entropy

MDPI

Article

Deconstructing Cross-Entropy for Probabilistic Binary Classifiers

Daniel Ramos *, Javier Franco-Pedroso, Alicia Lozano-Diez and Joaquin Gonzalez-Rodriguez

AuDIaS-Audio, Data Intelligence and Speech, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Calle Francisco Tomas y Valiente 11, 28049 Madrid, Spain; Javier.franco@uam.es (J.F.-P.); alicia.lozano@uam.es (A.L.-D.); joaquin.gonzalez@uam.es (J.G.-R.)

* Correspondence: daniel.ramos@uam.es; Tel.: +34-91-497-6206

Received: 22 February 2018; Accepted: 18 March 2018; Published: 20 March 2018

Abstract: In this work, we analyze the cross-entropy function, widely used in classifiers both as a performance measure and as an optimization objective. We contextualize cross-entropy in the light of Bayesian decision theory, the formal probabilistic framework for making decisions, and we thoroughly analyze its motivation, meaning and interpretation from an information-theoretical point of view. In this sense, this article presents several contributions: First, we explicitly analyze the contribution to cross-entropy of (i) prior knowledge; and (ii) the value of the features in the form of a likelihood ratio. Second, we introduce a decomposition of cross-entropy into two components: discrimination and calibration. This decomposition enables the measurement of different performance aspects of a classifier in a more precise way; and justifies previously reported strategies to obtain reliable probabilities by means of the calibration of the output of a discriminating classifier. Third, we give different information-theoretical interpretations of cross-entropy, which can be useful in different application scenarios, and which are related to the concept of reference probabilities. Fourth, we present an analysis tool, the Empirical Cross-Entropy (ECE) plot, a compact representation of cross-entropy and its aforementioned decomposition. We show the power of ECE plots, as compared to other classical performance representations, in two diverse experimental examples: a speaker verification system, and a forensic case where some glass findings are present.

Keywords: Bayesian; cross-entropy; probabilistic; classifier; discrimination; calibration; ECE plot

1. Introduction

Probabilistic approaches for data mining, machine learning and pattern recognition have proven their effectiveness both theoretically and practically in multiple applications [1]. As a consequence, one of the classical interests among the machine learning community has been to obtain reliable probabilities from a classifier [2–5]. More recently, this interest has been related to the active field of Deep Neural Networks [6]. We will call probabilistic classifiers to those classifiers that are designed to output probabilities.

The use of probabilistic classifiers presents many advantages against more deterministic outputs (e.g., class labels or non-probabilistic scores). First, reliable probabilistic classifiers can be easily compared because their outputs lay in the same domain. Second, it is much more straightforward to combine classifiers if they can be integrated into a theoretically sound probabilistic domain [7]. Third, a probabilistic classifier can be incorporated into a more complex model considering multiple sources of information, by the use of e.g., probabilistic graphical models [8]. Finally, probabilities are interpretable, and therefore probabilistic classifiers admit interpretation in many different domains where an end-user that operates the system as a black box wants to effectively use and understand the information it

gives. Moreover, probabilistic outputs of systems have proven to be useful in many other research and application areas such as clinical decision support systems [9], cognitive psychology [10,11], biometric systems [12–14], weather forecasting [15] and forensic science [16,17]. In all those areas, as well as it happens with classifiers in general, Bayesian decision theory [1] constitutes the formal framework to make optimal choices of courses of action.

In order to be useful, probabilities have to distinguish between different classes, a property that has been dubbed as refinement [15,18], sharpness [19] or discrimination [13,16]. However, it is not enough: probabilities must also present a good calibration [18,20]. According to this, a better-calibrated classifier would output more reliable probabilities, and this would lead to better decisions [4,13,16]. Calibration has been measured in the past by the use of proper scoring rules [18,21,22], i.e., functions that assess the performance or goodness of probabilities. One typical example of performance metrics based on proper scoring rules is the cross-entropy function, mainly due to its information-theoretical interpretation, its good mathematical behavior and its advantageous properties [13,23].

In this article, we deeply review the cross-entropy function, as a metric of performance and as a common objective function for probabilistic classifiers. In order to focus the topic, we restrict ourselves to two-class (or binary) classifiers in a Bayesian decision scenario. Then, this article contributes in several ways, summarized here:

- We analyze the contribution to cross-entropy of two sources of information: the prior knowledge about the classes, and the value of the observations expressed as a likelihood ratio. Apart from its advantages for general classifiers [13], this analysis is of particular interest in many applications such as forensic science [17,24], where prior probabilities and likelihood ratios are computed by different agents, with different responsibilities in the decision process. Moreover, this prior-dependent analysis allows system designers to work on the likelihood ratio computed by the classifier without having to be focused on particular prior probabilities or decision costs. This way of designing classifiers has been referred to as *application-independent* [12,13]. To the best of our knowledge, there are not many works in the literature offering tools to analyze the cross-entropy function under this dichotomy.
- We introduce a decomposition of cross-entropy into two additive components: discrimination loss and calibration loss, showing its information-theoretical properties. A better discrimination component of cross-entropy will indicate a classifier that separates the classes more effectively; and a better calibration component will indicate more reliable probabilities [18]. Moreover, the decomposition enables improving calibration without changing discrimination, by means of an invertible function, theoretically justifying approaches as in [2,3,5,12,13].
- We propose several interpretations of cross-entropy that offer useful communication tools to present results of systems performance to non-expert end users, which can be of great utility for many applications such as forensic science [24].
- We generalize the use of an analysis tool, the Empirical Cross-Entropy (ECE) plot, previously used in forensic science [16,25] for two-class probabilistic classifiers. We demonstrate its power in two different experimental examples.

The article is organized as follows. For the sake of clarity, in Section 2, we give some definitions and examples, and we briefly recall Bayesian decision theory. In Section 3, we review the important concept of calibration of probabilities. We define and deeply analyze cross-entropy in Section 4. Section 5 then introduces the Empirical Cross-Entropy (ECE) plot, as a compact representation to analyze cross-entropy. In Section 6, we will show the usefulness and power of the ECE plot with respect to other classical performance measures, where two classification approaches are explored: speaker verification and a forensic case involving glass findings. Finally, a discussion and some conclusions can be found in Section 7.

2. Review of Bayesian Decision Theory

2.1. Definitions, Notation and Examples

From now on, we will use capital letters to denote random variables (e.g., X) and small letters for particular values observed from those variables (e.g., $X = x$, or simply x). Random (multivariate) vectors will be denoted in capital bold (e.g., $\mathbf{X}$) and small bold letters will denote particular vectors observed from them (e.g., $\mathbf{X} = \mathbf{x}$, or simply $\mathbf{x}$). Probabilities will be referred to as $P(\cdot)$, and conditional probabilities as $P(\cdot|\cdot)$. Probability densities will be referred to as $p(\cdot)$, and $p(\cdot|\cdot)$ if they are conditional. Alternative probabilistic assignments will use the same notation, but replacing P with $\tilde{P}$, p with $\tilde{p}$, and so forth. Related to this notation, we may simply denote P and $\tilde{P}$ as different ways of assigning probabilities.

The classification problem is better illustrated by some examples:

Example 1. ***Speaker Verification****. In this example, the task is to decide whether a speaker is or not a genuine user of a given access control system. Imagine that a person wants to access a resource from a bank (it could be a bank account, or simply some personalized publicity). The telephone application of the bank has an authentication facility that includes voice biometrics. In order to decide what to do with this access attempt, a classifier must compare the speech of that person with some speech that has been previously stored and processed in the application, whose identity is known and claimed by the speaker. Several sources of information must be considered by the classifier:*

- Speaker recognition techniques [26] are used to perform a comparison between the speech of the person who tries to access the system and the speech stored in the system for that particular identity claimed. If the system yields a higher support that the person has the claimed identity, it will increase the chances of a successful access, and vice-versa.
- In order to make the decision, the speaker verification system must take also into account the prior probability, i.e., whether it is probable or not that an impostor attempt will happen for this particular application.
- Moreover, the bank policy should evaluate the consequences of an error when it occurs. Perhaps the application gives access to very sensitive information (e.g., operations with bank accounts), and then avoiding false attempts is critical, or, perhaps, the application yields access to personalized publicity for that particular client, and false rejections are to be avoided.

Example 2. ***Forensic case involving glass findings****. In this example, imagine the role of a trier of fact (a judge or a jury) in a legal trial. There is a suspect that is accused of some crime, and it is necessary to evaluate whether the suspect was involved in a burglary. Findings related to that question include some glass pieces found in the suspect's clothes, which could come from the window broken at the scene of the crime during the burglary. In order to evaluate those findings, the trier of fact considers two possible hypotheses: on the one hand, the glass found in the suspect's clothes and the glass in the window of the scene of the crime comes from the same source. On the other hand, they come from different sources (although it is not always the case, for simplicity, we will assume that the trier of fact establishes that any other hypothesis is so unlikely that its probability is negligible. Moreover, these hypotheses are known as* source-generic*, and they are simpler to understand, although other more complex hypotheses are much more typical in a forensic case. However, we believe that these two assumed simplifications will help in understanding the proposed approach). An ideal trier of fact, i.e., one who correctly uses Bayesian decision theory, should consider the following:*

- The report of a forensic examiner, who compares the glass in the burglary window with the glass in the clothes of the suspect. She or he uses analytical chemistry techniques for this, as well as forensic statistic models, in order to assign a quantitative value to those findings in the form of a likelihood ratio, according to recommendations from forensic institutions worldwide [24].

- In order to make the decision, the trier of fact must also take into account the prior probability that the glass in the suspect's clothes could be originated by the window at the scene of the crime, even before the findings are analyzed by the forensic examiner. Perhaps the suspect was arrested next to the house at the time of the burglary, and/or witnesses have seen her or him smashing the window, and/or she or he has been typically arrested in the past for similar crimes, etc., in which case the prior probability should increase. However, the suspect could present a convincing alibi, and/or witnesses could testify that they saw other people smashing the window, etc., and in those cases the prior probability should be low.
- Moreover, justice systems around the globe have legal standards and policies that influence decisions of triers of fact. In modern, advanced democracies, it is typical that a presumption of innocence is always respected, meaning that condemning a person must be supported by solid evidence. In this sense, it is naturally much more critical not to imprison an innocent person, even though this means that false acquittals may then be more probable. Thus, a trier of fact will only condemn a suspect if the probability that the suspect committed the crime is very high.

Here, we define the elements of a classification problem, contextualizing them into the described examples (speaker verification, and forensic case involving glass findings).

- Classification categories will be referred to as θ_1 and θ_2, and they will be assumed to be observed from a random variable Θ. It is assumed that both classes are complementary in probabilistic terms.
  - In the speaker verification example, θ_1 stands for *the person accessing is who she or he claims to be*, and θ_2 stands for *the person accessing is an impostor*.
  - In the forensic case example, θ_1 stands for *the glass of the suspect's clothes comes from the window at the crime scene*, and θ_2 stands for *the glass of the suspect's clothes does not come from the window at the crime scene*.
- The features observed in the classification problem will be assumed to be multivariate, and referred to as $\mathbf{x}$ generated by $\mathbf{X}$.
  - In the speaker verification example, $\mathbf{x} \equiv \{\mathbf{x}_a, \mathbf{x}_{id}\}$, where $\mathbf{x}_a$ are the speech features extracted from the utterance spoken by the person attempting to enter the system, and $\mathbf{x}_{id}$ are the features already stored in the system, which are known to come from the claimed identity.
  - In the forensic case example, $\mathbf{x} \equiv \{\mathbf{x}_r, \mathbf{x}_c\}$, where $\mathbf{x}_r$ are the chemical features extracted from the glass fragments *recovered* in the suspect's clothes, and $\mathbf{x}_c$ are the chemical features extracted from the window at the scene of the crime, known as *control* glass.
- The action of deciding θ_i will be denoted as α_i.
  - In the speaker verification example, α_1 means that the system decides to accept the speaker, and α_2 means deciding to reject the speaker.
  - In the forensic case example, α_1 means that the trier of fact decides that the suspect glass came from the window at the scene of the crime, and α_2 means deciding that the suspect glass came from a different source than the window at the crime scene.
- Decision costs will be referred to as $C(\alpha_i, \theta_j)$, where α_i is the decision made, but θ_j is the actual category to which a particular feature vector $\mathbf{x}$ belongs. Without loss of generality, costs are assumed to be non-negative. In addition, it is typically assumed that $C(\alpha_i, \theta_i) = 0$, i.e., right decisions are costless.
  - In the speaker verification example, the costs are defined depending on the risk policy of the bank, and depending on the application. For instance, if we talk about access to personalized

publicity, perhaps $C(\alpha_1, \theta_2) = 1$ and $C(\alpha_2, \theta_1) = 10$, being then flexible with false acceptances, but rigorous with false rejections. On the other hand, if the application involves accessing sensitive bank account data, perhaps the priority is to avoid false acceptances, and therefore we can set $C(\alpha_1, \theta_2) = 50$ and $C(\alpha_2, \theta_1) = 1$, for example.

– In the forensic case example, costs are typically designed to avoid false condemns, even though it means that false acquittals are more frequent. Thus, a possible selection of decision costs could be $C(\alpha_1, \theta_2) = 100$ and $C(\alpha_2, \theta_1) = 1$. In any case, it would correspond to the trier of fact to establish the values of the costs.

2.2. Optimal Bayesian Decisions

Bayesian decision theory leads to a well-known decision rule, expressed as:

$$\text{Decide } \alpha_1 \text{ iff: } \quad LR_{(2)}^{(1)} > \frac{C(\alpha_1, \theta_2)}{C(\alpha_2, \theta_1)} \frac{P(\theta_2)}{P(\theta_1)} = \tau_B, \quad \text{otherwise decide } \alpha_2, \tag{1}$$

where τ_B is the so-called Bayes threshold, and the likelihood ratio (LR) is defined as:

$$LR_{(2)}^{(1)} = \frac{p(\mathbf{x}|\theta_1)}{p(\mathbf{x}|\theta_2)}. \tag{2}$$

Equation (1) gives the general rule to make decisions if the following quantities are known:

- The likelihood ratio $LR_{(2)}^{(1)}$, expressing the value of the observation of $\mathbf{x}$ in support of each of the two classes θ_1 and θ_2.
- The prior probabilities for both classes, namely $P(\theta_1)$ and $P(\theta_2)$.
- The costs associated to each action α_i towards deciding a class θ_i, given that class θ_j is actually true, namely $C(\alpha_i, \theta_j)\ \forall i \neq j$.

Figure 1 illustrates this decision scheme graphically.

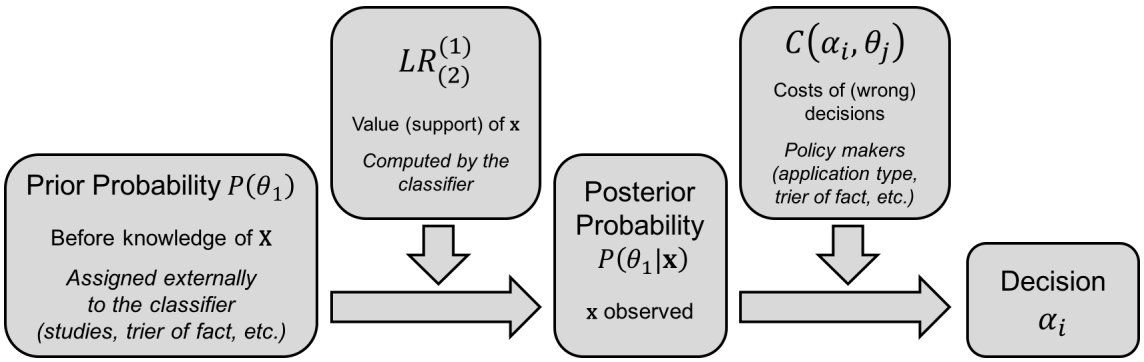


Figure 1. Decision scheme of a probabilistic classifier.

Additional insights can be obtained from the so-called *odds form* of the Bayes theorem:

$$\frac{P(\theta_1 \mid \mathbf{x})}{P(\theta_2 \mid \mathbf{x})} = LR_{(2)}^{(1)} \cdot \frac{P(\theta_1)}{P(\theta_2)}, \tag{3}$$

where the odds are defined as the quotient of complementary probabilities:

$$O(\cdot) = \frac{P(\cdot)}{1 - P(\cdot)} \tag{4}$$

and the following relation can be easily derived:

$$P\left(\theta_1|\,\mathbf{x}\right) = \frac{LR_{(2)}^{(1)} \cdot O\left(\theta_1\right)}{1 + LR_{(2)}^{(1)} \cdot O\left(\theta_1\right)}. \tag{5}$$

Moreover, we define $|\log(LR_{(2)}^{(1)})|$ as the *strength* of the support of the features (unless explicitly stated otherwise, logarithms are assumed to be natural, i.e. base-e).

2.3. Empirical Performance of Probabilistic Classifiers

Empirical performance measurement involves the use of a database where ground-truth labels are available. For instance:

- In speaker verification, imagine that we have a black-box speaker recognition system that receives two sets of speech files, where all utterances in each of the two sets belong to a given putative speaker. The system compares both sets and outputs a likelihood ratio $LR_{(2)}^{(1)}$, for θ_1 (same speaker), and θ_2 (different speakers). With a database of speech utterances, each one with a label indicating an index of speaker identity, we compare them according to a given protocol to generate LR values. The ground-truth labels of the experimental set of LR values are same-speaker ($\Theta = \theta_1$) or different-speakers ($\Theta = \theta_2$) labels.
- In forensic interpretation of glass samples, we can work analogously with a black-box glass interpretation model yielding likelihood ratios, and a database of feature vectors measured from glass objects, each vector with a label indicating a different index for each glass object. Thus, LR values generated are accompanied by ground-truth labels like same-source ($\Theta = \theta_1$) and different-source ($\Theta = \theta_2$).

The performance of such experimental set of LR values (with ground-truth labels) can then be measured as an expected cost:

$$\begin{aligned} E_{P(\alpha_i,\theta_j)}\left[C\left(\alpha_i,\theta_j\right)\right] &= P\left(\alpha_2,\theta_1\right) C\left(\alpha_2,\theta_1\right) + P\left(\alpha_1,\theta_2\right) C\left(\alpha_1,\theta_2\right) \\ &= P\left(\alpha_2|\theta_1\right) P\left(\theta_1\right) C\left(\alpha_2,\theta_1\right) + P\left(\alpha_1|\theta_2\right) P\left(\theta_2\right) C\left(\alpha_1,\theta_2\right), \end{aligned} \tag{6}$$

where $P\left(\alpha_i|\theta_j\right)$ is the probability of deciding action α_i when θ_j is the actual class of $\mathbf{x}$. If $i \neq j$, these probabilities are known as error probabilities. For our speaker recognition example, $P\left(\alpha_1|\theta_2\right)$ is known as *false acceptance probability*, and $P\left(\alpha_2|\theta_1\right)$ as *false rejection probability*. For the glass case examples, $P\left(\alpha_1|\theta_2\right)$ could be known as *false association probability*, and $P\left(\alpha_2|\theta_1\right)$ as *false exclusion probability*. Those error probabilities depend on the decision made with Bayes threshold τ_B (Equation (1)), for which prior probabilities and costs must be known.

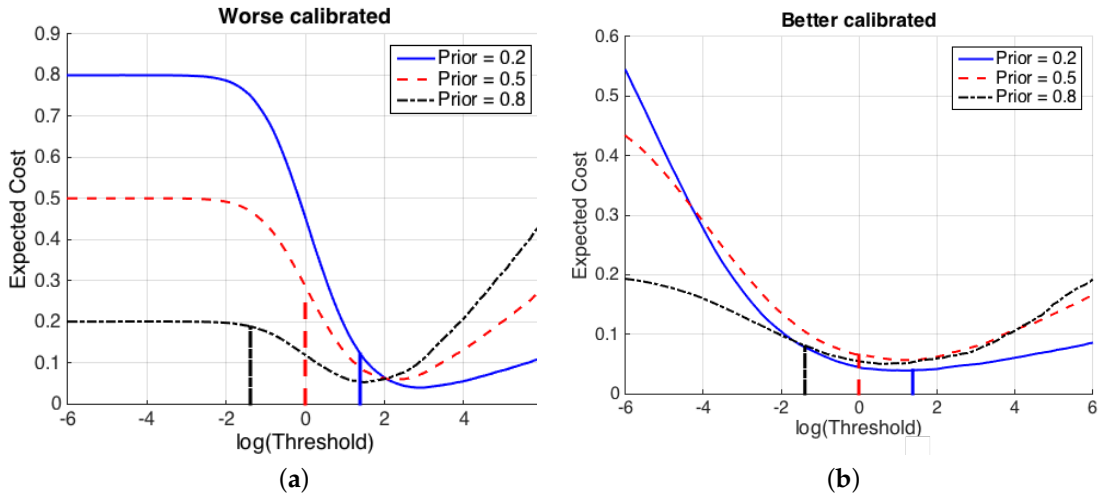
3. Calibration of Probabilities

Bayesian decision theory is based on the fact that, for the $LR_{(2)}^{(1)}$, the probabilities $p\left(\mathbf{x}|\theta_1\right)$ and $p\left(\mathbf{x}|\theta_2\right)$ are known to have generated $\mathbf{x}$. However, this is rarely the case for real applications, except for e.g., some simulated scenarios. Therefore, typically an alternative value $\tilde{LR}_{(2)}^{(1)}$ will be computed instead, as a result of directly computing the ratio in a discriminative way, or by assigning likelihoods to the data as $\tilde{p}\left(\mathbf{x}|\theta_1\right)$ and $\tilde{p}\left(\mathbf{x}|\theta_2\right)$. This will inevitably lead to suboptimal decisions in terms of expected cost because the decision rule in Equation (1) is assuming $LR_{(2)}^{(1)}$, not $\tilde{LR}_{(2)}^{(1)}$.

To address this problem, the designer of a probabilistic classifier might be tempted to keep the $\tilde{LR}_{(2)}^{(1)}$ values, and then to change the decision threshold τ from its theoretically optimal value τ_B. Thus, changing the decision threshold to τ^* will lead to different values of $P\left(\alpha_i|\theta_j\right)\ \forall i \neq j$, which can optimize the expected cost in Equation (6). Of course, it might be the case that $\tau^* \neq \tau_B$.

However, our objective as designers of a probabilistic classifier should not be choosing τ^*. For instance, in some applications, we could be even not aware of the values of the priors and decision costs, and therefore we could not compute the expected cost in order to get to τ^*. Conversely, our aim as designers is to compute values of $\tilde{LR}_{(2)}^{(1)}$ that are optimal when threshold τ_B is set according to the values of the prior probabilities and the decision costs.

Figure 2 illustrates the effects of a non-proper computation of $\tilde{LR}_{(2)}^{(1)}$ by a classifier. Two different speaker recognition systems that compute likelihood ratios for the features $\mathbf{x}$ are shown, and in both cases the value of the expected cost is represented for a range of thresholds $\log(\tau)$. For illustration, it is assumed that the costs of wrong decisions are both 1. Figure 2a shows a system computing $\tilde{LR}_{a\,(2)}^{(1)}$ values. Bayesian thresholds (τ_B), for each of several values of the prior probabilities, are represented as vertical lines. It is clearly observed that, for $\tilde{LR}_{a\,(2)}^{(1)}$, selecting the Bayesian threshold τ_B leads to a suboptimal value of the expected cost for all prior probabilities. However, Figure 2b shows a system where $\tilde{LR}_{b\,(2)}^{(1)}$ values have been computed, and it is shown that τ_B is near the optimum for all the represented values of $P(\theta_1)$. This is because the calibration of the likelihood ratios computed by $\tilde{LR}_{b\,(2)}^{(1)}$ is much better than for $\tilde{LR}_{a\,(2)}^{(1)}$. This property of calibration is of great importance in probabilistic classifiers, and it is described below.



(**a**) (**b**)

Figure 2. Value of the expected cost (Equation (6)) of two different sets of LR values from a speaker recognition system presenting (**a**) worse calibration ($\tilde{LR}_{a\,(2)}^{(1)}$ values), and (**b**) better calibration ($\tilde{LR}_{b\,(2)}^{(1)}$ values). Costs of erroneous decisions are set to 1 for simplicity. The x-axis shows possible values of a decision threshold $\log\tau$, and the logarithms of Bayes thresholds τ_B (Equation (1)) are shown as vertical lines.

Calibration and Discrimination of Posterior Probabilities

The concept of calibration of probabilities is not new in the statistics literature. In [18], it was introduced in order to evaluate so-called *forecasts*, which are indeed posterior probabilities of a given hypothesis (tomorrow, it will rain) elicited by a weather forecaster. This problem is equivalent to probabilistic binary classification as proposed here, where the analogous of the forecast is $P(\theta_1|\mathbf{x})$.

In [18,27], the accuracy of such a forecaster is assessed by means of strictly proper scoring rules. One example is the logarithmic scoring rule:

$$\begin{aligned} \theta_1 \text{ true}: &\ -\log_2\left(P\left(\theta_1|\,\mathbf{x}\right)\right), \\ \theta_2 \text{ true}: &\ -\log_2\left(P\left(\theta_2|\,\mathbf{x}\right)\right). \end{aligned} \tag{7}$$

Thus, strictly proper scoring rules may be seen as loss functions that assign a penalty to a given value of the posterior probability depending on the true value of the ground-truth label (see [22] for more examples of strictly proper scoring rules). The logarithmic scoring rule is illustrated in Figure 3.

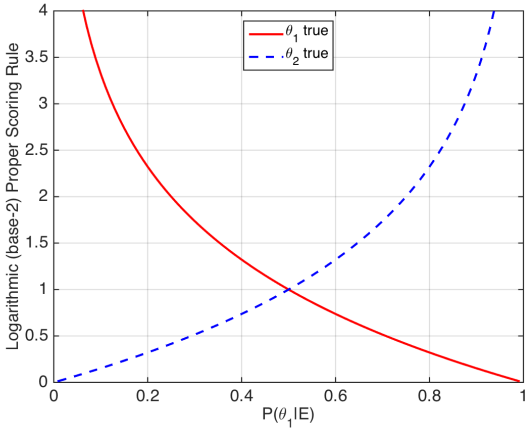


Figure 3. Logarithmic scoring rule. The x-axis represents the posterior probability of θ_1. The rule can also be applied to prior probabilities $P(\theta_1)$.

Strictly proper scoring rules have the following interesting properties:

- In our speaker verification example, a posterior probability $\tilde{P}(\theta_1|\mathbf{x})$ is obtained. We can define the mean value of the proper scoring rule with respect to a *reference* probability distribution $P(\theta_1|\mathbf{x})$, according to [18]. This reference probability can be viewed as a desired value P of the posterior probability distribution, to which the actual posterior probability distribution $\tilde{P}$ is compared. For the logarithmic scoring rule:

$$-P(\theta_1|\mathbf{x}) \cdot \log_2\left(\tilde{P}(\theta_1|\mathbf{x})\right) - (1 - P(\theta_1|\mathbf{x})) \cdot \log_2\left(1 - \tilde{P}(\theta_1|\mathbf{x})\right). \tag{8}$$

  By definition, the mean value of a strictly proper scoring rule is minimized if and only if $P(\theta_1|\mathbf{x}) = \tilde{P}(\theta_1|\mathbf{x})$ (for instance, this is easy to prove for the logarithmic rule by simply deriving Equation (8) with respect to $\tilde{P}(\theta_1|\mathbf{x})$). In other words, if a classifier aims at minimizing a strictly proper scoring rule, it should yield a likelihood ratio value as close as possible to $LR^{(1)}_{(2)}$, which will lead to the desired probability $P(\theta_1|\mathbf{x})$.

- In [18], the overall measure of goodness of an empirical set of posterior probabilities is defined as the empirical average of a strictly proper scoring rule. An example is the logarithmic score (LS):

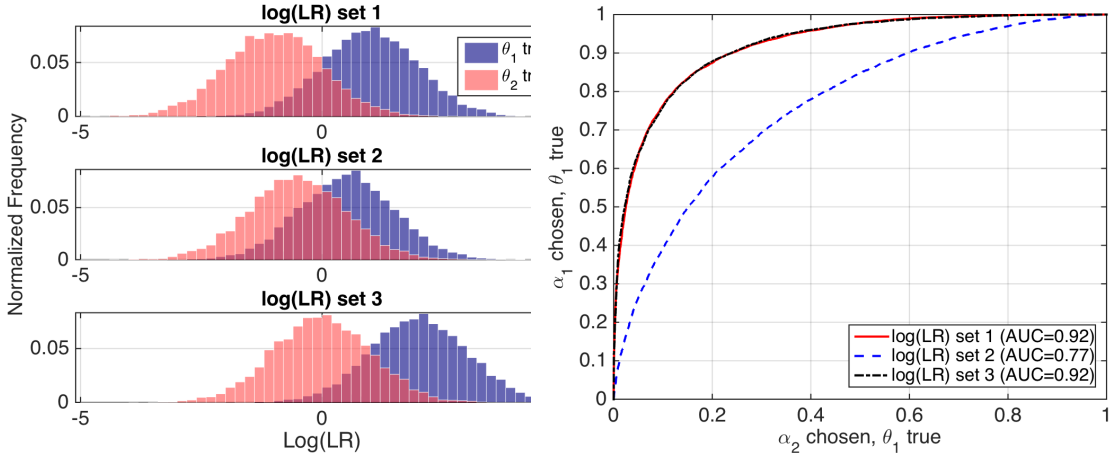
$$\begin{aligned} LS = \; & -\frac{1}{N_1} \sum_{\mathbf{x}:\theta_1 \text{ true}} \log_2 \tilde{P}(\theta_1|\mathbf{x}) \\ & -\frac{1}{N_2} \sum_{\mathbf{x}:\theta_2 \text{ true}} \log_2 \tilde{P}(\theta_2|\mathbf{x}), \end{aligned} \tag{9}$$

  where N_1 and N_2 are the number of comparisons where θ_1 or θ_2 are respectively true. Thus, LS is an overall loss. Moreover, it is also demonstrated in [18] that such a measure of accuracy can be divided into two components:

  1. A *calibration loss* component, which measures how similar the posterior probabilities are to the frequency of occurrence of θ_1. Low calibration loss means that, for a given range of values of $\tilde{P}(\theta_1|\mathbf{x})$ closely around a value k, the frequency of cases where $\Theta = \theta_1$ tends to be k.

2. A *refinement loss*, also known as *sharpness loss* or *discrimination loss*, component. It measures how sharp or how spread the posterior probabilities are. Roughly speaking, lower refinement loss means that, if the calibration loss is low, $\tilde{P}(\theta_1|\mathbf{x})$ will tend to be closer either to 0 or to 1, on average.

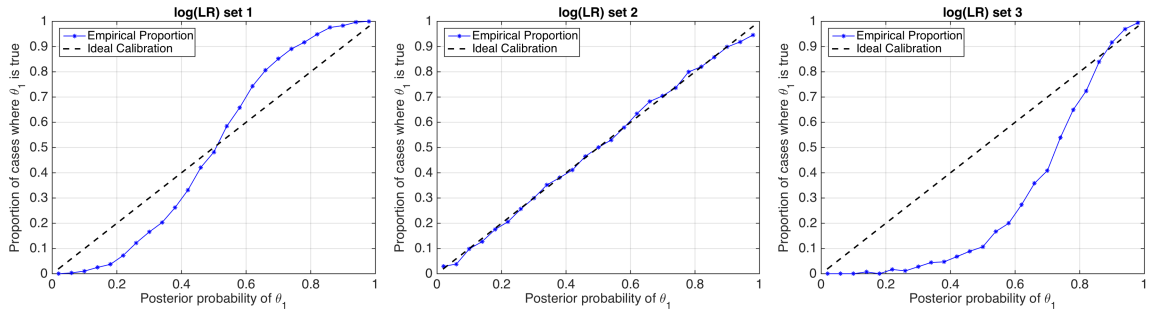
Refinement loss can be seen as discrimination performance, as measured by typical performance representations such as Receiver Operation Characteristic (ROC) curves, or values such as Area Under ROC Curve (AUC). For instance, if we consider our speaker verification example, for a fixed value of the prior probabilities, the refinement, sharpness or discrimination of a set of likelihood ratios will be better if the $\tilde{LR}_{(2)}^{(1)}$ values of the system overlap less when θ_1 and θ_2 are respectively true, and so the ROC and AUC measures will be better. An illustrating example is given in Figure 4.



(**a**) (**b**)

Figure 4. (**a**) histograms of $\log(LR)$ values for three simulated examples; and (**b**) their corresponding ROC curves, with AUC values.

A good calibration loss means that posterior probabilities actually represent the real empirical proportion of occurrence of each hypothesis. A popular technique to measure calibration involves the use of so-called empirical calibration plots [20], also known as reliability plots [3], where the frequency of occurrence of a given hypothesis θ_1 is represented against a binning of the posterior probabilities given by the system. Empirical calibration plots of simulated examples in Figure 4 are depicted in Figure 5.



(**a**) (**b**) (**c**)

Figure 5. Reliability plots of the simulated sets of $\log(LR)$ values shown in Figure 4. (**a**) $\log(LR)$ set 1; (**b**) $\log(LR)$ set 2; (**c**) $\log(LR)$ set 3.

The meaning of a good calibration loss is related to the way in which the information is presented to the decision maker, leading her or him to good decisions. This means that, if posterior probabilities are better calibrated, using the τ_B threshold to the likelihood ratio will yield closer-to-minimum expected cost [4,13].

4. Cross-Entropy: An Information-Theoretical Performance Measure

Information theory [28,29] states that the information obtained in an inferential process is determined by the reduction of the entropy, which measures the uncertainty about a given unknown variable in the light of the available knowledge. In our classification problem, the entropy represents the uncertainty that the decision maker has about the actual value of the hypothesis variable $\Theta \in \{\theta_1, \theta_2\}$.

In a given classification problem, even though the features are not observed yet, the uncertainty about Θ is given by the prior probabilities $P(\theta_1) = 1 - P(\theta_2)$. With this available knowledge, the entropy of the hypothesis, namely prior entropy or entropy of the prior, is determined by the following expression [29]:

$$H_P(\Theta) = -\sum_{i\in\{1,2\}} P(\theta_i) \log_2 P(\theta_i). \tag{10}$$

Once the features $\mathbf{x}$ are observed in a given single classification comparison, this may or may not reduce the uncertainty about Θ. However, it can be proven [29] that the expected value of the entropy of the posterior probability over all possible values of the features $\mathbf{x}$ cannot be greater than the prior entropy. This expected value is a conditional entropy, which will be denoted as a posterior entropy, computed as [29]:

$$H_P(\Theta|\mathbf{X}) = -\sum_{i\in\{1,2\}} \int_{\mathbf{x}} p(\theta_i, \mathbf{x}) \log_2 P(\theta_i|\mathbf{x})\, d\mathbf{x} \tag{11}$$

$$= -\sum_{i\in\{1,2\}} P(\theta_i) \int_{\mathbf{x}} p(\mathbf{x}|\theta_i) \log_2 P(\theta_i|\mathbf{x})\, d\mathbf{x}, \tag{12}$$

where the features $\mathbf{x}$ are integrated over their whole domain. As mentioned, $H_P(\Theta|\mathbf{X}) \leq H_P(\Theta)$ [29]. However, the computation of Equation (12) is usually non-practical, as it requires knowing the likelihoods $p(\mathbf{x}|\theta_i)$, which are not known in general.

This can be solved by comparing the posterior probabilities computed by the classifier with a reference probability distribution. The letter $\tilde{P}$ ($\tilde{p}$ for pdfs) will denote probabilities obtained using the classifier, and the letter P (p for pdfs) will denote those reference probabilities. Moreover, the expression $\tilde{LR}^{(1)}_{(2)}$ will denote the likelihood ratio computed by the classifier, a function of $\mathbf{x}$. Thus, $\tilde{LR}^{(1)}_{(2)}$ is the quotient between $\tilde{p}(\mathbf{x}|\theta_1)$ and $\tilde{p}(\mathbf{x}|\theta_2)$. The incorporation of the reference probabilities eliminates the dependence between $\tilde{P}(\theta_i|\mathbf{x})$ and $p(\mathbf{x}|\theta_i)$ in Equation (12), leading to the expression of the cross-entropy:

$$H_{P\|\tilde{P}}(\Theta|\mathbf{X}) = -\sum_{i\in\{1,2\}} P(\theta_i) \int_{\mathbf{x}} p(\mathbf{x}|\theta_i) \log_2 \tilde{P}(\theta_i|\mathbf{x})\, d\mathbf{x}. \tag{13}$$

It can be easily proven that the cross-entropy (Equation (13)) is decomposed into:

$$H_{P\|\tilde{P}}(\Theta|\mathbf{X}) = H_P(\Theta|\mathbf{X}) + D_{P\|\tilde{P}}(\Theta|\mathbf{X}), \tag{14}$$

where $D_{P\|\tilde{P}}(\Theta|\mathbf{X})$ is the well-known Kullback–Leibler (KL) divergence between the system's posterior distribution and the reference distribution [29] for all possible values of the evidence, defined as:

$$D_{P\|\tilde{P}}(\Theta|\mathbf{X}) = \sum_{i\in\{1,2\}} P(\theta_i) \int_{\mathbf{x}} p(\mathbf{x}|\theta_i) \log_2 \frac{P(\theta_i|\mathbf{x})}{\tilde{P}(\theta_i|\mathbf{x})} d\mathbf{x}. \tag{15}$$

Thus, the cross-entropy measures the complementary effect of two different magnitudes:

- $H_P(\Theta|\mathbf{X})$, the posterior entropy of the reference probability, which measures the uncertainty about the hypotheses if the reference probability distribution is used.
- $D_{P\|\tilde{P}}(\Theta|\,\mathbf{X})$, the divergence of the classifier posterior $\tilde{P}$ from the reference posterior P. This is an additional information loss because it was expected that the system computed P, not $\tilde{P}$.

4.1. Proposed Measure of Accuracy: Empirical Cross-Entropy (ECE)

An empirical approximation of cross-entropy (Equation (13)) is proposed here. With an empirical set of LR values and a fixed prior probability $P(\theta_1)$, we can approximate Equation (13) as follows:

$$\begin{aligned} H_{P\|\tilde{P}}(\Theta|\,\mathbf{X}) \simeq ECE = & -\frac{P(\theta_1)}{N_1}\sum_{\mathbf{x}:\theta_1 \text{ true}}\log_2\tilde{P}(\theta_1|\,\mathbf{x}) \\ & -\frac{P(\theta_2)}{N_2}\sum_{\mathbf{x}:\theta_2 \text{ true}}\log_2\tilde{P}(\theta_2|\,\mathbf{x}), \end{aligned} \tag{16}$$

where ECE stands for Empirical Cross-Entropy. Then, by Equation (5):

$$\begin{aligned} ECE = & \frac{P(\theta_1)}{N_1}\sum_{\mathbf{x}:\theta_1 \text{ true}}\log_2\left(1+\frac{1}{\tilde{LR}^{(1)}_{(2)}\cdot\frac{P(\theta_1)}{P(\theta_2)}}\right) \\ & +\frac{P(\theta_2)}{N_2}\sum_{\mathbf{x}:\theta_2 \text{ true}}\log_2\left(1+\tilde{LR}^{(1)}_{(2)}\cdot\frac{P(\theta_1)}{P(\theta_2)}\right), \end{aligned} \tag{17}$$

where $\tilde{LR}^{(1)}_{(2)}$ is dependent of $\mathbf{x}$, but we simplified the notation.

Figure 6 illustrates the information loss measured by cross-entropy in terms of its decomposition (Equation (14)). There, ellipses represent uncertainty, which can be viewed also as an information loss. On the left, the prior entropy $H_P(\Theta)$ is represented as a green ellipse, whereas, on the right, the cross-entropy $H_{P\|\tilde{P}}(\Theta|\,\mathbf{X})$ is represented, decomposed as $D_{P\|\tilde{P}}(\Theta|\,\mathbf{X})$ (blue area) plus $H_P(\Theta|\mathbf{X})$ (green area). The diagram illustrates that the observation of $\mathbf{X}$ does not increase the uncertainty about Θ, i.e., $H_P(\Theta) \geq H_P(\Theta|\mathbf{X})$. In other words, we will always gain information about Θ by the observation of $\mathbf{X}$, but this will only happen if reference probabilities are computed. Otherwise, the KL divergence term will contribute with an additional information loss to the cross-entropy. Thus, the cross-entropy could be arbitrarily larger than the prior entropy because $D_{P\|\tilde{P}}(\Theta|\,\mathbf{X})$ is positive and unbounded. This could lead to a potentially dramatic loss of information if a LR model computes $\tilde{LR}^{(1)}_{(2)}$ values that substantially diverge from the reference $LR^{(1)}_{(2)}$.

As the prior probability is considered a parameter, then $H_P(\Theta) = H_{\tilde{P}}(\Theta)$. In other words, the differences between $P(\theta_1|\mathbf{x})$ and $\tilde{P}(\theta_1|\mathbf{x})$ are assumed to be only because of a different LR, not because of a different prior probability, since we want to evaluate the classifier, and the prior probability is external to it. Therefore, from Equation (17), it is straightforward that the ECE is independent of the reference likelihoods p. This has the following interpretation: for a fixed value of ECE, changing the reference P implies that:

- $H_P(\Theta|\mathbf{X})$ increases (decreases) and
- $D_{P\|\tilde{P}}(\Theta|\,\mathbf{X})$ decreases (increases)

in order to keep ECE constant. This is depicted in Figure 6: the ellipse representing cross-entropy on the right always has the same size. However, the inner small ellipse representing posterior entropy of P may increase or decrease depending on the choice of the reference probability P.

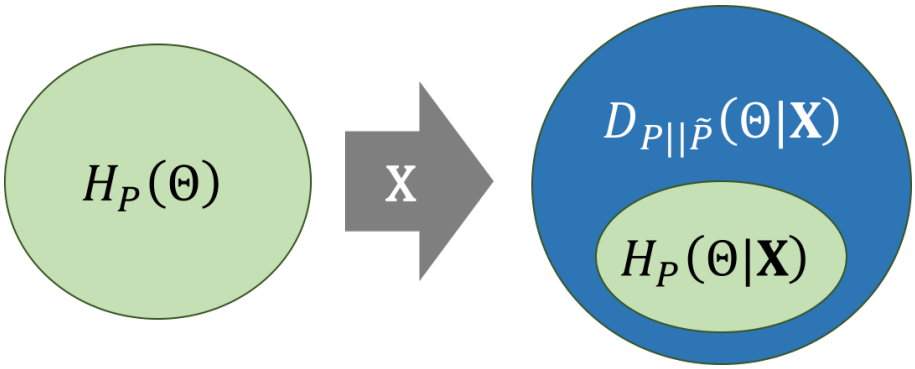


Figure 6. Scheme showing the decomposition of cross-entropy in Equation (14). Ellipses represent uncertainty, which can be viewed as an information loss. The ellipse on the left (green) is the prior entropy $H_P(\Theta)$. The ellipse on the right is the cross-entropy $H_{P||\tilde{P}}(\Theta|\mathbf{X})$ decomposed into its additive components: the posterior entropy $H_P(\Theta|\mathbf{X})$ (green) plus the KL divergence $D_{P||\tilde{P}}(\Theta|\mathbf{X})$ (blue).

4.2. Choosing a Reference Probability Distribution for Intuitive Interpretation

The selection of the reference probability P is constrained because Equation (14) must hold. Sensible choices of P should consider that, for some applications, simplicity and clarity is important, especially if one wants to communicate the results of the performance test to an end-user (like e.g., a trier of fact). In that sense, in this article, we give two proposals of reference P, detailed below.

4.2.1. Oracle Reference P_o

The *oracle* reference distribution is motivated as follows: the aim of every classification problem is finding the true value of the hypothesis Θ. This would be achieved if the decision maker assigns the following posterior probabilities P_o:

$$\begin{aligned} P_o(\theta_1|\mathbf{x}) &= 1\,, \ \theta_1 \text{ is true,} \\ P_o(\theta_1|\mathbf{x}) &= 0\,, \ \theta_2 \text{ is true.} \end{aligned} \tag{18}$$

If this oracle distribution is selected as a reference, then $H_{P_o}(\Theta|\mathbf{X}) = 0$, and therefore the ECE becomes solely the KL-divergence $D_{P_o||\tilde{P}}(\Theta|\mathbf{X})$. Thus, the higher the ECE value is, the higher the amount of information lost by the classifier in order to know the true value of the hypotheses. If the classifier does not yield oracle probabilities, then there will be an information loss. This could be also interpreted as the information needed to get to certain predictions.

4.2.2. PAV-Calibrated Reference P_{cal}

The reference probability P_{cal} will present the same discrimination loss as $\tilde{P}$ (i.e., the same ROC curve and AUC value), but it will be as perfectly calibrated as possible.

Fortunately, there is an algorithm that allows perfect calibration of probabilities, namely the Pool Adjacent Violators algorithm (PAV or PAVA) [30,31]. In fact, Ref. [32] provides a proof that PAV is the monotonic function that achieves the best possible value of a proper scoring rule in an empirical set of posterior probabilities. This also applies to ECE, as a weighted average of a proper scoring rule. The choice of P_{cal} as a reference posterior distribution allows the interpretation of ECE with its two components:

- The discrimination component, namely $ECE_{min} \simeq H_{P_{cal}}(\Theta|\mathbf{X})$, that represents the information loss due to a lack of discriminating power of the classifier.

- The calibration component, namely $ECE_{cal} \simeq D_{P_{cal}\|\tilde{P}}(\Theta|\mathbf{X})$, that represents the information loss due to a lack of calibration of the classifier.

5. The ECE Plot

In this paper, we propose representing ECE as a function of $P(\theta_1)$ in a so-called ECE plot. For each prior probability in a range centered around $P(\theta_1) = P(\theta_2) = 0.5$, posterior probabilities $\tilde{P}(\theta_1|\mathbf{x})$ are obtained using the $\tilde{LR}_{(2)}^{(1)}$ values computed by the classifier. The value of ECE (Equation (17)) is then represented as a function of the log-odds, namely $\log O(\theta_1)$. For the sake of interpretation, the x-axis of the ECE plots represents base-10 logarithms.

Figure 7 shows an example of ECE plots for the simulated sets of LR values used previously in Figures 4 and 5. The solid, red curve is the ECE (average information loss) of the LR values computed by the classifier. The higher this ECE curve, the higher the amount of information needed in order to know the true hypothesis, and therefore the worse the system.

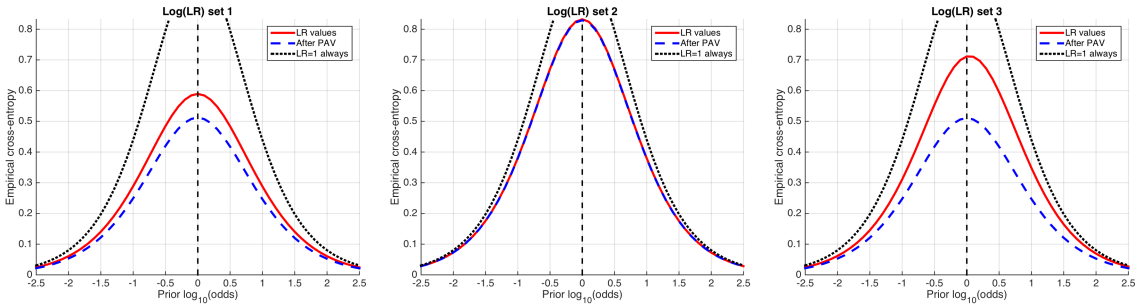


Figure 7. ECE plots of the simulated sets of $\log(LR)$ shown in Figures 4 and 5. (**a**) $\log(LR)$ set 1; (**b**) $\log(LR)$ set 2; (**c**) $\log(LR)$ set 3.

Two other curves are represented in each ECE plot. On the one hand, the dashed, blue curve represents the *best-calibrated* system; for each value of the prior probability, P_{cal} is obtained using the Pool Adjacent Violators (PAV) algorithm. The best-calibrated system can be seen as an optimum of performance in the sense of calibration.

On the other hand, the dotted curve represents the performance of a classifier always delivering $LR = 1$, referred to as a *neutral* system, for which the posterior probability is equal to the prior probability (Equation (3)), and its cross-entropy is simply the prior entropy (Equation (10)). Then, if the ECE value (red, solid curve) of the classifier is greater than the entropy of the neutral system, it will be better not to use the classifier.

As a summarizing measure, ECE at the prior log-odds of 0 can be used. This measure has been already proposed as *log-likelihood-ratio cost* (C_{llr}) in [12,13], where its interpretation in terms of expected cost can be found (as a matter of fact, C_{llr} is the expected cost of decisions, where the expectation is not only made on the action α_i and the true decision θ_j, but also over all possible values of the costs $C(\alpha_i, \theta_j)$. Thus, it is shown that optimizing the ECE value also optimizes the expected cost averaged over all values of decision costs, justifying the relationship between calibration and cost optimization as discussed in Section 3). Therefore, C_{llr} can be also decomposed as ECE does, and $C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$, where C_{llr}^{min} and C_{llr}^{cal} are, respectively, ECE_{min} and ECE_{cal} at the value of prior log-odds equal 0. In fact, the decomposition into discrimination and calibration by PAV was firstly proposed for C_{llr} [13], and generalized to ECE afterwards [25].

Freely available Matlab^{TM} software (release 2015b, The Mathworks inc., US.) to draw ECE plots can be found in http://arantxa.ii.uam.es/~dramos/software.html.

6. Experimental Examples

6.1. Speaker Verification

Here, we test the performance of our speaker verification example in terms of ECE. We use a state-of-the-art speaker recognition system following the approach in [33]. Roughly speaking, the speaker recognition system processes the speech utterances to form feature vectors in a 600-dimensional space, the so-called i-vectors. Then, a likelihood ratio is computed by Probabilistic Linear Discriminant Analysis (PLDA), as in [34].

Supposedly, the PLDA approach should yield likelihood ratios that present good calibration, but it is not typically the case. This can be due to the high dimensionality of the problem compared to the amount of data to compare, the extreme data sparsity, and the simplicity of the model compared to the complexity of the feature space. Therefore, a further calibration transformation is typically applied to the PLDA likelihood ratio, yielding a so-called calibrated likelihood ratio. A popular calibration technique involves a logistic regression model [3,35], also known as *Platt scaling* [2].

Regarding the data used for this setup, we used the framework provided by the American National Institute of Standard and Technology (NIST) for standard benchmark tests, known as Speaker Recognition Evaluations (SRE). In our case, the empirical set used to compute likelihood ratios to test the system is the SRE 2010 condition 5 (telephone, conversational speech in English) [36]. In addition, data from previous SRE and other speech databases were used to train the PLDA model and the i-vector extractor. The equivalent condition 5 of the SRE 2008 dataset has been used to train the logistic regression model for calibration purposes [37]. We restricted the test and calibration datasets to contain only female speech. The numbers of comparisons for class is $N_1 = 3704$ and $N_2 = 233{,}077$, yielding an empirical prior proportion of $\frac{N_1}{N_1+N_2} \simeq 0.016$.

Figure 8 shows the performance of the likelihood ratios computed by the PLDA stage, before and after logistic regression calibration. We show histograms of the values of $\tilde{LR}^{(1)}_{(2)}$; reliability plots of $\tilde{P}(\theta_1|\mathbf{x})$ computed from $\tilde{LR}^{(1)}_{(2)}$ at the empirical prior $P(\theta_1) = 0.016$; ROC curves with AUC values; and the proposed ECE plots for a wide range of prior probabilities.

Histograms in Figure 8 show equal discrimination performance because the degree of overlap between different empirical distributions is the same in both cases. This agrees with the fact that logistic regression is an invertible transformation (a sigmoid for probabilities, meaning a linear transformation for log-odds and $\log(LR)$ values), and, as a consequence, the ordering of the LR values is not altered, and therefore the discrimination performance is not changed either. However, the range of the $\log(LR)$ values is very different before and after logistic regression. This is due to the fact that the logistic regression transformation scales and shifts the histograms aiming to an improvement in the calibration of the LR values. However, the histograms after the application of logistic regression are not completely symmetric over $\log(LR) = 0$, and therefore we can expect some calibration loss. Although histograms are useful to spot tendencies in the LR values, they do not measure performance in an explicit way, and this is their main drawback as a performance representation. In reliability plots in Figure 8b, it is clear that the application of logistic regression dramatically improves calibration. As expected, the same ROC curves shown in Figure 8c indicate the same discrimination.

The previous performance measures present some problems. First, a clear decomposition between discrimination and calibration is not explicitly seen in a single figure. Although reliability plots and ROC curves separately measure both components, they are not comparable measures of discrimination and calibration loss. Finally, reliability plots only show calibration at the empirical prior.

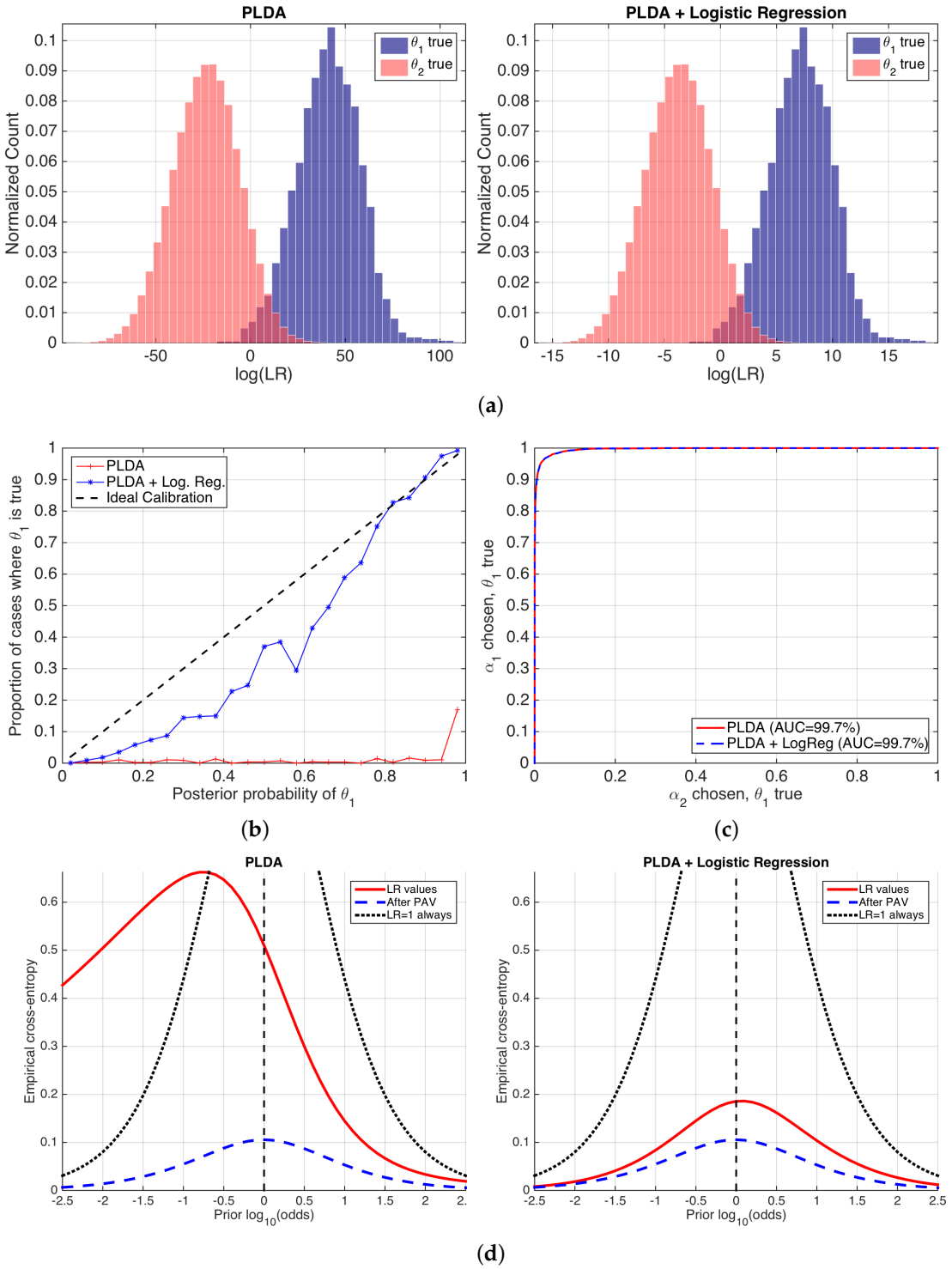


Figure 8. Performance of a speaker recognition system using PLDA, before and after logistic regression calibration. (**a**) histograms of $\log\left(\tilde{LR}^{(1)}_{(2)}\right)$ values; (**b**) reliability plots; (**c**) ROC curves; and (**d**) ECE plots.

ECE plots (Figure 8d) fix these drawbacks of previous performance representations. The performance of the set of LR values is computed for a wide range of prior log-odds (in the x-axis), giving a prediction of system performance for a different variety of possible applications defined by their prior probabilities. Moreover, the ECE curve is compared to the neutral curve (marked as $LR = 1$ *always*), showing the range of prior $\log_{10}$-odds where the classifier is useless. In our example,

Figure 8d shows that, for the PLDA model, the classifier will be better than doing nothing for all the range of log-odds from ca. -0.75 onwards, corresponding to $P(\theta_1) \geq 0.15$. However, it will be useless below this value, and end-users can be recommended not to use it in those applications. ECE plots also show that, after the application of a logistic regression stage, the classifier becomes much better because the red, solid ECE curve dramatically reduces for all the range of prior $\log_{10}$-odds analyzed. This means that the classifier will be informative for a wide range of applications. The latter is an important property, since it shows that the classifier will be useful in practically any application.

ECE plots also show the decomposition between the discrimination and calibration losses in a comparable way. The same ECE_{min} curves (blue, dashed) are equal in both sets of LR values, indicating the same discrimination loss. However, the calibration loss is much higher before than after logistic regression calibration. Nevertheless, logistic regression still presents some calibration loss, explained by the dataset shift between the NIST SRE 2010 testing data, and the NIST SRE 2008 data used to train the logistic regression stage. This further motivates better ways of training the logistic regression calibration stage.

Finally, we give an example of interpretation of ECE: if we use the classifier yielding LR values by PLDA + logistic regression, and the prior probability is $P(\theta_1) = 0.6$ for the given application (meaning prior $\log_{10}$-odds of ca. 0.8), the ECE of the classifier in that case is ca. 0.13. We can argue the following:

- We need 0.13 bits to know the true value of the ground-truth label of a comparison (i.e., we use P_o as the reference).
- The best possible calibrated classifier needs 0.07 bits to know the true value of the ground-truth label of a comparison. However, as our classifier is not so well calibrated, it will need 0.05 bits more (i.e., we use P_{cal} as reference).

In both cases, the values in bits can be also interpreted as minimum number of bits needed to describe the hypothesis variable Θ, according to information theory for data compression [29]. This has the additional insight that the higher the ECE, the more costly it is to represent the truth about the correct hypothesis, and the more information is needed to arrive at it.

6.2. Forensic Case Involving Glass Findings

In this section, two different models to compute likelihood ratios for glass findings are compared by ECE plots. First, a classical approach based on a multivariate model is used, with normal assumptions for the within-source variation of the features and a kernel density function used for the between-source variation, and it will be denoted as the Multivariate Kernel (MVK) model [38]. On the other hand, a model replacing the kernel density distribution with a Gaussian mixture is used, namely a Multivariate Gaussian Mixture Model (MVGMM) model [39].

A public database of glass features is used as in [38] (the dataset can be downloaded from http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9876/homepage/glass-data.txt (last accessed on January 2018)). It contains features computed as log-ratios involving three combinations of chemical elements measured on glass fragments (log(Ca/K), log(Ca/Si) and log(Ca/Fe)), using some analytical chemistry technique. The dataset contains the measurements of five fragments for each of the 62 different glass objects in the database. The numbers of comparisons for each of the classes are $N_1 = 3762$ and $N_2 = 7564$, which yields an empirical prior of $\frac{N_1}{N_1+N_2} = \frac{1}{3}$.

Figure 9 shows the performance of the likelihood ratios computed by MVK and MVGMM. We will reason analogously as in the previous example in Section 6.1. The histograms in Figure 9a show that the $\log\left(\tilde{LR}^{(1)}_{(2)}\right)$ values computed by both forensic glass evaluation models are similar. They show a quite asymmetric behavior, although the histograms for the MVGMM case are slightly more symmetric, which in some sense indicates a slightly better calibration performance. This is also made evident in the reliability plots in Figure 9b, where it is also seen that the calibration is quite defective in both cases. ROC curves in Figure 9c show almost the same discrimination performance.

ECE plots are shown in Figure 9d. Here, it is clear that the MVK and MVGMM models have the same discrimination loss, since the ECE_{min} curves (blue, dashed) are equal. Moreover, it is seen that the calibration loss is also comparable in both approaches, since the ECE_{cal} (difference between red solid and blue dashed curves) is also comparable, although slightly better in the MVGMM approach.

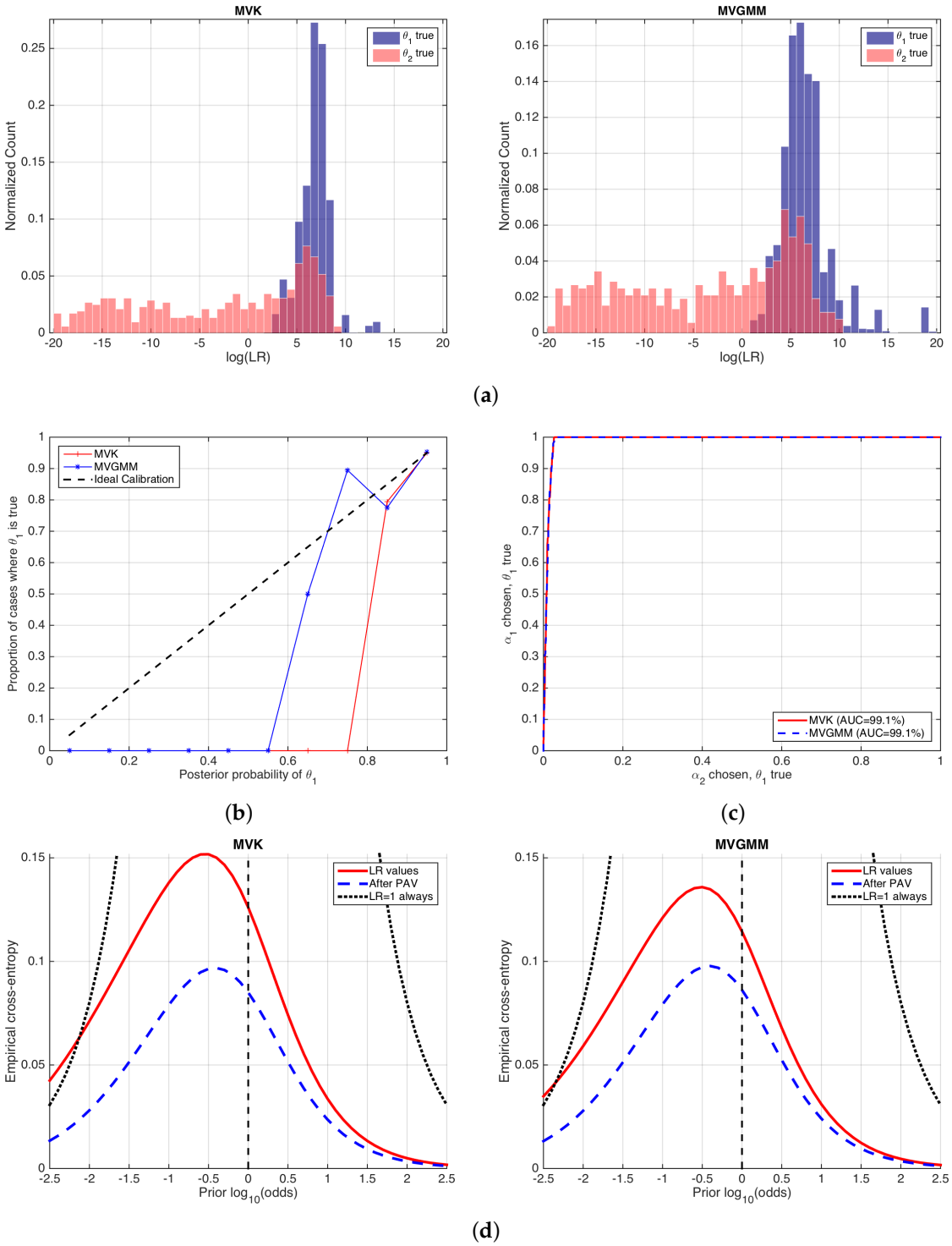


Figure 9. Performance of MVK and MVGMM glass evaluation models. (**a**) histograms of log $\left(\tilde{LR}^{(1)}_{(2)}\right)$ values; (**b**) reliability plots; (**c**) ROC curves; and (**d**) ECE plots.

Here, we highlight an important fact. It can be seen that both the MVK and the MVGMM classifiers have a performance worse than the neutral curve for some range of prior $\log_{10}$-odds, namely lower than -2.5 for MVK and lower than -2.6 for MVGMM. These correspond to prior probabilities respectively of $P(\theta_1) \leq 0.032$ and $P(\theta_1) \leq 0.025$. This means that, if the trier of fact assigns a prior probability lower than those values, it is better that she or he does not make use of the report on glass analysis given by the forensic examiner because it will be misleading. This is relevant in forensic science because the forensic examiner will seldom know the prior probability assigned by the trier of fact. However, she or he can, at least, warn the trier of fact about the range of prior probabilities where the forensic report should not be considered. This property of ECE plots is definitely not given by any of the other performance representations described in this article.

Information-theoretical interpretations can be also made as in the speaker verification example, taking P_o or P_{cal} as references. However, in this case, this is even more important because, in a forensic case, end-users are typically not proficient in mathematics and statistics, and care should be taken in order to explain the performance results if requested on trial (in any case, the authors discourage the use of complex performance representations on trial given the difficulties of human beings in general and triers of fact in particular to understand forensic statistics. A recent study about this can be found in [40]).

7. Discussion

In our opinion, this article presents sound contributions for the general fields of pattern recognition, data mining and machine learning. One of these contributions is showing the, typically independent, sources of information affecting cross-entropy: prior knowledge and value of the features (computed as a likelihood ratio). These are not typically taken into account by previous approaches using cross-entropy in classifiers, such as in [3,5,6], where it is most common that the empirical prior probability is used solely. Other related measures such as Confusion Entropy (CEN) or Matthews Correlation Coefficient (MCC) [41,42] work with decision errors rather than probabilities, which implies the selection of a threshold τ, and therefore they do not consider performance at different prior probabilities either (moreover, in [42] it is not recommended to use CEN in two-class problems, while ECE is perfectly suited to the task). However, the importance of this prior-dependent analysis is highly relevant, since we should aim at classifiers computing LR values that can work properly for very different prior probabilities and costs, as it has been shown critical in the described examples.

Another contribution of this article is the decomposition of cross-entropy into its discrimination and calibration components. Not only can this significantly help system designers to focus on the right problem, but it also theoretically justifies the use of invertible transformations of likelihood ratios as in [3,5,13]. In forensics, this is currently a trend in many disciplines such as chemistry [43], biometrics [14], or voice comparison [44,45], as well as in some more general classification problems such as speaker verification [13]. One of the motivations of this article is precisely to foster the use of these calibration approaches for any classifier with the help of ECE plots.

The interpretations provided in terms of information theory, thanks to the use of reference probability distributions for the cross-entropy, are also an important contribution. As no *true* probability can be invoked whatsoever as the reference probability to be compared in the cross-entropy formulation, those reference probabilities should be carefully established in order to yield a useful interpretation, as it happens with oracle and calibrated references.

We strongly believe that all these contributions of the present article will help scholars and system designers in the pattern recognition, data mining and machine learning areas. We tried to provide a better understanding of the importance and usefulness of cross-entropy, in order to take advantage of its power as a performance measure in a deeper way by the use of ECE plots, and to increase its use for improving modern artificial intelligence and facing new data mining challenges.

Acknowledgments: The current research has been supported by the Spanish Ministry of Economy and Competitiveness through project TEC2015-68172-C2-1-P.

Author Contributions: Daniel Ramos, Javier Franco-Pedroso, Alicia Lozano-Diez and Joaquin Gonzalez-Rodriguez conceived and designed the experiments; Daniel Ramos, Javier Franco-Pedroso and Alicia Lozano-Diez performed the experiments; Daniel Ramos analyzed the data; Daniel Ramos contributed analysis tools; Daniel Ramos, Javier Franco-Pedroso, Alicia Lozano-Diez and Joaquin Gonzalez-Rodriguez wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
2. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, Book ed.; Smola, A.J., Bartlett, P., Sholkopf, B., Shchuurmans, D., Eds.; MIT Press: Cambridge, MA, USA, 1999; Chapter 10, pp. 61–74.
3. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceeding of the Eight International Conference on Knowledge Discovery and Data Mining (KDD'02), Edmonton, AB, Canada, 23–26 July 2002.
4. Cohen, I.; Goldszmidt, M. Properties and benefits of calibrated classifiers. In *Lecture Notes in Computer Science*; Knowledge Discovery in Databases: PKDD 2004; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3202.
5. Niculescu-Mizil, A.; Caruana, R. Predicting Good Probabilities With Supervised Learning. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 7–11 August 2005; pp. 625–632.
6. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
7. Kittler, J.; Hatef, M.; Duin, R.; Matas, J. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 226–239.
8. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques—Adaptive Computation and Machine Learning*; MIT Press: Cambridge, MA, USA, 2009.
9. Sim, I.; Gorman, P.; Greenes, R.A.; Haynes, R.B.; Kaplan, B.; Lehmann, H.; Tang, P.C. Clinical Decision Support Systems for the Practice of Evidence-based Medicine. *J. Am. Med. Inform. Assoc.* **2001**, *8*, 527–534.
10. Tversky, A.; Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science* **1974**, *185*, 1124–1131.
11. Gigerenzer, G.; Hoffrage, U.; Kleinbölting, H. Probabilistic Mental Models: A Brunswikian Theory of Confidence. *Psychol. Rev.* **1991**, *98*, 506–528.
12. Van Leeuwen, D.; Brümmer, N. An introduction to application-independent evaluation of speaker recognition systems. In *Speaker Classification*; Müller, C., Ed.; Lecture Notes in Computer Science/Artificial Intelligence; Springer: Heidelberg/Berlin, Germany; New York, NY, USA, 2007; Volume 4343.
13. Brümmer, N.; du Preez, J. Application Independent Evaluation of Speaker Detection. *Comput. Speech Lang.* **2006**, *20*, 230–275.
14. Ramos, D.; Krish, R.P.; Fierrez, J.; Meuwly, D. From Bometric Scores to Forensic Likelihood Ratios. In *Handbook of Biometrics for Forensic Science*, Book ed.; Tistarelli, M., Champod, C., Eds.; Springer: Cham, Switzerland, 2017; Chapter 14, pp. 305–327.
15. Murphy, A.H.; Winkler, R.L. Reliability of Subjective Probability Forecasts of Precipitation and Temperature. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1977**, *26*, 41–47.
16. Ramos, D.; Gonzalez-Rodriguez, J. Reliable support: measuring calibration of likelihood ratios. *Forensic Sci. Int.* **2013**, *230*, 156–169.
17. Berger, C.E.H.; Buckleton, J.; Champod, C.; Evett, I.W.; Jackson, G. Expressing evaluative opinions: A position statement. *Sci. Justice* **2011**, *51*, 1–2.
18. DeGroot, M.H.; Fienberg, S.E. The Comparison and Evaluation of Forecasters. *Statistician* **1983**, *32*, 12–22.
19. Gneiting, T.; Balabdaoui, F.; Raftery, A.E. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B* **2007**, *69*, 243–268.
20. Dawid, A.P. The well-calibrated Bayesian. *J. Am. Stat. Assoc.* **1982**, *77*, 605–610.
21. Savage, L. The elicitation of personal probabilities and expectations. *J. Am. Stat. Assoc.* **1971**, *66*, 783–801.

22. Gneiting, T.; Raftery, A. Strictly Proper Scoring Rules, Prediction and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378.
23. Richard, M.D.; Lippmann, R.P. Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Comput.* **1991**, *3*, 461–483.
24. Willis, S. *ENFSI Guideline for the Formulation of Evaluative Reports in Forensic Science. Monopoly Project MP2010: The Development and Implementation of an ENFSI Standard for Reporting Evaluative Forensic Evidence*; Technical Report; European Network of Forensic Science Institutes: Wiesbaden, Germany, 2015.
25. Ramos, D.; Gonzalez-Rodriguez, J.; Zadora, G.; Aitken, C. Information-Theoretical Assessment of the Performance of Likelihood Ratio Models. *J. Forensic Sci.* **2013**, *58*, 1503–1518.
26. Kinnunen, T.; Li, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun.* **2010**, *52*, 12–40.
27. Brier, G. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3.
28. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
29. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley Interscience: New York, NY, USA, 2006.
30. Fawcett, T.; Niculescu-Mizil, A. PAV and the ROC convex hull. *Mach. Learn.* **2007**, *68*, 97–106.
31. Brümmer, N. Measuring, Refining and Calibrating Speaker and Language Information Extracted from Speech. Ph.D. Thesis, School of Electrical Engineering, University of Stellenbosch, Stellenbosch, South Africa, 2010. Available online: http://sites.google.com/site/nikobrummer/ (accessed on 31 January 2018).
32. Brümmer, N.; du Preez, J. The PAV Algorithm Optimizes Binary Proper Scoring Rules. Technical Report, Agnitio, 2009. Available online: https://sites.google.com/site/nikobrummer/ (accessed on 31 January 2018).
33. Dehak, N.; Kenny, P.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-End Factor Analysis for Speaker Verification. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 788–798.
34. Kenny, P. Bayesian speaker verification with heavy-tailed priors. In *Odyssey: The Speaker and Language Recognition Workshop*; International Speech Communication Association: Brno, Czech Republic, 2010.
35. Brümmer, N.; Burget, L.; Cernocky, J.; Glembek, O.; Grezl, F.; Karafiat, M.; van Leeuwen, D.A.; Matejka, P.; Scwartz, P.; Strasheim, A. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Trans. Audio Speech Signal Process.* **2007**, *15*, 2072–2084.
36. Martin, A.; Greenberg, C. The NIST 2010 speaker recognition evaluation. In Proceedings of the Interspeech 2010, Makuhari, Chiba, Japan, 26–30 September 2010; pp. 2726–2729.
37. Martin, A.; Greenberg, C. NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels. In Proceedings of the Interspeech 2009, Brighton, UK, 6–10 September 2009; pp. 2579–2582.
38. Aitken, C.G.G.; Lucy, D. Evaluation of trace evidence in the form of multivariate data. *Appl. Stat.* **2004**, *53*, 109–122; With corrigendum 665–666.
39. Franco-Pedroso, J.; Ramos, D.; Gonzalez-Rodriguez, J. Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data. *PLoS ONE* **2016**, *11*, e0149958.
40. Thompson, W.C.; Newman, E.J. Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law Hum. Behav.* **2015**, *39*, 332–349.
41. Wei, J.M.; Yuan, X.J.; Hub, Q.H.; Wangc, S.Q. A novel measure for evaluating classifiers. *Expert Syst. Appl.* **2010**, *37*, 3799–3809.
42. Jurman, G.; Riccadonna, S.; Furlanello, C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. *PLoS ONE* **2012**, *7*, e41882.
43. Corzo, R.; Hoffman, T.; Weis, P.; Franco-Pedroso, J.; Ramos, D.; Almirall, J. The Use of LA-ICP-MS Databases to Estimate Likelihood Ratios for the Forensic Analysis of Glass Evidence. *Talanta* **2018**, in press.
44. Gonzalez-Rodriguez, J.; Rose, P.; Ramos, D.; Toledano, D.T.; Ortega-Garcia, J. Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 2072–2084.
45. Morrison, G.S. Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Aust. J. Forensic Sci.* **2013**, *45*, 173–197.