \\ x_i^{''} = 1 \end{cases} \qquad \text{and} \qquad w_i \to \begin{cases} w_i^{'} = w_i M \frac{1-x_i}{M-1} \\ w_i^{''} = w_i \frac{Mx_i - 1}{M-1} \end{cases}$$

The substitution will not violate the constraints, if they did hold before. The weights add up and can not be negative because $x_i > 1/M$, $w_i = w_i^{'} + w_i^{''}$ and

$0 \leq w_i^{'}, w_i^{''} \leq 1$. The conditional accuracies obviously satisfy the constraints, $1/M \leq x_i^{'}, x_i^{''}, \leq 1$. The classification accuracy stays constant:

$$\begin{aligned} w_i^{'}x_i^{'} + w_i^{''}x_i^{''} &= w_i M \frac{1-x_i}{M-1} \cdot \frac{1}{M} + w_i \frac{Mx_i - 1}{M-1} \cdot 1 \\ &= w_i \frac{1-x_i}{M-1} + w_i \frac{Mx_i - 1}{M-1} \\ &= w_i \frac{(M-1)x_i + 1 - 1}{M-1} \\ &= w_i x_i \end{aligned}$$

Thus, the constraints are not violated. But the goal function is strictly larger as stated by Jensen's inequality for strictly convex functions. As a result all conditional accuracies $acc(S|r_i) = x_i$ must be $1/M$ or 1. They can not be further apart because conditional accuracies in this setting must be within $[1/M; 1]$ for signals with M response values.

Now, if all response values have a conditional accuracy of $1/M$ or 1 then there is only one ratio between these conditional accuracies that satisfies the constraint that $acc(S|R) = \sum_i w_i x_i$ is constant. Responses with conditional accuracy $acc(S|r_i) = x_i = 1/M$ must appear with probability $M(1-acc(S|R))/(M-1)$ and responses with $acc(S|r_i) = x_i = 1$ with $(Macc(S|R)-1)/(M-1)$ so that the classification accuracy $acc(S|R)$ is constant:

$$\begin{aligned} \sum_i w_i x_i &= \sum_{i \text{ with } x_i = 1/M} w_i \cdot \frac{1}{M} + \sum_{i \text{ with } x_i = 1} w_i \cdot 1 \\ &= \frac{M(1-acc(S|R))}{M-1} \cdot \frac{1}{M} + \frac{Macc(S|R)-1}{M-1} \cdot 1 \\ &= \frac{(M-1)acc(S|R) + 1 - 1}{M-1} \\ &= acc(S|R) = \sum_i w_i x_i \end{aligned}$$

Knowing the conditional accuracies, $1/M$ and 1, and their respective weightings, $M(1-acc(S|R))/(M-1)$ and $(Macc(S|R)-1)/(M-1)$, we can now apply proposition 2 to calculate the maximal mutual information. In order to do so we will use that the conditional probabilities of a response with conditional accuracy $acc(S|r) = 1/M$ must be the uniform distribution and the conditional entropy is $H(S|r) = \log(M)$. Responses with a conditional accuracy of $acc(S|r) = 1$ have a conditional entropy of $H(S|r) = 0$.

$$
\begin{aligned}
I(S;R) &= \sum_{r\in\Omega_R} P_R(r)\Big(H(S) - H(S|r)\Big) \\
&\leq \sum_{r\in\Omega_R \text{ with } acc(S|r)=1/M} P_R(r)\Big(\log(M) - \log(M)\Big) \\
&\quad + \sum_{r\in\Omega_R \text{ with } acc(S|r)=1} P_R(r)\Big(\log(M) - 0\Big) \\
&= \log(M)\left(1 - M\left(\frac{1-acc(R|S)}{M-1}\right)\right)
\end{aligned}
$$

((The last line has changed on January 6, 2022 after I was made away that there was a mistake in the old version.))

■

Proof of proposition 6 (Aggregating responses decreases classification accuracy and mutual information)

We start the proof by observing three properties which we will use throughout the proofs of (a)-(d). First, the aggregation function f maps r_1 and r_2 but nothing else onto r'. Therefore the joint probabilities with respect r' are just the sum of joint probabilities with respect r_1 and r_2.

$$\forall s \in \Omega_S : P(s,r') = P(s,r_1) + P(s,r_2) \tag{3.1}$$

Using this property we can derive that the marginal probability of r' is the sum of marginal probabilities of r_1 and r_2:

$$
\begin{aligned}
P_{f(R)}(r') &= \sum_s P(s,r') = \sum_s [P(s,r_1) + P(s,r_2)] = \sum_s P(s,r_1) + \sum_s P(s,r_2) \\
&= P_R(r_1) + P_R(r_2)
\end{aligned} \tag{3.2}
$$

Additionally, we can show that the conditional probabilities given r' are a weighted mean of the two conditional probabilities given r_1 and r_2.

$$
\begin{aligned}
P(s,r') &= P(s,r_1) + P(s,r_2) \\
P_{f(R)}(r')P(s|r') &= P_R(r_1)P(s|r_1) + P_R(r_2)P(s|r_2) \\
P(s|r') &= \frac{P_R(r_1)}{P_{f(R)}(r')}P(s|r_1) + \frac{P_R(r_2)}{P_{f(R)}(r')}P(s|r_2)
\end{aligned} \tag{3.3}
$$

(a) We start by solving the equality between classification accuracies $acc(S|R)$ and $acc(S|f(R))$ which we want to prove. We use that the classification accuracy is a weighted mean of its conditional accuracies and the property in equation (3.1):

$$
\begin{aligned}
acc(S|R) &= acc(S|f(r)) \\
\sum_{r\in\Omega_R} P_R(r) \max_s P(s|r) &= \sum_{r\in\Omega_{f(R)}} P_{f(R)}(r) \max_s P(s|r) \\
\sum_{r\in\Omega_R} \max_s P(s,r) &= \sum_{r\in\Omega_{f(R)}} \max_s P(s,r) \\
\max_s P(s,r_1) + \max_s P(s,r_2) &= \max_s P(s,r') \\
\max_s P(s,r_1) + \max_s P(s,r_2) &= \max_s \Big(P(s,r_1) + P(s,r_2) \Big)
\end{aligned}
$$

From the assumption $C^*(r_1) \cap C^*(r_2) \neq \varnothing$ follows that there is a signal value s^* which is predicted for both responses, $\exists s^* : s^* \in C^*(r_1) \cap C^*(r_2)$. By definition of the Bayes classifier follows that s^* maximizes all terms, $P(s^*, r_1) = \max_s P(s, r_1)$ and $P(s^*, r_2) = \max_s P(s, r_2)$. Now we can resolve the equality and see that it holds.

$$
P(s^*, r_1) + P(s^*, r_2) = P(s^*, r_1) + P(s^*, r_2)
$$

(b) We start by solving the inequality the same way as in (a).

$$
\begin{aligned}
acc(S|R) &> acc(S|f(r)) \\
\max_s P(s,r_1) + \max_s P(s,r_2) &> \max_s \Big(P(s,r_1) + P(s,r_2) \Big)
\end{aligned}
$$

From the assumption $C^*(r_1) \cap C^*(r_2) = \varnothing$ follows that if $s^{'} \in C^*(r_1)$ and $s^{''} \in C^*(r_2)$ then $s^{'} \neq s^{''}$, for any choice of $s^{'}$ and $s^{''}$. But then, no matter which signal value s^* maximizes the right side, the inequality holds. This is simply because the aggregates responses do not agree on the optimal signal value.

$$
\begin{aligned}
s^* = s^{'} &\implies P(s^{'}, r_1) = P(s^*, r_1) \wedge P(s^{''}, r_1) > P(s^{'}, r_1) \\
s^* = s^{''} &\implies P(s^{'}, r_1) > P(s^*, r_1) \wedge P(s^{''}, r_1) = P(s^{'}, r_1) \\
s^* \in \Omega_S \setminus \{s^{'}, s^{''}\} &\implies P(s^{'}, r_1) > P(s^*, r_1) \wedge P(s^{''}, r_1) > P(s^{'}, r_1)
\end{aligned}
$$

(c) To solve the equality between $I(S;R)$ and $I(S;f(R))$ we use that by proposition 2 the mutual information is a weighted mean of the reduced entropy.

$$
\begin{aligned}
I(S;R) &= I(S;f(R)) \\
\sum_{r\in\Omega_R} P_R(r)\Big(H(S)-H(S|r)\Big) &= \sum_{f(r)\in\Omega_{f(R)}} P_{f(R)}(f(r))\Big(H(S)-H(S|f(r))\Big) \\
\sum_{r\in\{r_1,r_2\}} P_R(r)\Big(H(S)-H(S|r)\Big) &= P_{f(R)}(r')\Big(H(S)-H(S|f(r))\Big) \\
-\sum_{r\in\{r_1,r_2\}} P_R(r)H(S|r) &= -P_{f(R)}(r')H(S|f(r))
\end{aligned}
$$

Now, we plug in the definition of conditional entropy on both sides. Then, we substitute with $g(x) = x\log(1/x)$.

$$
\begin{aligned}
\sum_{r\in\{r_1,r_2\}} P_R(r)\sum_{s\in\Omega_S} P(s|r)\log_2\frac{1}{P(s|r)} &= P_{f(R)}(r')\sum_{s\in\Omega_S} P(s|f(r))\log_2\frac{1}{P(s|f(r))} \\
\sum_{r\in\{r_1,r_2\}} P_R(r)\sum_{s\in\Omega_S} g(P(s|r)) &= P_{f(R)}(r')\sum_{s\in\Omega_S} g(P(s|r')) \\
\sum_{s\in\Omega_S}\sum_{r\in\{r_1,r_2\}} \frac{P_R(r)}{P_{f(R)}(r')} g(P(s|r)) &= \sum_{s\in\Omega_S} P_{f(R)}(r')g(P(s|r'))
\end{aligned}
$$

We can use the property in equation 3.3 again.

$$
\sum_{s\in\Omega_S}\sum_{r\in\{r_1,r_2\}} \frac{P_R(r)}{P_{f(R)}(r')} g(P(s|r)) = \sum_{s\in\Omega_S} g\left(\frac{P_R(r_1)P(s|r_1)}{P_{f(R)}(r')} + \frac{P_R(r_2)P(s|r_2)}{P_{f(R)}(r')}\right)
$$

To simplify the terms we substitute $x_1 = P(s|r_1)$, $w_1 = P_R(r_1)/P_f(R)(r')$ and $x_2 = P(s|r_2)$, $w_2 = P_R(r_2)/P_f(R)(r')$.

$$
\sum_{s\in\Omega_S}\Big[w_1 g(x_1) + w_2 g(x_2)\Big] = \sum_{s\in\Omega_S} g(w_1x_1 + w_2x_2)
$$

We will show that this equality holds for each $s \in \Omega_S$.

$$\forall s \in \Omega_S : w_1 g(x_1) + w_2 g(x_2) = g(w_1 x_1 + w_2 x_2)$$

Eventually, we can apply Jensen's inequality to the strictly concave function g. Here in (c) we can assume that the conditional probabilities are equal, $x_1 = x_2$, and thus the equality holds. Since this is true for every $s \in \Omega_S$ the equality $I(S;R) = I(S;f(R))$ holds as we wanted to prove.

(d) Following the same first steps as in (c) we must switch the inequality sign once because we drop the minus sign:

$$I(S;R) > I(S;f(R))$$
$$\sum_{s \in \Omega_S} \Big[w_1 g(x_1) + w_2 g(x_2) \Big] < \sum_{s \in \Omega_S} g(w_1 x_1 + w_2 x_2)$$

Note, that $g(x)$ is a strictly concave function because $x \log(1/x)$ is strictly concave (MacKay, 2003, p. 35). By Jensen's inequality for strictly concave functions we know that:

$$\forall s \in \Omega_S : w_1 g(x_1) + w_2 g(x_2) \leq g(w_1 x_1 + w_2 x_2)$$

Additionally, we know that at least for one s the conditionally probabilities are not equal $P(s|r_1) \neq P(s|r_2) \iff x_1 \neq x_2$ because otherwise we would be in case (c) and not here in (d). For such signal values s, Jensen's inequality states that the left term is strictly smaller than the right term.

$$\exists s \in \Omega_S : w_1 g(x_1) + w_2 g(x_2) < g(w_1 x_1 + w_2 x_2)$$

Taking together that the terms in the left sum are less or equal than the terms in the right sum and that there is at least one term that is strictly smaller, the left sum is strictly smaller than the right sum. As a result, we get $I(S;R) > I(S;f(R))$ as we wanted to prove.

■

3.4 Generalizations

In this section we want to review our four assumptions. Only assumption 2 (equal weights) is essential to our findings. We have already generalized over assumption 1 (binary signal) in proposition 5. Additionally, if assumptions 3 (discrete noise) and 4 (independent noise) do not hold, our results still apply. This is satisfying because, in most applications, we can ensure assumptions on the signal but not on the noise.

3.4.1 Assumption 1 (binary signal) – non-binary signals

Assumption 1 (binary signal) requires the signal to be binary. We have already extended our results to the non-binary case so that this assumption has been dealt with. We want to briefly recap the difference between binary vs. non-binary case here.

We interpret our results in the way that the mutual information captures other aspects of the observation than the classification accuracy. While the classification accuracy is interested in the proportion of correct predictions, the mutual information additionally captures the variation in conditional accuracies (aspect 1) and, in the non-binary case, the distribution over the remaining response values (aspect 2). So the non-binary case adds another aspect that is measured by the mutual information.

3.4.2 Assumption 2 (equal weights) – relative mutual information

Assumption 2 (equal weights) requires the signal to be equally weighted. This is a necessary assumption for our results and there is, to our knowledge, no simple solution to compare responses for signals with different weightings.

When the signal is not equally weighted, a problem with the classification accuracy arises. Unequal weights may lead to misinterpretations when comparing different classification tasks (Provost et al., 1997, Congalton, 1991). For example, compare the following two binary classification tasks. Task 1 has an equally weighted signal S_1 and task 2 has unequal weights $P(S_2 = -1) = 0.99$ and $P(S_2 = 1) = 0.01$. In task 2 a completely uninformative response can achieve a classification accuracy of $acc(S_2|R) = 0.99$ by simply always predicting the signal to be -1. This problem of the classification accuracy is handled by taking into consideration other measures such as the specificity, sensitivity and receiver-operator-characteristic (ROC) curves (Baldi et al., 2000, Fawcett, 2006).

IT takes a different approach to that problem. While S_1 has a entropy of $H(S_1) = 1bit$, S_2 only has a entropy of $H(S_2) \approx 0.081bit$. Thus, the highest possible amount of information that a response can convey about S_2 is $I(S_2; R) \leq 0.081bit$. This leads to the definition of a normalized version of the mutual information (Kononenko and Bratko, 1991). The concept is known under the term

uncertainty coefficient (Press et al., 2007). The uncertainty coefficient gives the relative reduction in uncertainty about S when R is observed.

Definition 10 (**Certainty coefficient**). *The uncertainty coefficient $U(S|R)$ for S given R is the mutual information between S and R relative to the entropy of S (Press et al., 2007):*

$$U(S|R) = \frac{I(S;R)}{H(S)}$$

Surprisingly, the uncertainty coefficient does not entirely correct for unequal weights. For example, consider table 3.4. Here we used the same noise distribution and compared the equally vs. the unequally weighted case. Obviously, the classification accuracy and the mutual information differ, $acc(S_1|R_1) = 0.75$ and $acc(S_2|R_2) = 0.9925$ and $I(S_1;R_1) = 0.311bit$ vs. $I(S_2;R_2) = 0.0227bit$. But the certainty coefficient also decreases, $U(S_1|R_1) = 0.311bit$ vs. $U(S_2|R_2) = 0.281bit$. This seems to be counter-intuitive because one may think that the certainty coefficient eliminates the influence of S. Nevertheless, the weightings of S in fact influence the marginal distribution of R. To our knowledge the relation between the weightings of S and the mutual information or uncertainty coefficient has not been investigated yet. Due to the limited time frame of this thesis we can not solve this problem here. We wanted to point out that there is not a simple solution in the framework of IT.

		R_1				
		-3	-1	1	3	
S_1	1	0.125	0.250	0.125	0	0.5
	1	0	0.125	0.250	0.125	0.5

(a)

		R_2				
		-3	-1	1	3	
S_2	1	0.2475	0.4950	0.2475	0	.99
	1	0	0.0025	0.0050	0.0025	.01

(b)

Table 3.4: Two classification settings with the same noise distribution but different weightings of the signal. In (a) according to the equal weights assumption the weights are equal, $P(S_1 = -1) = P(S_1 = 1) = 0.5$. In (b) the weights are not equal, $P(S_2 = -1) = 0.99$ and $P(S_2 = 1) = 0.01$. The weightings have impact not only on the mutual information $I(S_1;R_1) = 0.311bit$ vs. $I(S_2;R_2) = 0.0227bit$ but also on the certainty coefficient, $U(S_1|R_1) = 0.311bit$ vs. $U(S_2|R_2) = 0.281bit$.

In conclusion, the weightings of the signal play a relevant and non-trivial role for both, the classification accuracy and the mutual information. Hence, assumption

2 (equal weights) is essential to our findings. When comparing different responses, their signals should exhibit the same weightings. Otherwise, the performance of the responses can not be easily compared.

3.4.3 Assumption 3 (discrete noise) – generalization to the continuous case

Assumption 3 (discrete noise) requires the noise to have a finite or countably infinite support. Note, that we never used this assumption in our proofs and therefore our results extend to the case of continuous noise. The reason we chose to introduce this assumption is to deal with the continuous case separately in this section. Here, we want to provide a brief explanation of why our results are unaffected by the problems that emerge with continuous variables in IT.

With continuous variables the definition of entropy has to be reconsidered replacing the sum by an integral.

Definition 11 (**Differential entropy**). *The differential entropy h of a continuous random variable S with probability density function f is (Cover and Thomas, 2006, p. 243):*

$$h(S) = \int_{-\infty}^{\infty} f(s) \log_2 \frac{1}{f(s)} ds \ \textit{[bit]}$$

When $f(s) = 0$ *then* $f(s) \log_2 1/f(s) \equiv 0$ *because* $\lim_{x \to 0} x \log 1/x = 0$.

The definitions of conditional entropy and mutual information for continuous variables follow alike. While most properties transfer from the discrete to the continuous case, two problems emerge.

(1) Probability distributions on continuous random variables are density distributions. As density distribution functions can have values greater than 1, the differential entropy can be negative. For example the differential entropy of a continuous random variable S with a uniform density distribution with support $[0, 0.5]$ is $h(S) = -1bit$. It is not clear how to interpret this negative entropy.

(2) Unlike the discrete entropy, the differential entropy is not independent of the measuring unit. For example, let the said uniformly distributed random variable S with support $[0, 0.5]$ be measured in cm. When measuring in mm the range is scaled up by a factor of $a = 10$. The uniform distribution is stretched to the support of $[0, 5]$ which yields a differential entropy of $h(aS) \approx 2.322bit$. This is because $h(aS) = h(S) + log|a|$, e.g. $h(10S) = -1bit + 3.322bit$ (Cover and Thomas, 2006, p. 254). This does not happen to the discrete entropy when different measuring units are used.

Nevertheless, these problems are not relevant to our proofs because at no point we applied assumption 3 and therefore we can simply leave it out of the proofs

and replace the sums by integrals where necessary. Still, we want to provide an explanation of why these seemingly problematic aspects do not affect our results.

(1) The mutual information is non-negative even for continuous variables (Cover and Thomas, 2006, p. 253).

(2) With a scaling factor a the continuous entropy changes from $h(S)$ to $h(aS) = h(S) + log|a|$. Thus scaling does not change the mutual information.

$$\begin{aligned} I(aS;R) &= h(aS) - h(aS|R) \\ &= h(S) + log|a| - (h(S|R) + log|a|) \\ &= h(S) - h(X|R) \\ &= I(S;R) \end{aligned}$$

It is also noteworthy that continuous random variables can be quantized to discrete random variables. The mutual information of quantized variables converges to the mutual information of the continuous variables (Cover and Thomas, 2006, p. 251). Thus, we can treat the mutual information in the continuous case as the limit of the discrete mutual information of quantized variables. In conclusion, continuous variables pose no problem to our results.

3.4.4 Assumption 4 (independent noise) – extreme cases with independent noise

Assumption 4 (independent noise) requires that the noise is independent of the signal. We did not use this assumption in any of our proofs. Therefore, our results generalize to the dependent noise case. But when our bounds even hold for the dependent noise case one may ask whether there are stronger bounds in the independent noise case. In this section we provide independent noise examples for the minimal vs. maximal mutual information proving that our bounds are tight at least for the binary case.

We start with the example of maximal mutual information for a given classification accuracy. In this case the noise is uniform. The following noise seems rather artificial and complex but we only want to prove the existence of a case that maximize the mutual information with independent noise (see table 3.6):

$$P(N=n) = \begin{cases} c_1 & \text{if } n \in \left\{ \frac{(k-0.5)2c_1}{acc(S|R)-0.5} \mid k \in \mathbb{Z} \wedge -\frac{1}{4c_1} < |k| \leq \frac{1}{4c_1} \right\}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{where } c_1 = \frac{1 - acc(S|R)}{2} T\left(\frac{acc(S|R) - 0.5}{1 - acc(S|R)}\right)$$

Here, T is Thomae's function (Bartle and Sherbert, 1999, p. 122). The constant c_1 represents the fractions of the probability mass. For example a classification accuracy of $acc(S|R) = 0.8$ leads to $c_1 = \frac{1}{20}$ and the contingency table 3.5. This table represents two overlapping uniform probability distribution that can be aggregated to obtain the introductory example. The aggregation has the same classification accuracy and mutual information because only responses with equal conditional probabilities are aggregated.

		R															
		$-\frac{15}{6}$	$-\frac{13}{6}$	$-\frac{11}{6}$	$-\frac{9}{6}$	$-\frac{7}{6}$	$-\frac{5}{6}$	$-\frac{3}{6}$	$-\frac{1}{6}$	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{5}{6}$	$\frac{7}{6}$	$\frac{9}{6}$	$\frac{11}{6}$	$\frac{13}{6}$	$\frac{15}{6}$
S	-1	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	0	0	0	0	0	0
	1	0	0	0	0	0	0	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$

		$f(R)$		
		-1	0	1
S	-1	0.3	0.2	0
	1	0	0.2	0.3

Table 3.5: Contingency table with uniform noise and its aggregation with equal classification accuracy $acc(S|R) = 0.8$ and maximal mutual information $I(S;R) = 0.6bit$.

The second example with minimal mutual information has symmetric, discrete, exponential-like noise. Again, what is important here is not the mathematical details of this function but that exponential-like noise leads to the minimal mutual information in the independent noise case (see table 3.6):

$$P(N = n) = \begin{cases} c_2 e^{-\lambda|n|} & \text{if } n \in \{2k | k \in \mathbb{Z}\}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{where } c_2 = \frac{e^{2\lambda} - 1}{e^{2\lambda} + 1} \text{ and } \lambda = 0.5 \cdot \log_e \left(\frac{acc(S|R)}{1 - acc(S|R)} \right)$$

Here, c_2 is the normalization constant and λ is the parameter that determines the classification accuracy. All conditional accuracies are equal. For example, a classification accuracy of $acc(S|R) = 0.8$ leads to $c_2 \approx 0.433$, $\lambda \approx 1.386$ and the contingency table 3.6. All conditional accuracies are $acc(S|r) = 0.8$. This example can be aggregated into the introductory example with minimal mutual information. Again, the classification accuracy and mutual information stay unchanged because only responses with equal conditional probabilities are aggregated.

		R						
		...	-3	-2	-1	1	2	3	...
S	-1		0.019	0.075	0.300	0.075	0.019	0.005	
	1		0.005	0.019	0.075	0.300	0.075	0.019	

		$f(R)$	
		-1	1
S	-1	0.4	0.1
	1	0.1	0.4

Table 3.6: Contingency table with symmetric, discrete, exponential distribution and its aggregation with equal classification accuracy $acc(S|R) = 0.8$ and minimal mutual information $I(S;R) \approx 0.278bit$.

At this point we assume that concerns may arise about the support sizes. First, the support sizes of the two responses in the introduction were not equal. And now the second example requires an infinite support size in order to minimize the mutual information in the independent noise case. In principle, a support size of 2 is sufficient to either display minimal mutual information (two equal conditional accuracies) or to display maximal mutual information (conditional accuracies of 0.5 and 1). But it is more complicated in the independent noise case. We suspect that by introducing restrictions on the support size of the response Ω_R one could obtain stronger bounds for the independent noise case. But because of the limited time frame of this thesis we can not solve this problem here.

We have shown that even with independent noise our bounds are tight in the binary case. In the non-binary case the bounds are tight for dependent noise and we conjecture that they are in the independent case, too. Although in this case R would have to be a vector of independent dimensions to make the independent noise work. Because of the complexity in this case we chose not to incorporate it into this thesis. However, restrictions on the support size of the response (and, in the non-binary case, on the dimension of R) may lead to stronger bounds.

Chapter 4

Reinterpreting the indirect task advantage

In this chapter we reveal a flawed interpretation in psychological research. This interpretation is based on the comparison between a binary vs. a continuous response. We argue that the binary response can be seen as an aggregation of the continuous response and therefore the information is strictly smaller. Using our theoretical results we can show that the continuous response outperforms the binary because of the aggregation and not because of underlying differences.

In 2014, ten Brinke et al. have claimed that unconscious processing is superior to conscious processing. They compare conscious vs. unconscious performances in the direct vs. the indirect task. Then, they observe better performances in the indirect task indicating superior unconscious processing (*indirect task advantage*). But the problem is that they use a binary response for the direct task in contrast to a continuous response in the indirect task. In fact, even if we assume that conscious and unconscious performances are equal the direct task forces an aggregation into a binary response. This potentially decreases the mutual information leading to a worse response in the direct task and to a better performance in the indirect task. Thus, we argue that their interpretation is flawed because the differences they find may not come from superior unconscious processing but may have only been caused by restricting the response to be binary in the direct task.

First, we will describe their findings that display differences between direct and indirect tasks (section 4.1). Then, we will show that even for equal underlying performances these differences appear (section 4.2). Thus, in section 4.3 we conclude that interpreting a superior unconscious performance from this difference is incorrect. The indirect task advantage in their results is a response metric advantage.

4.1 Findings from ten Brinke et al. (2014)

In experiment 2 ten Brinke et al. (2014) compare conscious vs. unconscious lay lie detection. They show videos to participants in which a suspect pleads to be not

guilty. Some of these pleas are lies and some are the truth. Then the participants are asked to classify each video as a lie or the truth. This is the direct task because it represents conscious evaluation. In this direct task the participants did not perform above chance level with a mean accuracy of 49.62%, $t(65) = -0.27$, $p = .79$, $d = -0.01$. In fact, the literature supports the notion that participants are close to chance level (54%) when it comes to conscious detection of lies (Bond and DePaulo, 2006).

The indirect task representing unconscious processing involves an implicit association task with priming (Greenwald et al., 1998, 2003). In this task participants first saw the videos of the pleading suspects. Then, they had to classify words such as "deceitful" or "valid" into one of two semantic categories. The first category is related to lies whereas the second category is related to the truth. Shortly before this task participants were primed. The masked prime was a frame of a video the participants saw beforehand. In the congruent condition the participants were primed with a frame of a video that was related to the same semantic category as the subsequent word. For example, participants were primed with a frame of a video in which the actor lied and then had to classify "deceitful" into the lie category. Therefore, the content of the prime was semantically congruent to the content of the word. In the incongruent condition the semantic content of prime and word differed (e.g. a picture of a video in which the actor lied and the word "valid"). In this task, participants classified the words faster in the congruent condition compared to the incongruent condition. For single reaction times the difference between congruent and incongruent condition was $d = 0.03$ standard deviations ($SD = 0.11$). The averages of reaction times over multiple trials (64 per video pair, 12 videos) yielded a significant result, $t(65) = 2.26$, $p = .027$, $d = 0.27$.

This reflects a better performance in the indirect task compared to the direct task because the effect size in the indirect task ($d = 0.27$) is higher than in the direct task ($d = -0.01$). From this ten Brinke et al. (2014, 2016) interpret that the unconscious performance is superior to the conscious performance.

4.2 Model with equal conscious and unconscious performance

From our understanding a better performance in the indirect task does not indicate superior unconscious performance. Assume for a moment that the conscious and unconscious performance is equal. The direct task poses a restriction because it requires a binary response. Participants are forced to internally dichotomize their response. This dichotomization is an aggregation that will lead to a decrease in mutual information and in the direct task performance even when the underlying performances are equal.

To elaborate on this argument we now apply IT by modelling the problem in terms of signals, noise and responses. In the direct task the signal is the validity

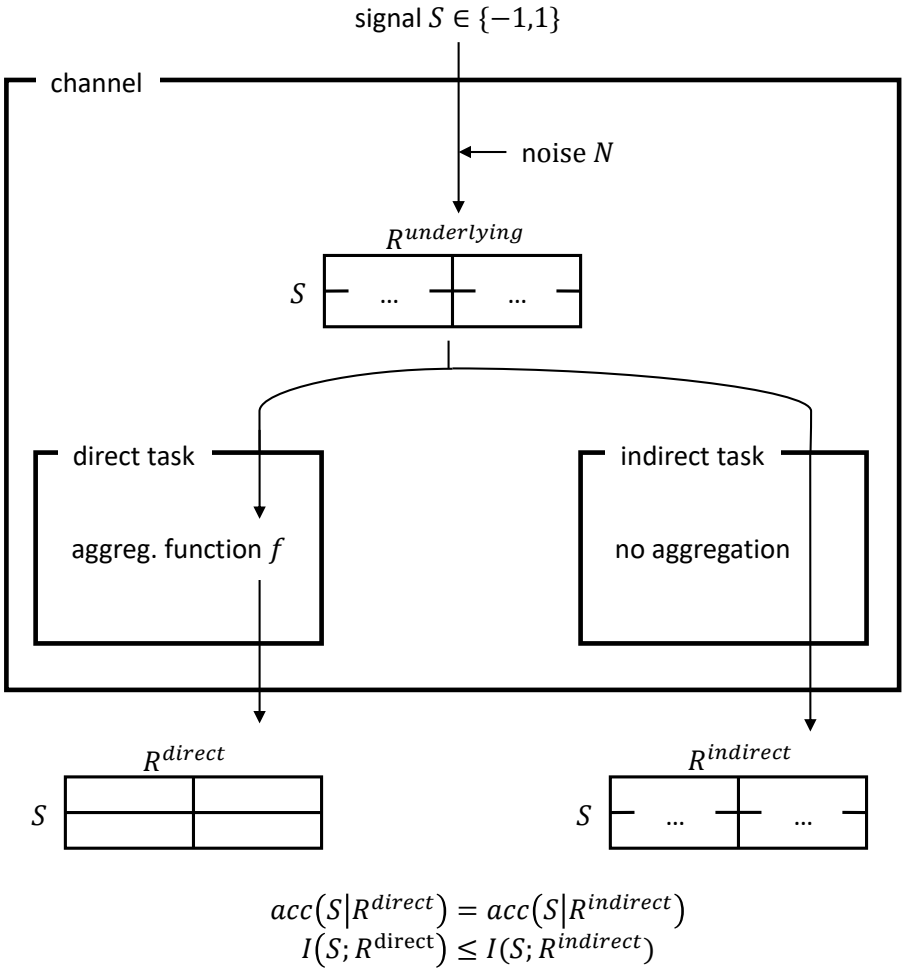


Figure 4.1: The channel model of direct and indirect task with equal underlying response. Since the direct task restricts the response format to be binary the mutual information may decrease.

of the video, $S^{direct} = -1$ for a lie and $S^{direct} = +1$ for the truth. In the indirect task the signal is the congruency condition, $S^{indirect} = -1$ for incongruent and $S^{indirect} = +1$ for congruent conditions. Both signals can be treated as equal because despite their semantic differences they are binary and equally weighted. We now make our crucial assumption that there is an equal underlying response $R^{underlying}$ in both tasks. In the direct task this response has to be internally dichotomized by the participants, $R^{direct} = f(R^{underlying})$. Whereas, in the indirect task the response is equal to the underlying response, $R^{indirect} = R^{underlying}$. For an illustration of our model see figure 4.1. The "channel" is the participant who, given a signal, produces an imperfect response either in the direct or in the indirect task.

We will now go into the mathematical details of our model. We will assume Gaussian noise N and choose median split as the optimal dichotomization f. With

this we will dichotomize the responses in the indirect task (reaction times) from the data of of ten Brinke et al. (2014). As a result, the mutual information will decrease at least by 35.8% and with that the effect size will shrink from $d = 0.03$ to $d = 0.01$. This reveals that the binary response of the direct task necessarily performs worse than the continuous response of the indirect task. Thus, even if the underlying performances are equal, the direct task allows for less information to be conveyed leading to a lower effect size.

We assume the noise N to be independent and Gaussian. Then, the underlying response as well as the indirect response $R^{indirect}$ are mixtures of two Gaussians (see figure 4.2). Without loss of generality assume $E(N) = 0$ and $Var(N) = \sigma^2$. This is not a restriction because a linear transformation relates this setting to any other setting with the same effect size, Cohen's $d = \frac{|s1-s2|}{\sigma}$. Hence, the mutual information $I(S; R^{indirect})$ is a function of d. There is no closed form for calculating $I(S; R^{indirect})$ in this case but there are approximations available (Nasif and Karystinos, 2005). Thus, we can plot d against the mutual information of the indirect task, see table 4.3.

For the direct task we have to apply the optimal dichotomization f which, in this case, is a median split (see figure 4.2).

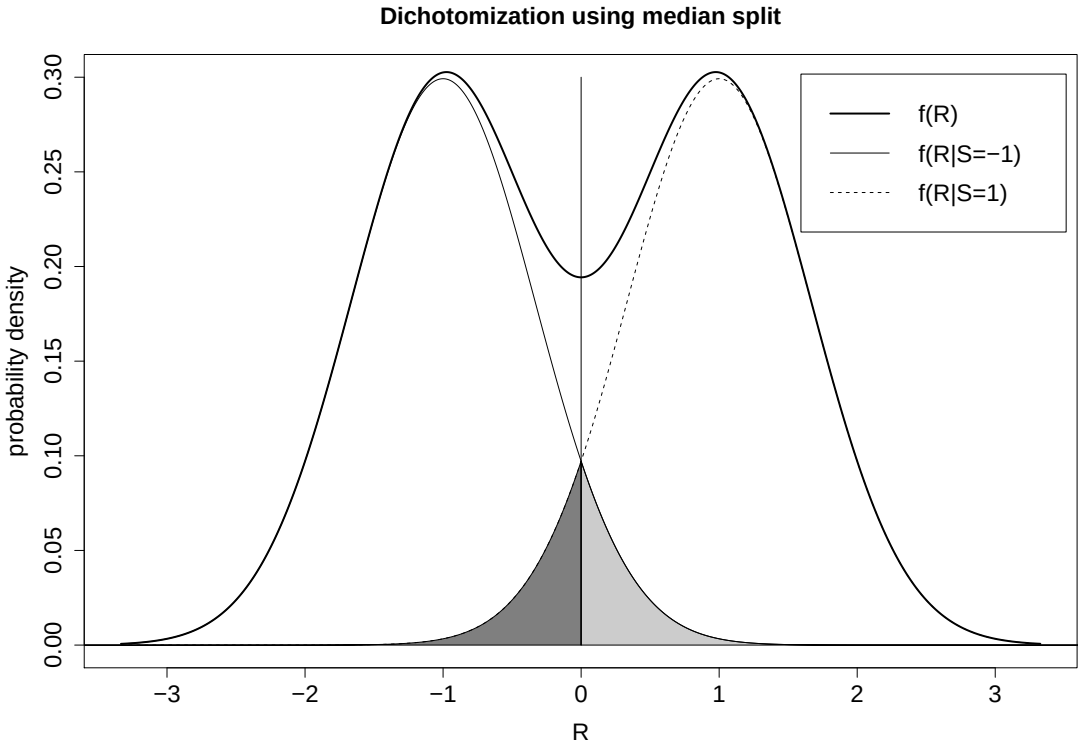


Figure 4.2: A mixture of two Gaussians can be optimally dichotomized by a median split. The light and dark grey areas indicate the error probability $\Phi(d/2)$, where d is Cohen's effect size parameter and Φ is the cumulative normal distribution.

$$f(R) = \begin{cases} 1 & \text{if } R \geq 0, \\ -1 & \text{if } R < 0. \end{cases}$$

From proposition 6 we can deduce what happens to the mutual information upon this aggregation. For Gaussian noise the conditional accuracies are $acc(S|R^{underyling} = r) = [1 + e^{-2|r|}]^{-1}$ and they are strictly monotone for $R \geq 0$ and $R < 0$, respectively. This implies that the conditional probabilities of aggregated responses values are not equal and therefore the mutual information strictly decreases, $I(S; R^{direct}) = I(S; f(R^{indirect})) < I(S; R^{indirect})$. Figure 4.3 plots the mutual information in the direct vs. indirect task and shows this noteworthy difference as predicted by our proposition 6.

Now consider $d = 0.03$ because that was mean effect size of a single indirect task response in the experiment of ten Brinke et al. (2014). This leads to $I(S; R^{indirect}) \approx 0.00016 bit$ for the indirect task and $I(S; R^{direct}) \approx 0.0001 bit$ for the direct task. The dichotomization discards 35.8% of the mutual information. This loss of information is accompanied by a smaller effect size and test power according to Cohen (1983, 1988). In fact, dichotomizing a mixture of two Gaussians with $d = 0.03$ results in a decreased effect size of only $d = 0.01$. To make this argument as clear as possible: Even if the participants knew their actual unconscious responses from the indirect task, they would still perform worse in the direct task because it restricts their response to be binary!

We have assumed Gaussian noise for the reaction times. But reaction times are often considered to have a log-normal noise distribution. When we use log-normal noise, estimate the parameters based on the data from ten Brinke et al. (2014) and then perform the same analysis we get even stronger results. The information loss upon dichotomization in this case is at 59.7%, $I(S; R^{indirect}) \approx 0.000256 bit$ and $I(S; R^{direct}) \approx 0.000103 bit$ (We have tried different parameter estimates and this was the one losing the least amount of information).

4.3 No evidence for superior unconscious performance

To summarize the last two sections: (4.1) there are different effect sizes in the direct vs. indirect task; and (4.2) this difference occurs even when underlying conscious vs. unconscious performances are equal because the direct task forces a binary response. In conclusion, a better indirect task performance (indirect task advantage) does not necessarily indicate superior unconscious performance. To support our argument we want to cite other findings related to this issue.

First, Cohen (1983, 1988) explained how dichotomization decreases information, effect size and test power. Our contribution to his argument is that we calculated how much information in *bit* is lost in our case. Even though Cohen was considering the dichotomization that is done explicitly by researchers in their

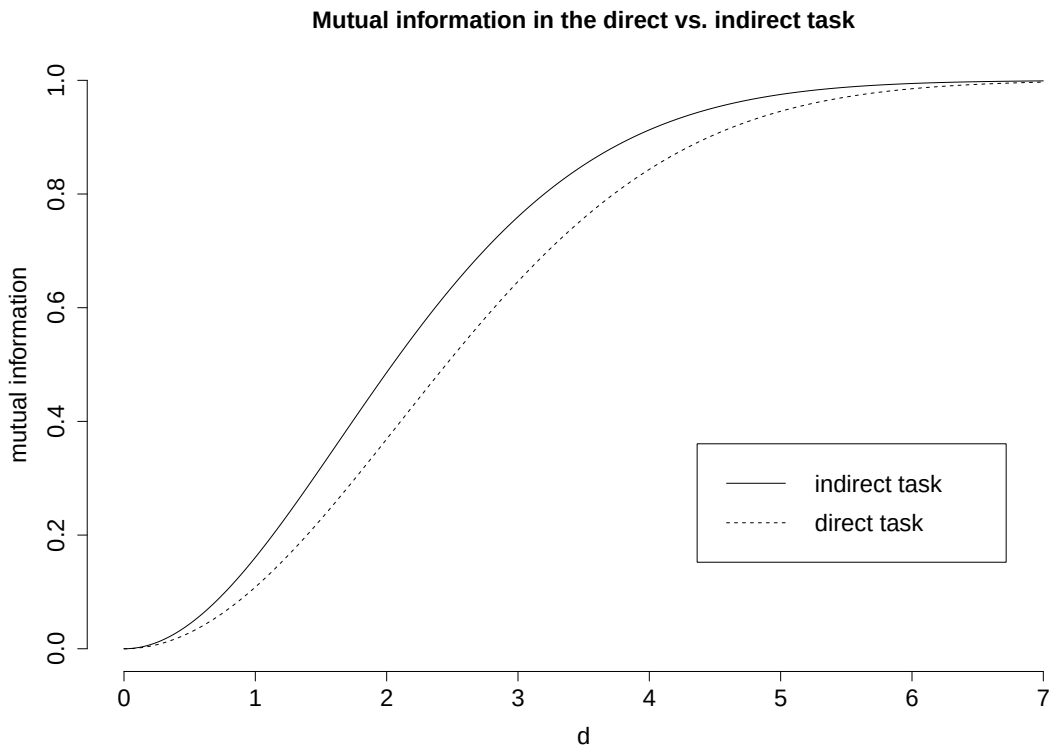


Figure 4.3: The mutual information in the direct and indirect task assuming independent Gaussian noise and the optimal dichotomization.

statistical analyses, we argue that the same argument holds when researchers force participants to dichotomize internally. The decreased information due to this dichotomization explains the stronger effect size in the indirect task.

Second, Reingold and Merikle (1988) have repeatedly recommended that when comparing conscious and unconscious performances the direct and indirect task have to use the same response metric (Merikle and Reingold, 1991, 1992). They refer to Eriksen (1956) who criticised Lazarus and McCleary (1951) for using a discrete response in the direct task and a continuous response in the indirect task. The critique is that the difference between the direct and indirect task can be attributed to the difference in the response metric and not necessarily to a superior unconscious performance. The same critique applies to the lie detection experiment from ten Brinke et al. (2014). They used a binary response in the direct task ("lie" or "truth") and a continuous response in the indirect task (reaction times). We agree with this critique entirely.

Third, our understanding is based on and supported by the findings of Franz and von Luxburg (2015). They reproduced the results from ten Brinke et al. (2014) based on the same data. Their finding was that indirect task does not allow for a better classification accuracy than the meta-analysis by Bond and DePaulo (2006) has shown (54%). Among other analyses, they dichotomized the reaction times in the data of using a designated median for each participant. This is the

explicit dichotomization that we assume the participants had to do internally. The resulting classification accuracy was 50.6% ($SD = 2.65\%$) compared to the classification accuracy in the direct task 49.62% ($SD = 11.36\%$). The classification accuracies are essentially equal. This fits to our model because we assume that the dichotomization decreases the mutual information but keeps the classification accuracy equal as we see here. Based on this we doubt that there really is an underlying difference between conscious vs. unconscious performance.

Other researchers show the same pattern of results. They – at least implicitly – compare binary responses and continuous responses. For example, Dehaene et al. (1998) show that participants can not reliably differentiate masked primes (numbers) with a duration of 43ms in the direct task (binary response). In the indirect task, they find larger effect sizes measuring reaction times (continuous responses). The same comparison is done by Pessiglione et al. (2007) but with pictures of coins instead of numbers. They too find larger effect sizes in the indirect task using a continuous response compared to the direct task where they use a binary response. In both cases the participants might have been able to perform better if the direct task had not restricted them to binary responses. Nevertheless, Pessiglione et al. (2007) interpret from a missing significant result in the direct task that "the analysis could then be restricted to all situations where subjects guess at chance level about stimulus identity". But a missing significant result does not mean that the participants in the direct task perform on chance level, especially when the test power is diminished due to the binary response metric.

The binary response metric of the direct task may have lead a decrease in the mutual information and effect size compared to the indirect task. This indirect task advantage is no evidence for an underlying difference in the performances if the response metrics are different. Yes, the performance in the indirect task is better than in the direct task as indicated by larger effect sizes. But this may only be because it is harder to achieve large effect sizes with a binary response due to lower mutual information even though the classification accuracy is equal. In conclusion, lower effect sizes in the direct task compared to the indirect task are no evidence for superior unconscious processing if different response metrics are used (e.g. binary vs. continuous).

Chapter 5

Loss and risk

In this chapter we will investigate the performance of observations with different mutual information and equal classification accuracy. As we have explained before, the mutual information captures not only the classification accuracy but also two other properties, the variation of conditional accuracies (aspect 1) and the entropy over the remaining response values (aspect 2). We will characterize one scenario in which these aspects are irrelevant and another in which aspect 1 is relevant by defining a loss functions for each scenario. In both scenarios we will show the performance of observations with minimal vs. maximal mutual information.

For this investigation we will jump back to the notation of observations X and labels Y instead of responses R and signals S. We do so because this is the common terminology when describing loss functions. We will consider binary classification tasks in which an observation X is used to predict label $\hat{Y}$ while the true label is $Y \in \{-1, 1\}$. The prediction may be incorrect, $\hat{Y} \neq Y$, and we measure the cost of incorrect predictions with a loss function, $l(X, Y, \hat{Y})$. We want to choose observations X in order to minimize the expected loss, or risk: $R_l(X) = E[l(X, Y, \hat{Y})]$. Because we want to compare different observations we use the optimal strategy π to predict $\hat{Y} = \pi(X)$. Thus, π minimizes $R_l(X)$. Depending on the loss function $\pi(X)$ may be the Bayes Classifier $C^*(X)$ predicting the most probable label. But it can also be that $\pi(X) = 0$ because a prediction is no accurate enough so that we choose to abstain from making a prediction by choosing the value 0. We will now define two loss functions and compare observations with constant classification accuracy, $acc = acc(Y|X)$, and minimal, $X^{\text{acc}}_{\text{min I}}$, vs. maximal mutual information, $X^{\text{acc}}_{\text{max I}}$. As we have shown before the minimal mutual information is obtained when all conditional accuracies are equal. The maximal mutual information occurs when all conditional accuracies are 0.5 or 1. This is illustrated in table 5.1. For these two observations we derive and compare their respective risks and see for which loss functions $X^{\text{acc}}_{\text{max I}}$ performs better.

		$X^{\text{acc}}_{\text{min I}}$	
		-1	1
Y	-1	$\frac{acc}{2}$	$\frac{1-acc}{2}$
	1	$\frac{1-acc}{2}$	$\frac{acc}{2}$

		$X^{\text{acc}}_{\text{max I}}$		
		-1	0	1
Y	-1	$\frac{2acc-1}{2}$	$1-acc$	0
	1	0	$1-acc$	$\frac{2acc-1}{2}$

Table 5.1: Observations with minimal, $X^{\text{acc}}_{\text{min I}}$, vs. maximal mutual information, $X^{\text{acc}}_{\text{max I}}$, for a fixed classification accuracy $acc = acc(Y|X)$. These examples are representative because minimal mutual information requires all conditional accuracies to be on average and maximal mutual information requires them to be at 0.5 and 1.

5.1 0-1-loss

The first scenario is characterized by the 0-1-loss. An incorrect prediction leads to a loss of 1 and a correct prediction to 0. A prediction has to be made for all observations.

$$l_1(X, Y, \hat{Y}) = \begin{cases} 0 & \text{if } \hat{Y} = Y, \\ 1 & \text{otherwise.} \end{cases}$$

Here, the optimal strategy is to always predict the most probable label. In this scenario the risk simply boils down to $R_{l_1}(X^{\text{acc}}_{\text{min I}}) = R_{l_1}(X^{\text{acc}}_{\text{max I}}) = 1 - acc$ and is independent of the mutual information. Thus, in scenario in which one always has to predict only the classification accuracy is relevant.

5.2 0-λ-1-loss

There are scenarios in which it might be more desirable to make no prediction instead of an unreliable prediction. This includes scenarios where it is better to take no action than to take an action based on unreliable information. With this, we aim to support the findings by Hu (2014) who argues that the mutual information is the preferred measure in cases in which an observation can be "rejected", instead of making an unreliable prediction. For such scenarios we will define a 0-λ-1-loss. It is a 0-1-loss including the option to make no prediction or to take no action. But using this option costs λ because e.g. it costs time and effort to apply the measurement even though it does not produce the desired prediction.

$$l_2(X, Y, \hat{Y}) = \begin{cases} 0 & \text{if } \hat{Y} = Y, \\ \lambda & \text{if } \hat{Y} = 0, \\ 1 & \text{otherwise.} \end{cases}$$

In this scenario the optimal strategy depends on λ and $acc(Y|X)$ because depending on how much it costs to abstain from making a prediction it may be favorable to risk making a prediction or not. We can distinguish four sub-scenarios (a)-(d) as shown in figure 5.1. In (a) the optimal strategy is to never predict, $\pi(X) = 0$. In (d) the optimal strategy is to always predict as in the 0-1-loss. But in (b)-(c) we deal with what we call "risky" scenarios. In these cases it is only favorable to make a prediction if the conditional accuracy of a particular observation is greater than $1 - \lambda$. Because $X^{\text{acc}}_{\text{max I}}$ only has observations with conditional accuracy of 0.5 or 1 the optimal strategy is to always predict on the informative observations and never predict on the non-informative observations. In contrast, $X^{\text{acc}}_{\text{min I}}$ only has observations with conditional accuracy equal to acc so that the strategy depends on whether $acc > 1 - \lambda$ or not. In (b) the best strategy is to never predict. But in (c) it is optimal to always predict.

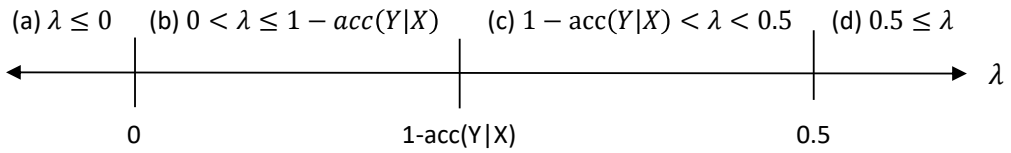


Figure 5.1: Axis for λ representing the different scenarios (a)-(d) in which different strategies π are optimal. From left to right (for higher λ) it becomes more favorable to make predictions even for unreliable observations.

The optimal strategies π lead to the risks calculated in table 5.2. In (b)-(c) $\lambda < 0.5$ and $acc(Y|X) > 0.5$ so that the risk is strictly smaller for the observation with maximal mutual information, $X^{\text{acc}}_{\text{max I}}$, compared to $X^{\text{acc}}_{\text{min I}}$. Thus, the observation with the maximal mutual information performs better than the observation with the minimal mutual information.

	$R_{l_2}(X^{\text{acc}}_{\text{min I}})$		$R_{l_2}(X^{\text{acc}}_{\text{max I}})$
(a)	λ	$=$	λ
(b)	λ	$>$	$2\lambda(1 - acc(Y\|X))$
(c)	$1 - acc(Y\|X)$	$>$	$2\lambda(1 - acc(Y\|X))$
(d)	$1 - acc(Y\|X)$	$=$	$1 - acc(Y\|X)$

Table 5.2: Risk of 0-λ-1-loss for scenarios (a)-(d). The risk is measures for observations with minimal vs. maximal mutual information but with the same classification accuracy $acc = acc(Y|X)$.

As an example, consider the lie detection task from chapter 4. Say, there are two

lie detectors, one with minimal and one with maximal mutual information. Both have a classification accuracy of $acc(Y|X) = 0.8$. When we plug these values into table 5.1 we obtain our introductory examples from chapter 1: The first lie detector with the minimal mutual information will always predict with a constant accuracy of 80%. So 1 out of 5 times we get an incorrect prediction. The second lie detector will make a perfectly accurate prediction in 60% of the cases. In the remaining 40% of the cases this lie detector will make an unreliable prediction that is only in 50% of the cases correct. Now say, the cost of applying the lie detector is 10 (in an arbitrary unit). If we make a correct prediction we are happy and consider the cost to be 0, but if we make a wrong prediction the cost is 100 in total. We can rescale this scenario to a 0-0.1-1-loss. Then, we deal with sub-scenario (b). The risk of the first lie detector is $R_{l_2}(X^{\text{acc}}_{\text{min I}}) = 0.1$ and for the second it is $R_{l_2}(X^{\text{acc}}_{\text{max I}}) = 0.04$. In order to perform equally well on the risk, the observation with minimal mutual information would require a classification accuracy of $acc(Y|X) = 0.9$ instead of 0.8. So in this case, the observation with the mutual information performs notably better and it is worth checking the mutual information to evaluate this observation even if the classification accuracy would have not been in favor of it.

We conclude that the mutual information is relevant in "risky" scenarios (0-λ-1-loss with $0 < \lambda < 0.5$). In these scenarios observations with maximal mutual information strictly outperform observations with minimal mutual information at equal classification accuracies. But depending on the scenario it could also be that an observation with higher mutual information is outperformed by an observation with lower mutual information but higher classification accuracy. As always, the general conclusion is that for different scenarios one should apply the appropriate measures. Here we have argued that in risky scenarios the mutual information is relevant and should be taken into consideration.

Chapter 6

Conclusion

In this thesis we have investigated the relation between the classification accuracy and the mutual information. The two measures are loosely related to each other because they capture different properties of the classification task. The classification accuracy measures the probability of a correct prediction but the mutual information additionally captures the variation of conditional accuracies (aspect 1) and the distribution over the remaining observation values (aspect 2). Their bounded relation is mediated through conditional accuracies. We found that when observations are further processed by aggregating observations, the mutual information can decrease while the classification accuracy stays constant. With this theoretical result we have shown that comparing binary vs. continuous observations is inherently biased towards a better performance of the continuous observation. Neglecting that continuous observations are expected to outperform binary observations has lead to a flawed interpretation in psychological research. Additionally, we have shown that in risky scenarios (such as using a lie detector as evidence in court) the mutual information is relevant and should be taken into consideration. In general, it is important to consider the measure of mutual information next to the classification accuracy.

Our theoretical analysis considered the Bayes classifier and assumed that the true probabilities are known. This is not the case in real world applications. In these applications the classification accuracy and the mutual information have to be estimated. But estimators for the mutual information are inherently biased and the bias also depends on the true probability distribution that is to be estimated (Grassberger, 2003, Paninski, 2003, Kraskov et al., 2004). This poses substantial difficulties on measuring the mutual information.

A way to deal with this problem might be to develop statistical tests for the difference between the mutual information of two observations with their respective labels. Then, even though we could not make an unbiased estimation, we would be able to decide which observation conveys more information. Developing suitable statistical tests for this application remains an open research question.

Regarding our critique on the flawed interpretation in psychological research we want to clarify that we do not rule out the possibility that unconscious human processing may outperform conscious processing. But we argue that the findings by

ten Brinke et al. (2014) are not to be considered as evidence for it. In line with the recommendations by Reingold and Merikle (1988) we suggest that evidence for superior unconscious processing must be based on comparisons with equal response metrics. For example researchers could dichotomize the continuous response so that both tasks feature a binary response and compare those. Another possibility to achieve equal response metrics is to give participants an n-point Likert scale in the direct task. This would allow them to give a response with a higher mutual information compared to a binary response because they can now indicate how certain they are about their prediction. This reflects variation in conditional accuracy which allows for a higher mutual information. Then, the continuous response from the indirect task would have to be transformed into a n-ary response as well in order to make a fair comparison. It is also conceivable to allow participants to use a continuous scale in the direct task. These so called "visual analogue scales" are commonly used in pain research and a detailed description can be found in the doctor thesis by Funke (2010).

That being said, we want to thank ten Brinke, Stimson and Carney for making their data accessible. This allows for a transparent discussion and facilitates scientific progress.

Potentially, there is another way to increase the mutual information in the direct task of lay lie detection. Until now we have considered the direct task to be a binary classification task and videos being either a lie or the truth. But the videos are not binary objects, they are complex, multidimensional and continuous objects reflecting for example the stress level of the pleading suspect in it. Thus, future psychological research could model the videos as continuous or at least non-binary signals allowing to capture the influence of these stimuli inhomogeneities on the direct and indirect task performance.

A particularly interesting open question is: How is boosting affected by mutual information? We understand that combining different observations with higher (vs. lower) mutual information leads to better performances even when the classification accuracy is initially equal. To see this, consider the introductory examples and perform boosting on three independent observations. Predicting the label based on three observations with minimal mutual information and classification accuracy of 80% leads to a combined classification accuracy of $0.8^3 + 3 \cdot 0.8^2 \cdot 0.2 = 89.6\%$. In contrast, three observations with maximal mutual information lead to a combined classification accuracy of $1-(1-0.6)^3 = 93.6\%$. Thus, the performance of boosting seems to be influenced, not only by the classification accuracy, but also by the mutual information. We think that a theoretical investigation on boosting in the framework of information theory is a fruitful future research topic.

In line with Hu (2014) we want to conclude by stating that, at least in some cases, the mutual information formalizes better what humans are concerned about, compared to the classification accuracy.

Bibliography

Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

Robert G Bartle and Donald R Sherbert. Introduction to real analysis, 1999.

Charles F Bond and Bella M DePaulo. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234, 2006.

Jacob Cohen. The cost of dichotomization. *Applied Psychological Measurement*, 7:249–253, 1983.

Jacob Cohen. Statistical power analysis for the behavioural sciences. Hillside. *NJ: Lawrence Earlbaum Associates*, 1988.

Russell G Congalton. A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, 37(1):35–46, 1991.

Thomas M Cover and Joy A Thomas. *Elements of information theory 2nd edition*. Wiley-interscience, 2006.

Stanislas Dehaene, Lionel Naccache, Gurvan Le Clec'H, Etienne Koechlin, Michael Mueller, Ghislaine Dehaene-Lambertz, Pierre-Francois van de Moortele, and Denis Le Bihan. Imaging unconscious semantic priming. *Nature*, 395(6702):597–600, 1998.

Charles W Eriksen. Subception: Fact or artifact? *Psychological Review*, 63(1):74, 1956.

Robert M Fano and David Hawkins. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8): 861–874, 2006.

John W Fisher, Michael Siracusa, and Kinh Tieu. Estimation of signal information content for classification. In *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th*, pages 353–358. IEEE, 2009.

Volker H Franz and Ulrike von Luxburg. No evidence for unconscious lie detection a significant difference does not imply accurate classification. *Psychological science*, 26:1646–1648, 2015.

Frederik Funke. Internet-based measurement with visual analogue scales. an experimental investigation. *Online im Internet: http://nbn-resolving. de/urn: nbn: de: bsz*, 2010.

Peter Grassberger. Entropy estimates from insufficient samplings. *arXiv preprint physics/0307138*, 2003.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

Anthony G Greenwald, Brian A Nosek, and Mahzarin R Banaji. Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology*, 85(2):197, 2003.

Bao-Gang Hu. What are the differences between bayesian classifiers and mutual-information classifiers? *IEEE transactions on neural networks and learning systems*, 25(2):249–264, 2014.

Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

Igor Kononenko and Ivan Bratko. Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6(1):67–80, 1991.

Vladimir A Kovalevsky. The problem of character recognition from the point of view of mathematical statistics. *Character Readers and Pattern Recognition*, (eds. V.A. Kovalevskij):3—-30, 1967.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Richard S Lazarus and Robert A McCleary. Autonomic discrimination without awareness: A study of subception. *Psychological review*, 58(2):113, 1951.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Philip M Merikle and Eyal M Reingold. Comparing direct (explicit) and indirect (implicit) measures to study unconscious memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(2):224, 1991.

Philip M Merikle and Eyal M Reingold. Measuring unconscious perceptual processes. *Perception without awareness: Cognitive, clinical, and social perspectives*, pages 55–80, 1992.

Ahmed O Nasif and George N Karystinos. Binary transmissions over additive gaussian noise: A closed-form expression for the channel capacity. In *Proc. 2005 Conf. Inf. Sci. and Syst.(CISS), Baltimore, MD, USA*, 2005.

Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Mathias Pessiglione, Liane Schmidt, Bogdan Draganski, Raffael Kalisch, Hakwan Lau, Ray J Dolan, and Chris D Frith. How the brain translates money into force: a neuroimaging study of subliminal motivation. *Science*, 316(5826):904–906, 2007.

WH Press, SA Teukolsky, WT Vetterling, and BP Flannery. Conditional entropy and mutual information. *Numerical Recipes 3rd Edition: The Art of Scientific Computing, Cambridge University Press, New York*, 2007.

Foster J Provost, Tom Fawcett, et al. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In *KDD*, volume 97, pages 43–48, 1997.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org`.

Eyal M Reingold and Philip M Merikle. Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, 44(6):563–575, 1988.

Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 4:379–423, 1948.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

Leanne ten Brinke, Dayna Stimson, and Dana R Carney. Some evidence for unconscious lie detection. *Psychological science*, page 1098–1105, 2014.

Leanne ten Brinke, Kathleen D Vohs, and Dana R Carney. Can ordinary people detect deception after all? *Trends in Cognitive Sciences*, 20(8):579–588, 2016.

Yong Wang and Bao-Gang Hu. Derivations of normalized mutual information in binary classifications. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*, volume 1, pages 155–163. IEEE, 2009.

Erklärung der Urheberschaft

Ich versichere an Eides statt, dass ich die Master thesis im Studiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Ort, Datum Unterschrift

Erklärung zur Veröffentlichung

Ich erkläre mein Einverständnis mit der Einstellung dieser Master thesis in den Bestand der Bibliothek.

Ort, Datum Unterschrift