Utility-Based Accuracy Measures to Empirically Evaluate Credal Classifiers

Marco Zaffalon
IDSIA, Switzerland
zaffalon@idsia.ch

Giorgio Corani
IDSIA, Switzerland
giorgio@idsia.ch

Denis Mauá
IDSIA, Switzerland
denis@idsia.ch

Abstract

Predictions made by imprecise-probability models are often indeterminate (that is, set-valued). Measuring the quality of an indeterminate prediction by a single number is important to fairly compare different models, but a principled approach to this problem is currently missing. In this paper we derive a measure to evaluate the predictions of credal classifiers from a set of assumptions. The measure turns out to be made of an objective component, and another that is related to the decision-maker's degree of risk-aversion. We discuss when the measure can be rendered independent of such a degree, and provide insights as to how the comparison of classifiers based on the new measure change with the number of predictions to be made. Finally, we empirically study the behavior of the proposed measure.

Keywords. Credal classification, indeterminacy, empirical evaluations, discounted accuracy, utility, risk-aversion.

1 Introduction

When we use an imprecise-probability model to make predictions, we meet one of the most striking differences of imprecise probability in comparison to precise probability: the imprecise-probability model can issue indeterminate predictions. That is, among the set of possible options, the model may drop some of them as sub-optimal, while keeping the entire remaining set as its prediction. The prediction is generally indeterminate as such a set is not necessarily a singleton. Indeterminate predictions are a crucially important feature of imprecise-probability models: they allow credible, and reliable, predictions to be obtained no matter how scarce is the information available to build a model.

Yet, we should have a way to *measure* how good is an indeterminate prediction. A major reason is that we need to compare imprecise- with precise-probability models: we should have a clear, simple, and possibly shared, way to say which one is better, in a given application. The same consideration applies, of course, when we compare two imprecise-probability models. Ideally, we would like to be able to reward each, determinate or indeterminate, prediction by a single number. Much probably this would speed up progress in the field, as it would enable comparisons to be automatized over a large number of test applications.

In the case of precise-probability models, there are well-consolidated measures to do so. Let us consider the field of *pattern classification* [4], which is the focus of this paper (Section 2 gives a brief introduction to classification problems). In this case, the predictive models are called (precise) *classifiers*. A classifier predicts one out of a finite set $\mathscr{C}$ of so-called *classes*. In this case, correct predictions may be rewarded with 1 and incorrect ones with 0, thus giving rise to the measure of performance called the *predictive accuracy* of a classifier: i.e., the proportion of correct predictions it makes.

The situation is very different with *credal classifiers*, that is, classifiers that issue set-valued predictions. One of the very few proposals to evaluate an indeterminate prediction by a single number can be found in [2]: a prediction made of a set $\mathscr{K}$ of k classes is rewarded with $1/k$ if it contains the actual class, and with 0 otherwise. This gives rise to the measure called *discounted accuracy*, which was borrowed from the field of multi-label classification [11]. The problem here is that no justification is given for discounted accuracy, as the work in [2] somewhat points out. In [7], it is proposed to evaluate classifiers which return indeterminate classifications through the F-metric, originally designed for information retrieval problems; but also here the measure is not justified. Other than these, the proposals are either explicitly non-numerical, as the rank test in [2], or require a vector of parameters to evaluate the performance, as in [1]. The latter approach is actually meaningful, but was conceived to compare credal with precise classifiers, and cannot be easily generalized to the more general case; moreover, it is a method that needs supervision so that it does not easily lend itself to be run automatically on many test cases.

On our view, the scarcity of principled numerical evaluation methods for credal classifiers is not accidental: in fact, it is not easy to assign a single number to an indeterminate

prediction. Consider the following case: there is a *vacuous* classifier, which every time predicts the set of all classes $\mathscr{C}$, and a *random* one, which picks up a class from $\mathscr{C}$ through the uniform distribution. If $\mathscr{C}$ is made of two classes (we say that the classification problem is binary), and we use the predictive accuracy, the random classifier has an expected reward equal to $1/2$. What should be the expected reward of the vacuous classifier? Both classifiers do not know how to predict the class, but only the vacuous classifier declares it. From this, one might argue that the latter should be rewarded with more than $1/2$. On the other hand, it is clear also that the vacuous classifier cannot predict the class better than the random one, so that one might argue that it should be rewarded with $1/2$ too.

In the attempt to address these kinds of problems in the most objective way, we found it useful to regard classifiers as bettors. In the betting framework introduced in Section 3, we assume we only know how to value determinate predictions, in particular by 0-1 rewards. In Section 4, we extend the framework, in a kind of least-committal way, to credal classifiers, in the attempt to see what are the implications of determinate rewards alone on indeterminate ones: we show that, under certain assumptions, indeterminate predictions should be valued according to discounted accuracy.

Note that, in the previous example, discounted accuracy would value the vacuous and the random classifiers the same. This kind of (questionable) effect can be traced back to having deliberately avoided introducing subjective considerations in the evaluation. Still, subjective preferences should be accounted for: we introduce in Section 5 a decision-maker in charge of selecting the 'best' classifier in the next bet, and show that preferences can enter the picture through his utility, as a function of discounted accuracy. This defines the utility-based accuracy measure we propose to evaluate credal classifiers. More generally, this shows in a very definite sense how the reliability of a classifier is tightly related to the variability of its predictions, and that the aversion to this variability is what makes some people prefer credal classifier to precise ones.

In Section 6 we discuss an important case where the evaluation can still be made in quite an objective way despite the decision-maker's preferences, and we relate this to the amount of indeterminacy produced by a credal classifier.

In Section 7 we analyze how the picture changes if we focus on evaluating classifiers in the next $m \geq 1$ bets. We show that the difference between precise and credal classifiers decreases with growing m, so that the relative benefits of credal classification are less important with large m.

Finally, in Sections 8 and 9 we make some empirical analysis of our utility-based measure, by comparing *naive Bayes* [3] with *naive credal classifier* [1] on binary classification problems. We show that the decision-maker's utility can be defined very easily in this case, and that the credal classifier becomes superior to the precise one even with relatively small preferences of the decision-maker towards reliable predictions.

2 Classification Problems

A classification problem is made of objects described by *attribute* (or *feature*) variables, which we group into the single variable A, and a class variable C. The class variable represents the object's category. There are finitely many possible categories, which we identify with their indexes to simplify notation: $\{1,\ldots,n\} =: \mathscr{C}$. We denote by c the generic element of $\mathscr{C}$. The attribute variable represents some characteristics of the object that are related to the class. Variable A takes values in the set $\mathscr{A}$; we denote by a its generic element. As an example, objects might be patients; A would represent information about a patient, such as personal information as well as outcomes of medical tests; $\mathscr{C}$ would index the patient's possible diseases.

Usually, some values of (A,C) are sampled in an independent and identically distributed way according to a law that is not known a priori. The so-called *learning set* $\mathscr{L}$ records those values, which are also called *instances* of (A,C). The goal of classification is to learn from the learning set a function that maps attributes into classes. We call this function a (*precise*) *classifier*.

A classifier is applied to predict the class of new objects based on their attributes. Predictions are rewarded through a *reward matrix* $\mathbb{R}$. This is an $n \times n$ matrix whose generic element r_{ij} is a number representing the reward obtained by predicting class i when the actual class is j. Equivalently, we can regard the reward matrix as a set of *gambles* (i.e., bounded random variables) $\mathbb{R}_i$, $i = 1,\ldots,n$, each one corresponding to a row of $\mathbb{R}$: gamble $\mathbb{R}_i$ represents the uncertain reward obtained by predicting class i and is defined by $\mathbb{R}_i(j) := r_{ij}$, with $j \in \mathscr{C}$. The reward matrix is an input of the classification problem, in the sense that it is given.

In classification, at least with respect to the machine learning practice, rewards are usually measured in a linear utility scale: although this point is often left implicit, we can deduce it from the observation that the performance of a classifier is usually identified with its expected reward.

The most frequent practice consists also in using just a 0-1 valued reward matrix, which we denote by $\mathbb{I}$. In this case, the gamble corresponding to the i-th row of the matrix coincides with the indicator function of set $\{i\}$, which yields $\mathbb{I}_i(i) = 1$, and $\mathbb{I}_i(j) = 0$ for $i \neq j$. Accordingly, the performance of a classifier corresponds to the probability of predicting the actual class. Such a probability is called the *predictive accuracy* (or simply the *accuracy*) of a classifier.

The term 'accuracy' is used also for the sample estimate of such a probability. In fact, a classification problem usually comes with a test set $\mathscr{T}$. This set contains a number of

sampled instances of (A,C) that are used to evaluate the classifier's predictive performance by measuring its accuracy on them. And in fact the predictive accuracy is by far the most frequently used empirical index to compare classifiers, despite a careful elicitation of rewards would arguably lead in many cases to a reward matrix more general than $\mathbb{I}$. Such a widespread use has probably been favored by the simple interpretation of predictive accuracy; a more substantial reason could be that the predictive accuracy is particularly convenient to make extensive comparisons of classifiers over many data sets, which is a key component of the machine learning practice. Accordingly, in this paper we focus on the 0-1 valued reward matrix $\mathbb{I}$.

So far we have introduced the traditional view of classification, where the predictions issued by (precise) classifiers are made of single classes. This view has been generalized through the introduction of credal classifiers [13, 14]. A *credal classifier* is also a function learned from set $\mathscr{L}$, but it maps the attributes of an instance into a set $\mathscr{K} \subseteq \mathscr{C}$ of $k := |\mathscr{K}|$ classes in general. We call this a set-valued classification. We also say that the classification is *determinate* when $k = 1$, and *indeterminate* otherwise. When a classification is fully indeterminate, that is, when $\mathscr{K} = \mathscr{C}$, we call it *vacuous*. Similarly, the *vacuous classifier* is the one that always issues vacuous predictions. To each credal classifier it is possible to associate a determinate classifier that outputs predictions by choosing every time a class uniformly at random[1] from the output set $\mathscr{K}$ of the credal classifier. We call this the *$\mathscr{K}$-random classifier*; when the related credal classifier is the vacuous one, we just call it the *random classifier*.

3 Introducing the Betting Framework

In order to make the comparison of credal classifiers as objective as possible, we introduce the idea of a betting framework. We define the framework for a traditional problem of classification, where classifiers issue determinate predictions. In Section 4 we will extend the framework to credal classification.

In the framework under consideration, we have two classifiers, which we would like to compare, that have already been inferred from data (so that there is no further learning, only an evaluation stage). These classifiers are regarded as bettors. Bets correspond to instances of the problem of classification: a bet is set up by sampling an instance of the problem. Classifiers are required to bet by predicting the actual class of the instance, and are rewarded according to matrix $\mathbb{I}$. The process is repeated for ever, and the performance of classifiers is taken to be their predictive accuracy.

[1] Throughout the paper we use the word 'random' to mean *uniformly* random.

Let us make the betting framework more precise by describing the two types of actors that play a role there:

Bettors: each of the two classifier we aim at comparing is regarded as a bettor.

House: rewards are delivered to bettors by an artificial entity that we call House. House only accepts determinate bets, which are rewarded according to matrix $\mathbb{I}$.

These actors are characterized by clarifying their relationship with the rewards, that is, with the utility scale involved. To start with, based on the discussion made in Section 2, we can readily state our first assumption concerning the betting framework:

(A1) Utility of bettors is linear in the rewards.

This assumption simply states explicitly what is current practice in classification.

The second assumption concerns House. We want to model House as an agent whose only aim is to reward correct predictions. In other words, House should not introduce any subjective bias in the process of rewarding bettors because of a risk-averse or risk-seeking attitude; it should just be risk-neutral:

(A2) Utility of House is linear in the rewards.

4 Betting with Credal Classifiers

Now we would like to extend the betting framework to credal classifiers. The crucial point here is that House only accepts determinate bets, while a credal classifier outputs set-valued classifications in general. Therefore, if we want to allow a credal classifier to play, we should find a way to extend the reward matrix to set-valued classifications in a way that both House and bettor find acceptable.

The first step in this direction is to recognize that any negotiation between the credal classifier and House can be made only on the basis of determinate bets, which are the only language that House understands. In order to enable the credal classifier play as a determinate bettor, we state the following assumption:

(A3) The credal bettor accepts betting on any single class from its set-valued prediction, if forced to make a determinate bet, and on no class outside that set.

This assumption is satisfied whenever the classes in the output set of the credal classifier are incomparable, and the other ones represent dominated options. This is the case when credal classifiers are obtained using sets of probabilities and decision criteria like maximality or e-admissibility

(see, e.g., [12, Section 3.9]). We state the assumption explicitly in order to allow the framework to be used also by credal classifiers created in a different way.

The next assumption formalizes the idea that the framework is run for ever:

(A4) Every possible bet is repeated infinitely many times in the betting framework by sampling the problem instances.

This assumption, together with the previous one, enable the credal classifier to actually adopt a randomized strategy over the k classes in its output set $\mathscr{K}$. A randomized strategy is a mass function $\sigma = (\sigma_i)_{i\in\mathscr{K}}$ that represents the (determinate) betting behavior of the credal classifier in the limit.

At this point House knows that the credal classifier has the freedom to implement any randomized betting strategy: this means that the credal classifier can actually force House to undergo any expected loss that can follow from the choice of the strategy.

Let us call a prediction $\mathscr{K}$ 'successful' if the actual class belongs to $\mathscr{K}$. We restrict the attention to successful predictions as they determine House's expected loss: in fact, an unsuccessful prediction always yields a zero loss, by definition of $\mathbb{I}$, irrespective of the randomized strategy adopted. Let $\theta = (\theta_j)_{j\in\mathscr{C}}$ be the vector of chances, that is, the population proportions, for the classes conditional on the prediction being successful (this means that $\theta_j = 0$ if $j \notin \mathscr{K}$). House's expected loss conditional on a successful predictions equals

$$\sum_{i\in\mathscr{K}}\sum_{j\in\mathscr{C}} \mathbb{I}_i(j)\sigma_i\theta_j = \sum_{i\in\mathscr{K}} \sigma_i\theta_i,$$

where we are assuming that the strategy is chosen independently of the chances.

The loss depends on σ, which is chosen by bettor, and on θ. The latter models the specific problem under consideration. But House knows that the betting system will be applied, in principle, to every possible problem. House should then be enabled to consider every possible scenario:

(A5) In the determination of the expected loss, House has the freedom to choose any value for θ.

At this point we are ready to derive the extended reward matrix:

Theorem 1. *Let $\mathscr{K} \subseteq \mathscr{C}$ be a set-valued prediction made of k classes, $\mathbb{I}_\mathscr{K}$ be the indicator function of set $\mathscr{K}$, and j the actual class. The corresponding value in the extended reward matrix that is uniquely consistent with (A1)–(A5) is the discounted accuracy:*

$$\frac{\mathbb{I}_\mathscr{K}(j)}{k}. \tag{1}$$

Proof. If $\mathscr{K}$ is unsuccessful, then any randomized strategy will yield a zero loss. Let us focus on successful predictions. Let Δ be the $n-1$ probability simplex. We formulate the problem in a game-theoretic setting. The two players are just bettor and House. Bettor can choose $\sigma \in \Delta$, while House can choose $\theta \in \Delta$. What we get is a zero-sum game with a gain for bettor defined by $\sum_{i\in\mathscr{K}} \sigma_i\theta_i$. This is a continuous linear function in σ for all $\theta \in \Delta$, as well as in θ for all $\sigma \in \Delta$, and moreover Δ is a compact convex set. The minimax theorem (see, e.g., [10, Theorem 6.7.3]) allows us to deduce that there is an optimal solution to the game with expected reward equal to $\max_{\sigma\in\Delta}\min_{\theta\in\Delta}\sum_{i\in\mathscr{K}} \sigma_i\theta_i$. It is easy to see that that is equal to $1/k$: once a strategy σ is fixed, the minimum is achieved by setting $\theta_{i_*} := 1$ on any $i_* = \operatorname{argmin}_{i\in\mathscr{K}} \sigma_i$; then the problem becomes $\max_{\sigma\in\Delta}\min_{i\in\mathscr{K}} \sigma_i = 1/k$. The related optimal strategy σ^* is uniform, $\sigma^*_i := 1/k$ for all $i \in \mathscr{K}$; this means that bettor and House agree that credal bettor should act like the $\mathscr{K}$-random classifier.

Now, remember that, according to (A1)–(A2), both bettor and House are risk-neutral. This means they agree that an unsuccessful prediction is rewarded by the certain value 0 and a successful one by the certain value $1/k$. This is achieved by setting the reward equal to the discounted accuracy. □

It is useful to comment on this result from a few different viewpoints.

One thing is that the the discounted accuracy implements a kind of least-committal reward system for House, in the sense that House gives bettor only what is certainly due to it. In fact, if the credal bettor does implement strategy σ^*, the expected reward that it achieves is indeed $1/k$, irrespective of the chances. Therefore the established reward is what House knows already that bettor can make for sure. For the same reason, it would be implausible to expect that credal bettor accepts any smaller reward. It is also interesting to observe that playing as the $\mathscr{K}$-random bettor (i.e., classifier) is the only way for credal bettor to have a sure reward.

The next consideration is again based on the observation that credal bettor is evaluated exactly as the $\mathscr{K}$-random bettor. This has important implications for the comparison of classifiers through the discounted accuracy: the main point is that the $\mathscr{K}$-random bettor is actually taken as a baseline to compare classifiers. Consider, for the sake of explanation, a determinate classifier whose output class is always contained in that of a certain credal classifier. The determinate classifier will be evaluated better than the credal classifier as soon as it exploits, to any (even a very tiny) degree, the credal classifier's set of output classes better than the $\mathscr{K}$-random one. Looking at this from another side, it means that the credal classifier can be better than the determinate one only if it behaves worse than the $\mathscr{K}$-random classifier! In practical applications, this will imply that a credal classifier will almost *never* be superior to a determinate classifier whose output is included in the credal's one. This discussion should make clear that the discounted accuracy, although it is a reasonable criterion, is probably

the most unfavorable way (among the reasonable ones) to evaluate credal classifiers, as a credal classifier cannot do better than isolating a set of classes that is impossible to compare.

This points to an aspect of the evaluation that the discounted accuracy certainly fails to capture. Let us focus on the simplest possible setup, using the following example. You are trying to evaluate two physicians based on some recorded diagnostic performance of theirs. In your records, the first physician always issues a vacuous diagnosis, that is, the entire set $\mathscr{C}$ of possible diseases. The second always issues a determinate diagnosis. But when you measure the second physician's predictive accuracy, you realize that his predictions are random. In this case, the discounted accuracy values the two physicians the same: $1/n$. But it is clear that the first physician provides you with something more than the second, because, in a sense, he delivers what he promises. How to precisely value this 'something more' appears to be quite a subjective matter. In this sense, it should not be too surprising that discounted accuracy does not value it at all, as it has been created trying to keep subjectivity out of consideration. And yet, subjectivity matters, and should be taken into account. The next section shows that this can be done in a very natural way.

5 Comparing Credal Classifiers

We have two classifiers f, g. We focus on selecting the classifier whose expected performance in the next instance (i.e., next bet) is greater than the other's. To this end, we start identifying classifiers with gambles: gambles f and g yield the discounted-accuracy reward achieved by classifiers f and g, respectively, in the next instance. There is uncertainty about these gambles because we assume that the instance has yet to be sampled.

The comparison of gambles f and g needs a (rational) decision-maker, whom we call 'you'. By definition of the gambles, you will compare them based on discounted-accuracy rewards. We model your attitude towards these rewards through the following assumption:

(A6) Your utility function[2] $u(\cdot)$ is concave in the discounted-accuracy rewards,

which means that you are risk-averse, or at most neutral, in these rewards.

This appears quite a reasonable assumption, at least in the common setup where the original rewards (the ones used to define the 0-1 reward matrix $\mathbb{I}$) are measured in a utility scale that is linear for you. In fact, imagine that you are explicitly asked to extend the reward matrix to take into account your attitude towards set-valued classifications. Can we say something about the values you would use to define such an extended matrix? On the one hand, we argue that the rewards you would put there should be greater than or equal to the discounted-accuracy rewards. This follows from the discussion at the end of Section 4, which shows that it would be unreasonable to use values smaller than the discounted accuracy. On the other hand, values strictly greater than that would be reasonable: these allow you to express a preference in favor of a set-valued classification in comparison to the related $\mathscr{K}$-random prediction. These considerations imply that your utility function is in general non-linear in the discounted-accuracy (that is, discounted accuracy can be regarded as defining a new utility scale out of the original one). We take your utility in particular to be concave to express a consistent preference for set-valued classifications in comparison to the related $\mathscr{K}$-random predictions (note that this includes the extreme case of a linear utility function, in which the two options are equally valued).

Going back to the comparison of classifiers, it follows immediately from (A6) and decision-theoretic arguments that you will choose the one with maximum expected utility: $h^* := \operatorname{argmax}_{h\in\{f,g\}} E[u(h)]$.

Re-consider the example of the vacuous and the random classifier, discussed at the end of Section 4, as they are emblematic of the differences that arise in the evaluation of credal and precise classifiers when using utility.

Proposition 2. *The random and the vacuous classifiers have the same expected reward on the next instance, but the expected utility of the vacuous is greater under any strictly concave utility function.*

Proof. Denote the random classifier by r, and the vacuous classifier by v. As usual, we identify the classifiers with the corresponding gambles, which represent uncertain discounted-accuracy rewards for the next bet. The vacuous classifier gets on any instance the deterministic reward $1/n$. Thus, under any utility function:

$$E[u(v)] = u\left(\frac{1}{n}\right) = u(E[v])\,.$$

The random classifier r samples the predicted class from $\mathscr{C}$ according to the uniform mass function σ^*, independently of the actual class. Let us denote, as usual, by $\theta = (\theta_j)_{j\in\mathscr{C}}$ the vector of chances for the actual classes. We obtain that

$$E[r] = \sum_{i\in\mathscr{C}}\sum_{j\in\mathscr{C}} \mathbb{I}_i(j)\sigma_i^*\theta_j = \sum_{i\in\mathscr{C}} \sigma_i^*\theta_i = 1/n\,.$$

This shows that $E[v] = E[r]$. In addition, using Jensen's inequality leads to

$$E[u(r)] < u(E[r])) = u(1/n) = E[u(v)]\,,$$

whenever u is a strictly concave function. □

To better analyze this point, it is useful to approximate the expected utility by a second-order Taylor series. Let h be a

[2] We assume that the usual regularity conditions for utility hold, and in particular that it is strictly increasing, and that it has first and second derivatives (see, e.g., [9]).

generic classifier (and hence, a gamble):

$$E[u(h)] \simeq u(E[h]) + \overbrace{u'(E[h])E(h-E[h])}^{=0} + \\ + \frac{1}{2}u''(E[h])E\left[(E[h]-h)^2\right] = \\ = u(E[h]) + \frac{1}{2}u''(E[h])Var[h], \quad (2)$$

where u', u'' are the first and second derivative of the utility function, and $Var[h]$ denotes the variance of h. Well-known papers in finance [6, 8] have shown that this a very accurate approximation.

Remember that $u''(E[h]) \leq 0$ for every concave utility function (moreover, $u''(\cdot)$ is related to the degree of risk aversion of the utility assessor). Therefore what Equation (2) tells us is that the expected utility increases by increasing the expectation of rewards and decreasing their variance. It is clear now why the vacuous classifier, with variance equal to zero, is preferred to the random one. In other words, the 'something more' that the vacuous classifier is providing is its inherent reliability in earning rewards, which, using discounted accuracy, has a very clear numerical counterpart in its variance. The value that you give to this is indeed personal, and is formalized through your utility function. In the extreme case when you are risk-neutral in the discounted-accuracy rewards, the value is zero, and in this case there seems to be little room for credal classifiers in your interests. Bigger values express stronger preferences for reliable predictions.

It is also interesting to briefly consider the case where you are risk-averse in the original rewards defining $\mathbb{I}$. This would much probably be the case if those rewards represented amounts of money. In particular, if the discounted-accuracy rewards were the actual money payed by a betting system, then you would be 'natively' risk-averse in them; as a natural byproduct, you would prefer the more reliable (i.e., less variable) credal classifier to its $\mathscr{K}$-random counterpart.

All the above considerations can be turned into a remarkably simple procedure to empirically compare credal classifiers in practice. Remember that in a classification problem we usually have a test set $\mathscr{T}$, that is, a collection of instances used to evaluate the performance of a classifier. We need to estimate $E[u(h)]$ for a certain classifier h. Let us denote by $\mathscr{U}$ the set of values that gamble $u(h)$ can take. Set $\mathscr{U}$ has $(2^n-1)\cdot n$ elements, as they are in one-to-one correspondence with the elements of the reward matrix extended through discounted accuracy. If we estimate the chance of a value $u_h \in \mathscr{U}$ by its sample proportion $\#(u_h)/|\mathscr{T}|$ in the test set, we obtain:

$$E[u(h)] \simeq \sum_{u_h \in \mathscr{U}} u_h \frac{\#(u_h)}{|\mathscr{T}|} = \frac{1}{|\mathscr{T}|} \sum_{(a,c)\in\mathscr{T}} u(h(a,c)).$$

This is equivalent to evaluating the performance of a credal classifier using the $(2^n-1)\times n$ reward matrix obtained by applying function $u(\cdot)$ point-wise to the matrix extended through discounted accuracy.

A final consideration is that the comparison can be, perhaps more conveniently, made also using $u^{-1}(E[u(h)])$, the so-called *certainty equivalent*. This brings the performance index back to the range $[0,1]$ so that it can still be interpreted as a predictive accuracy, although one that is distorted through the utility function.

6 The Case for an Objective Winner

Equation (2) is useful because it gives us a very accurate approximation to the expected utility while releasing us from having our considerations narrowed down by the specific form of the utility function considered. To this end, in the following, we will repeatedly refer to (2) as if it was our actual expected utility.

In particular, an interesting consideration suggested by Equation (2) is that in one case the comparison of classifiers can be done minimizing subjective considerations: when the two classifiers have equal expected reward. In this case, the classifier with minimum variance wins under every strictly concave utility function: that is, no matter how tiny (but non-zero) is your degree of risk-aversion. This can be implemented in practice by defining a range where the difference of the expected rewards is deemed irrelevant, and estimating their variances from the test set.

In the following, we investigate whether we can relate the variance of a classifier with its *determinacy*, that is, with a measure of the amount of imprecision in the output. Intuitively, we expect such a relationship to exist because both measures are related to the reliability of a classifier, and moreover, we expect that larger indeterminacy corresponds to smaller variance.

The gamble h corresponding to a classifier's performance in the next bet can be decomposed in two other gambles h_D and h_I such that $h = h_D + h_I$ and $h_D h_I = 0$ (element-wise). Intuitively, h_D and h_I represent the rewards for f when it returns, respectively, a determinate and an indeterminate classification. The following identities follow from the decomposition under discounted accuracy:

$$E[h^2] = E[h_D^2] + E[h_I^2], \\ E[h_D^2] = E[h_D], \\ E[h_I] \geq E[h_I^2]. \quad (3)$$

In Equation (3) the equality is obtained only if $E[h_I] = E[h_I^2] = 0$, which implies that either h is a precise classifier or that indeterminate predictions of h contain the actual class with probability zero.

Let f and g denote two generic classifiers with the same

expected discounted accuracy: $E[f] = E[g]$. Using the identities above, one can show that the difference of variances is thus

$$\begin{aligned}\Delta Var &:= Var[g] - Var[f] \\ &= E[g_D] + E[g_I^2] - E[f_D] - E[f_I^2]. \end{aligned} \quad (4)$$

Let us start by considering the important case where we compare a credal classifier with a precise one:

Proposition 3. *Consider a credal classifier and a precise classifier with the same expected reward. Then the credal classifier is preferable to the precise classifier under any strictly concave utility function.*

Proof. Let us denote by f the credal classifier and by g the precise one. We know by Equation (2) that we prefer the classifier with smaller variance under any strictly concave utility function. Thus, it suffices to show that $\Delta Var \geq 0$. Since $E[f_I^2] \leq E[f_I]$, it follows from Equation (4) that

$$\begin{aligned}\Delta Var &= E[g_D] - E[f_D] - E[f_I^2] \\ &\geq E[g_D] - E[f_D] - E[f_I] \\ &= E[g] - E[f], \end{aligned}$$

which equals zero, since f and g have equal expected reward. Note the inequality is strict (i.e., there is strict preference) if the credal classifier is not always determinate and its indeterminate predictions are successful with positive probability. □

Now, let H_D be the event that equals 1 when the generic classifier h is determinate on the next instance, and 0 otherwise. We define the *determinacy* of classifier h as the probability that h is determinate: $P(H_D)$. This definition allows us to settle the problem for the next case:

Proposition 4. *Consider two credal classifiers that are vacuous whenever they are indeterminate and that have the same expected reward. Then the more indeterminate classifier is preferable under any strictly concave utility function.*

Proof. Let us denote by f and g the two credal classifiers, assuming f to be more indeterminate than g: $P(G_D) > P(F_D)$. It suffices to show that $\Delta Var > 0$. Any generic classifier h that is vacuous whenever it is indeterminate is rewarded with $1/n$ for any indeterminate prediction. Hence,

$$E[h_I] = \frac{1 - P(H_D)}{n}, \qquad E[h_I^2] = \frac{E[h_I]}{n}.$$

From these identities and Equation (4) we have that

$$\begin{aligned}\Delta Var &= E[g_D] + E[g_I]/n - E[f_D] - E[f_I]/n \\ &= -E[g_I] + E[g_I]/n + E[f_I] - E[f_I]/n \\ &= \frac{n-1}{n}\left(-E[g_I] + E[f_I]\right) \\ &= \frac{n-1}{n^2}\left(P(G_D) - P(F_D)\right), \end{aligned}$$

which is strictly positive by the initial assumptions. □

This proposition is particularly useful as it allows us to solve the problem in the case of binary classification problems, where any indeterminate prediction is necessarily vacuous.

One might be tempted to think that the previous result extends to non-vacuous classifiers as well, that is, that the more determinate a classifier the higher its variance (and therefore the more preferable it is). Unfortunately, this is not the case, as the following example shows.

Example 1. *Consider a three-class classification problem. Let H_k denote the event that equals 1 if the generic classifier h returns a set of k classes that contains the actual one, and 0 otherwise. Likewise, let H_k^c be the event that equals 1 if h outputs k incorrect classes, and 0 otherwise. Note that $\sum_{k=1}^{3} H_k + H_k^c = 1$ and $H_3^c = 0$. We can define the relevant expectations in terms of H_k, H_k^c:*

$$P(D_h) = P(H_1) + P(H_1^c), \qquad E[h] = \sum_{k=1}^{3} \frac{1}{k} P(H_k),$$

$$E[h^2] = \sum_{k=1}^{3} \frac{1}{k^2} P(H_i), \qquad 1 = \sum_{k=1}^{3} P(H_i) + P(H_i^c).$$

Assume that $P(F_1) = P(G_1) + \varepsilon$, $P(G_1^c) = P(F_1^c) + 2\varepsilon$, $P(G_2) = P(F_2) + 2\varepsilon$, $P(F_2^c) = P(G_2^c) + 3\varepsilon$, and $P(F_3) = P(G_3)$, for some small $\varepsilon > 0$. Then we have from the identities above that $E[f] = E[g]$. Similarly, we have that $E[f^2] = E[g^2] + \frac{\varepsilon}{2}$. Hence, $\Delta Var = E[g^2] - E[f^2] < 0$, and g is preferred over f even though g is more determinate than f: $P(D_f) = P(D_g) - \varepsilon$.

Alternatively, we might measure the indeterminacy of a classifier h by the expected number of classes it outputs: $\sum_{k=1}^{n} k\left[P(H_k) + P(H_k^c)\right]$. Thus, in the example, we would have

$$\sum_{k=1}^{n} k\left[P(F_k) + P(F_k^c)\right] = \sum_{k=1}^{n} k\left[P(G_k) + P(G_k^c)\right] + 4\varepsilon,$$

and g is preferred over f even though the former has a smaller expected number of output classes than the latter. ♦

7 Comparison Over the Next m Bets

So far, we have considered the expected reward and utilities for the next *single* classification; this setting fits for instance the case of a patient, who asks a doctor for a diagnosis and who is concerned only about the utility generated by the very next classification (his diagnosis). Conversely, an online trader, who performs m trading operations every day, might accept to lose some money in the very next transaction, provided that the set of m transactions generated at the end of day has high enough utility. In this case, expected rewards and expected utilities should be computed over the next m bets. In the following, we compare the random classifier r and the vacuous classifier v on the next m bets; we denote by v_m and r_m the rewards of the vacuous and the random ones over the next m instances.

Gamble v_m has deterministic value m/n and thus:

$$E[u(v_m)] = u\left(\frac{m}{n}\right).$$

To compute $E[u(r_m)]$, let us consider that classifier r yields utility $u(k)$ when it correctly predicts ℓ outcomes in the next m bets; considering that classifier r issues a correct classification with probability $1/n$ (see Proposition 2), the probability of correctly predicting ℓ instances out of the next m is the binomial:

$$Bin(\ell, m, \frac{1}{n}) = \binom{\ell}{m} \frac{1}{n}^{\ell} \left(1 - \frac{1}{n}\right)^{m-\ell}.$$

The expected utility produced by the random classifier over the next m bets is thus:

$$E[u(r_m)] = \sum_{\ell=1}^{m} u(\ell) Bin(\ell, m, \frac{1}{n}). \tag{8}$$

It is not immediate to compare the expected utilities of the random and vacuous classifiers using Equation (8); a clear understanding can be obtained through the second-order approximation given by Equation (2). In the following, we analyze in this way the logarithmic and the exponential utility. The second-order approximation of both the logarithmic and the exponential utility is very good, having relative absolute error consistently smaller than 1%.

7.1 Logarithmic Utility

The logarithmic utility is $u(x) := \log(1+x)$, whence $u''(x) = -\frac{1}{(1+x)^2}$; applying Equation (2), we get:

$$u(E[r_m]) + \frac{1}{2}u''(E[r_m])Var(r_m) = u(E[r_m]) - \frac{Var(r_m)}{2(E[r_m]+1)^2} = u\left(\frac{m}{n}\right) - \frac{m\frac{1}{n}\left(1-\frac{1}{n}\right)}{2(\frac{m}{n}+1)^2},$$

where in the last passage we introduced the analytical expression of the variance for a binomial distribution.

Thus, the (approximated) difference between the expected utility of the random and the vacuous over the next m bets is

$$d(m) = E[u(v_m)] - E[u(r_m)] = \frac{\frac{m}{n}\left(1-\frac{1}{n}\right)}{2(\frac{m}{n}+1)^2} \propto \frac{m}{(\frac{m}{n}+1)^2}, \tag{9}$$

where in the last passage we removed the proportionality constant $\frac{1}{2n}\left(1-\frac{1}{n}\right) > 0$. Function $d(m)$ is shown in Fig. 1.

The first derivative of $d(m)$ is:

$$d'(m) = \frac{1}{(\frac{m}{n}+1)^2} - 2\frac{\frac{m}{n}}{(\frac{m}{n}+1)^3} \propto 1 - \frac{m}{n}, \tag{10}$$

where the last passage is obtained considering that $(\frac{m}{n}+1)^3 > 0$. From Equations (9) and (10), we can figure out that

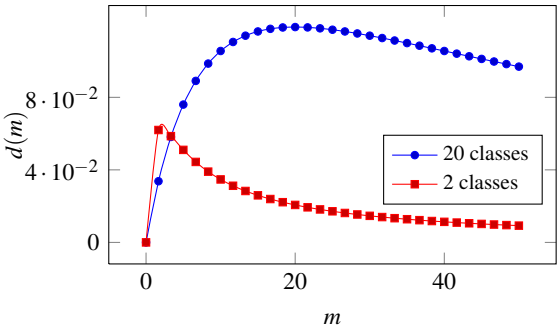

Figure 1: Function $d(m)$ for *logarithmic* utility, under different number of classes.

$d(m)$ will monotonically increase up to $m < n$ (*inversion point*), to then indefinitely decrease, so that $d(m) \to 0$ for $m \to \infty$; if expectations of utilities are computed over a long enough number of bets, the expected utility produced by the two classifiers is the same. It also follows that increasing n delays the convergence of the expected utilities to the same value, as also shown in Fig. 1.

7.2 Exponential Utility

The exponential utility is $u(x) := 1 - \exp(-ax)$, where a is a coefficient of risk-aversion. Noting that $u''(x) = -a^2\exp(-ax)$, the second-order approximation yields:

$$u(E[r_m]) + \frac{1}{2}u''\left(\frac{m}{n}\right)Var(r_m) = u\left(\frac{m}{n}\right) - \frac{1}{2}a^2\exp\left(-a\frac{m}{n}\right)m\frac{1}{n}\left(1-\frac{1}{n}\right),$$

whence

$$d(m) = -\frac{1}{2}a^2\exp\left(-a\frac{m}{n}\right)m\frac{1}{n}\left(1-\frac{1}{n}\right) \propto \propto -\exp\left(-a\frac{m}{n}\right)m,$$

where the proportionality constant is $\frac{a^2}{2}\frac{1}{n}\left(1-\frac{1}{n}\right) > 0$.

We have

$$d'(m) = \exp\left(-a\frac{m}{n}\right)\cdot\left(a\frac{m}{n} - 1\right).$$

Function $d(m)$ has qualitatively the same behavior of the logarithmic case, but the inversion point is now located at $m = \frac{n}{a}$. Moreover, the difference between the expected utility of the two classifiers depends also on the risk-aversion coefficient a; higher risk-aversion delays the convergence of the expected utilities, thus emphasizing the difference in favor of the vacuous on small m.

8 Experiments on Artificial Data Sets

In the following, we denote the naive Bayes classifier as NBC [3] and the naive credal classifier as NCC [1]. We

compare the utility generated by NBC and NCC on the next *single* bet. In a first set of experiments, we generated artificial data sets, considering a binary class and 10 binary features; we set the marginal chances of classes as uniform, while we drew the conditional chances of the features under the constraint $|\theta_{i1\ell} - \theta_{i2\ell}| \geq 0.1 \; \forall i, j$, where $\theta_{ij\ell}$ denotes the chance of feature A_i to be in state ℓ when $C = j$; the constraint forced each feature to be truly dependent on the class. We drew θ 80 times uniformly at random and we consider the sample sizes: $s \in \{25, 50, 100\}$. We did not consider larger sample sizes, under which NCC would have been almost completely determinate, and thus not really different from NBC. For each pair (θ, s) we generated 50 training sets; we then evaluate the trained classifiers on a test set of 10000 instances. In the following, the instances indeterminately classified by NCC are referred to as the *area of ignorance*. We denote as NBC(NCC-I) the accuracy of NBC on the area of ignorance. For each sample size, we thus perform 80 θ * 50 trials = 4000 training/test experiments.

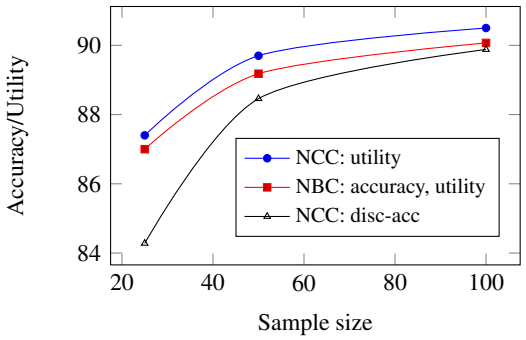

Figure 2: Experimental results with artificial data; each point shows the median over 4000 experiments, performed with the same sample size s. For NBC, accuracy and utility coincide. For NCC, the curve of utility rises the values of discounted accuracy.

We set the utility of a determinate and successful classification as $u(1) := 1$; the utility of a non-successful classification (determinate or indeterminate) as $u(0) := 0$. This is the case, for instance, if you are risk-neutral in the scale the original rewards are measured. It remains to set the utility $u(0.5)$ of an indeterminate classification (notice that for a data set with two classes, an indeterminate classification has necessarily discounted accuracy of 0.5). We think that in general the value of $u(0.5)$ could reasonably lie between 0.6 and 0.8; in our experiments, we set $u(0.5) := 0.65$. As a term of comparison, determinate and indeterminate classifiers have been compared in [7] through the F_1 metric, which is widely used in information retrieval. Under the F_1 metric, on a dataset with 2 classes, the vacuous classifier gets the same score of a precise classifier with 66% accuracy; this gives further support to our choice.

As expected, NBC has higher discounted accuracy than NCC (see Fig. 2); this means that, on the area of ignorance, it is doing better a than the $\mathscr{K}$-random guesser. Yet, NCC produces slightly higher utility than NBC at each sample size. The determinacy of NCC rises steadily with the sample size; interestingly, at the same time the value of NBC(NCC-I) decreases; this means that NCC is getting better at identifying instances which are really hard to classify. For instance, NBC(NCC-I) is 64% for $s = 25$, and 54% for $s = 100$; this explains why the gap of utility tends to slightly increase with the sample size. Note however that the restriction of the area of ignorance (20% for $s = 25$, and only 4% for $s = 100$) works against enlarging the gap between NCC and NBC. Results similar to those shown here are obtained also using logarithmic utility; however we find it clearer in this simple setting to reason about the only point to elicit, $u(0.5)$, rather about the whole utility function.

9 Experiments on the kr-kp Data Set

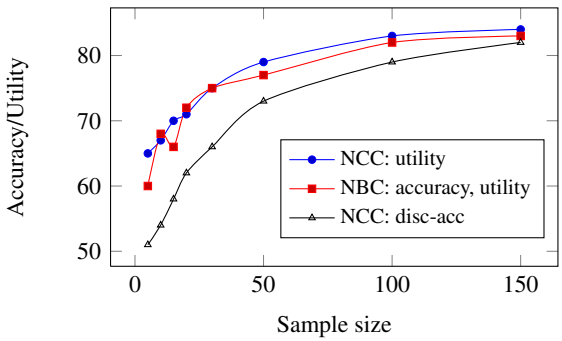

Figure 3: Utility and discounted accuracy generated by NBC and NCC for downsampled versions of kr-kp.

We then performed some experiments on the kr-kp data set (2 classes, 36 binary features, 3200 instances) from the UCI repository. To evaluate the sensitivity of the performance on the sample size, we worked by *downsampling* the kr-kp data set. In particular, we generated training sets of size $s \in \{5, 10, 15, 20, 30, 50, 100, 150\}$; for each sample size, we generated 100 different training sets; for each training set, the corresponding test set is given by the instances left in the original data set. All training and test sets are *stratified*, namely the proportion among the two classes matches that of the original data set. For each sample size, we report the average results over all splits; the results are shown in Tab. 1 and Fig. 3. The determinacy of NCC steadily increases with the sample size, as well its discounted accuracy and the accuracy of NBC. For NBC, notice that accuracy and utility have the same value. For very small s (e.g., $s = 5$), NCC is almost always indeterminate; in this case, its utility corresponds to $u(0.5)$ and thus is 0.65; in the same situation, NBC is almost randomly guessing, and

s	NCC: Determ (%)	NBC: NBC(NCC-I) (%)
5	2	59
10	10	65
15	25	60
20	29	64
30	41	64
50	60	62
100	78	60
150	85	59

Table 1: Results for the kr-kp experiment; Determ. indicates the % of instances determinately classified by NCC.

thus its utility is close to 50%. Both the utility of NBC and NCC smoothly increases with s; the utility of NCC remains however slightly superior. In fact, under a data set with two classes, whether NCC or NBC produces a higher utility can be realized by comparing NBC(NCC-I) with $u(0.5)$; if $u(0.5) <$NBC(NCC-I), then NCC produces higher utility than NBC, and vice versa. However, the outcome of the comparison would be slightly in favor of NBC by (conservatively) setting $u(0.5) = 0.6$, as can be deduced from Tab. 1; in fact, once utility is introduced in the evaluation of the classifiers, it also plays a role in the final decision about which of the considered classifiers is better. This also implies that to generate sensible results when using utility-based metrics, it is fundamental to *carefully* elicit the decision maker's utility.

10 Conclusions

In this paper, we have tried to define in a principled way a measure to empirically evaluate credal classifiers. In our proposal, any such measure is made of two main components: the discounted accuracy, which represents a kind of objective performance of a classifier, and its variance, which represents the unreliability of the classifier, and whose contribution to the overall measure has to be weighted through subjective considerations of risk-aversion. Our measure can be implemented very easily in practice, and in fact is shown to empirically lead to some interesting results. Future work could (i) explore generalizations to rewards more general than 0-1 ones; (ii) exploit what appear to be natural connections between our measure and finance, in order to evaluate credal classifiers (some recent work connecting utility and machine learning, that could be useful to consider in that respect, has also recently appeared [5]); and also (iii) deepen the empirical study in order to verify the possibility to define some kind of 'general purpose' utility functions for machine learning aims.

Acknowledgements

The research in this paper has been partially supported by the Swiss NSF grants n. 200020_134759 / 1, 200020-121785 / 1, 200020-132252 and by the Hasler foundation grant n. 10030.

References

[1] G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: the naive credal classifier 2. *Journal of Machine Learning Research*, 9:581–621, 2008.

[2] G. Corani and M. Zaffalon. Lazy naive credal classifier. In J. Pei, L. Getoor, and A. de Keijzer, editors, *First ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*, pages 30–37. ACM, 2009.

[3] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2001. 2nd edition.

[5] C. Friedman and S. Sandow. *Utility-Based Learning from Data*. Chapman & Hall/CRC, Baca Raton, Fl, 2011.

[6] W. Hlawitschka. The empirical nature of Taylor-series approximations to expected utility. *The American Economic Review*, 84(3):713–719, 1994.

[7] J. Jose del Coz and A. Bahamonde. Learning nondeterministic classifiers. *Journal of Machine Learning Research*, 10:2273–2293, 2009.

[8] H. Levy and H.M. Markowitz. Approximating expected utility by a function of mean and variance. *The American Economic Review*, 69(3):308–317, 1979.

[9] D. G. Luenberger. *Investment Science*. Oxford University Press, New York, 1998.

[10] J. Stoer and C. Witzgall, editors. *Convexity and Optimization in Finite Dimensions*. Springer-Verlag, Berlin, 1970.

[11] G. Tsoumakas and I. Vlahavas. Random k-Label sets: An Ensemble Method for Multilabel Classification. In J. N. Kok, J. Koronacki, R. López de Mántaras, S. Matwin, D. Mladenic, and A. Skowron, editors, *Proceedings of ECML 2007, 18th European Conference on Machine Learning*, volume 4701 of *Lecture Notes in Computer Science*, pages 406–417. Springer, 2007.

[12] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[13] M. Zaffalon. A credal approach to naive classification. In G. de Cooman, F. G. Cozman, S. Moral, and P. Walley, editors, *ISIPTA '99: Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications*.

[14] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.