AOS4 : Article summary (Sections 1-7)

Evaluating credal classifiers by utility-discounted predictive accuracy
Marco Zaffalon, Giorgio Corani, Denis Mauá

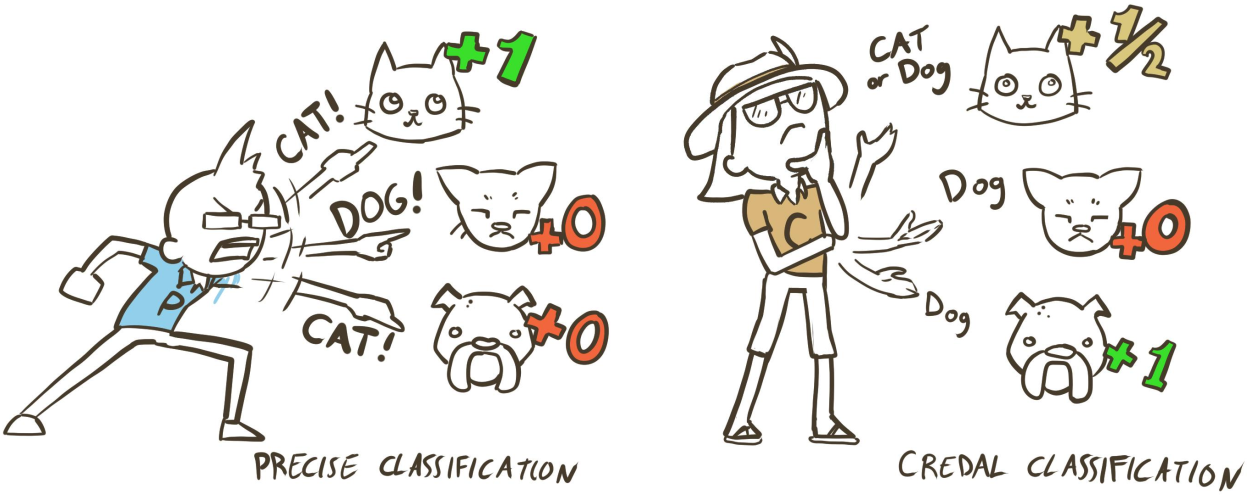
Hugo Martin, Robin Monje
8 janvier 2023

Machine Learning (ML) is a booming field at the intersection of many other fields, among which mathematics, statistics and computer science, which focuses on building models from data to accurately predict and/or explain a variable. One of the main types of problems in ML is **classification**, which is the problem of building a model (called a **classifier**) that outputs the correct class that an input belongs to. For instance, such a problem could be to output the correct species of an animal given an image of this animal. This may not seem like it, but this specific problem is a very difficult one that has been puzzling computer scientists for decades. A classifier cannot always be right : even our brain sometimes confuses animals, for instance we might mistake some big dogs with wolves. This is why there is a need to **evaluate** the performance of a model, so we can compare them and pick the best one. This is usually not a very difficult problem, even though a lot of approaches exist. We often rely on the **predictive accuracy** of the model : we feed it a lot of inputs it has not yet seen, called the **test data**, and we count the number of correct predictions it made[1]. It is equivalent to giving them a "reward" of 1 when they accurately predict the class, and 0 otherwise.

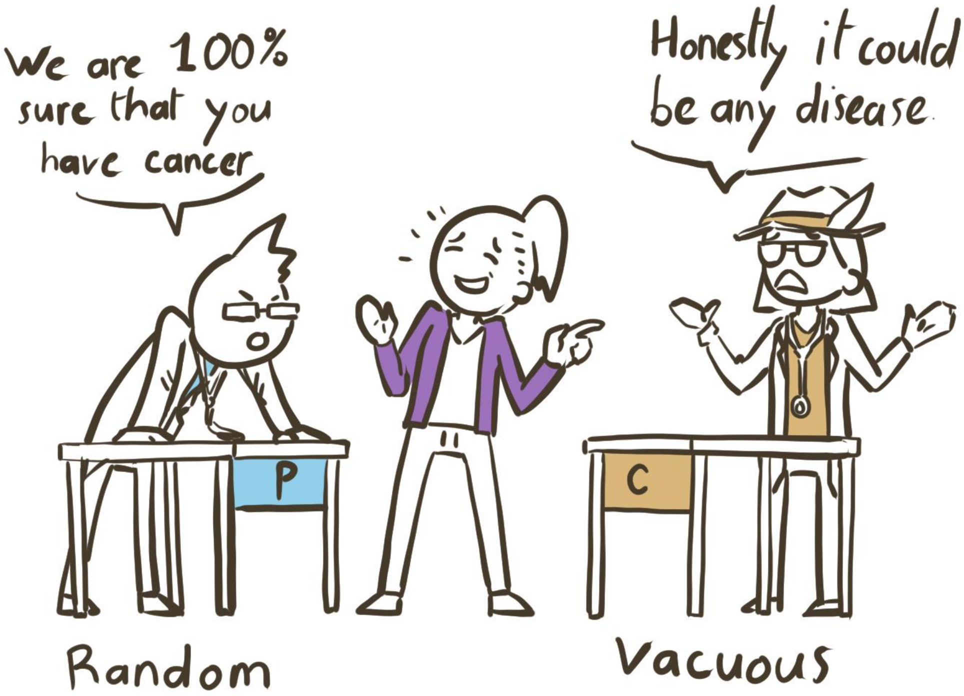
This works well for most classifiers. However, there exists a specific type of classifier for which this approach cannot be directly applied : **credal** classifiers. For each input, these models do not output one single class but a **set** of classes they think the input could correspond to. When they output more than one class, we say it is an **indeterminate** prediction. The most common way to evaluate credal classifiers is then to reward them with a value of *1/k* if the correct class is in the predicted set, *k* being the number of classes in this set, and 0 otherwise. The authors of the article summarized here derive this formula from a

[1] Usually we also divide by the number of examples in the test data, so that classifiers tested on data with different sizes can be compared.

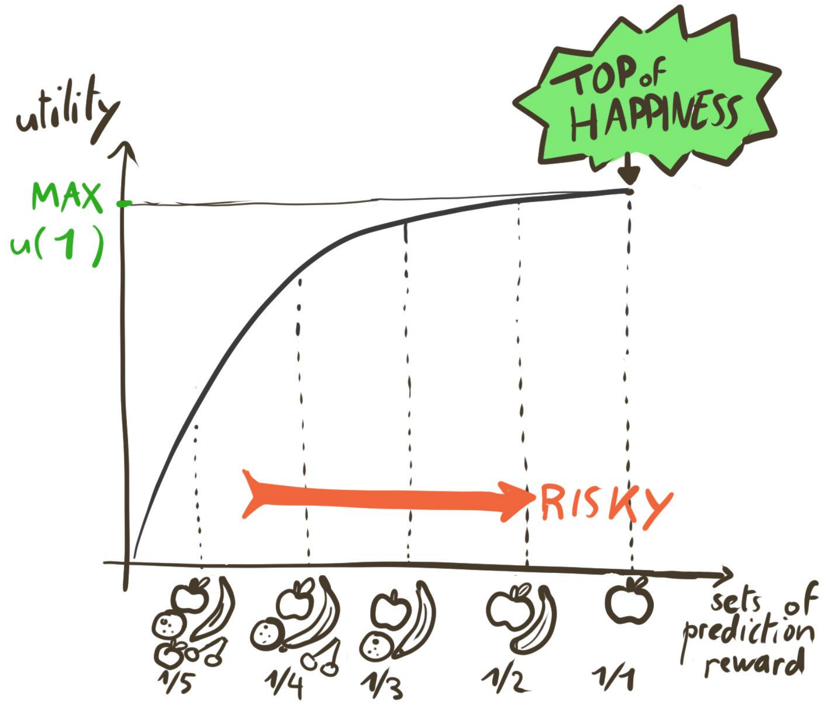
set of assumptions and in particular show this is the only way of rewarding these classifiers when we have "no subjective preference" for indeterminate predictions.



To understand this last statement, consider two classifiers. The first one (called vacuous) always outputs a set containing all the possible classes. The second one (called random) always outputs one class, chosen at random regardless of the input. If we reward them according to the method previously seen, you can convince yourself they will both receive a reward of *1/c* on average, *c* being the number of classes in total. However, we might want to favor the first classifier, since it at least acknowledges that it does not know which class is correct. If it was a doctor, you would probably prefer it to be clueless and open about it rather than clueless and ready to diagnose you with a random disease. If we prefer this first classifier, then we have a subjective preference for indeterminate predictions.


We are 100% sure that you have cancer
Honestly it could be any disease
P
C
Random
Vacuous

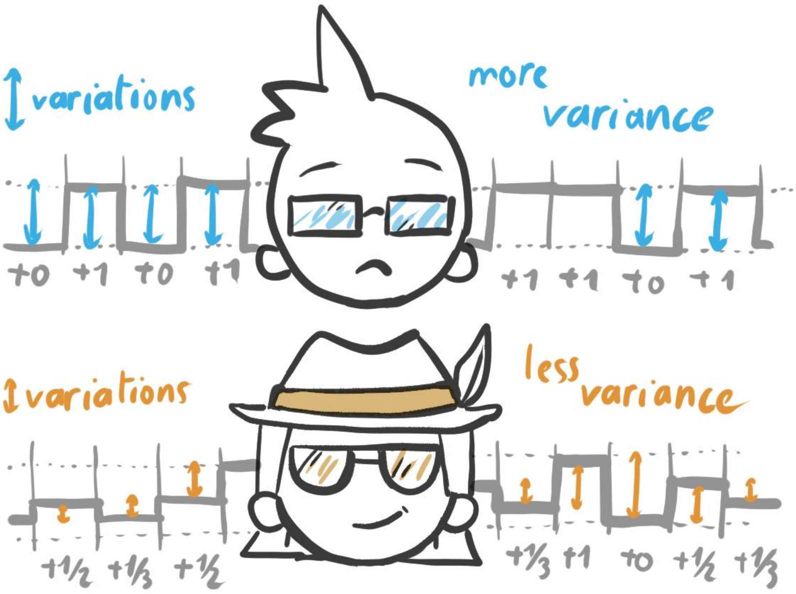
Favoring indeterminate predictions implies that in an imaginary scenario where you would start from a predicted set which contains all the classes possible and you slowly empty it class by class, at first you would increase your happiness a lot with each class removed because your prediction would get more precise. However, the happiness you receive when removing a class will grow smaller and smaller, until you get to a point where you will not get that much happier after removing a class, because you realize you risk accidentally removing the correct class, and that is too costly for you. This is called being risk averse, and can be modeled via non-decreasing concave functions. The "non-decreasingness" is because you will never lose happiness when you have one less class and the correct guess is in the predicted set, and the concavity is because the happiness gains will go smaller and smaller with the number of classes removed. In this context, "Happiness" is usually called **utility** instead.



When classifiers include the correct class in their predicted set, we then do not reward them with *1/k*, but with *u(1/k)* instead, with *u* being a concave function called utility function like the one depicted above. With these rewards, we do not have a standard accuracy anymore, but a "utility-discounted" one. The authors show that for risk averse people, the classifier that always predicts all the classes has an advantage over the random classifier : its consistency in winning rewards every time. The variance of its rewards is indeed lower (actually it is literally 0 because it always receives the same reward). Risk averse agents do not like high variance, because you can get a very good reward or a very poor one, which is risky.

Based on this insight, the authors prove two interesting propositions. Both those propositions assume that your utility function is strictly concave. They add the "strictly" requirement here to rule out the linear utility function (since linear functions are concave), which would imply we are neutral to risk.

The authors first show that if two classifiers, one credal and the other precise, have the same reward on average, then the credal one is preferable. The idea is that the variance of the credal classifier's rewards are necessarily lower or equal to those of the precise one, since it can predict more classes and thus include the correct class more often. The values of the rewards are closer to each other for this classifier than for the precise one, which always gets 0's or 1's.



Finally, the authors show that among two classifiers that predict all the classes when they are indeterminate and have the same expected rewards, the one being the most often indeterminate is preferable. This second property allows us to compare two credal classifiers in the case where there are two variables (and thus every indeterminate prediction includes both classes). Notice that it does not mean that a classifier always predicting all the classes will be better than any other, because it also has to have a high enough reward on average.

In conclusion, the authors developed a metric allowing us to compare credal classifiers between themselves and between precise classifiers by allowing the person building the model to inject their own personal preference towards risk, depending on the type of problem and the subjective preference of this person.

It's just a cold. You should have come to see me earlier!
I know. I met two weird doctors on my road. They weren't any big help...