 doi:10.20944/preprints202108.0521.v1

The costs and potential benefits of introducing the "I don't know" answer in binary classification settings

Damjan Krstajic[1]

[1] Research Centre for Cheminformatics, Jasenova 7, 11030 Beograd, Serbia
* Correspondence: DK: damjan.krstajic@rcc.org.rs

Abstract: We are of the opinion that during the design of a binary classifier one ought to consider adding an "I don't know" answer. We provide the case for the introduction of this third category when a human needs to make a decision based on the answer from a binary classifier. We discuss the costs and potential benefits of its introduction. Colloquially, we have used the term "I don't know", but formally we refer to it as NotAvailable. A procedure to define NotAvailable predictions in binary classifiers, called all leave-one-out models (ALOOM), is presented as proof of the concept. Furthermore, we discuss the potential benefits of applying ALOOM in real life applications.

Keywords:: non-applicability domain; binary classification; ignorance; decision-making

1. Background

In our practice we have come up with the following situation. We have been given a task to create a binary classification model to predict the existence of a cancer based on certain markers. We have been provided with a training set to build the model and a blind test set for prediction purposes. The test set consisted of 100 samples with known values of markers, but unknown status regarding the cancer. After a thorough process of model selection and assessment [1], we selected a support vector machine model. We built the model using the training set, predicted 100 test samples and noted the predictions.

However, we accidentally repeated the same process of building the model and predicting samples, and we found that three samples have different predicted categories. Upon closer inspection, we found that the problematic samples had predicted probabilities close to 0.5. Due to random choice inside the algorithm of our model building technique (SVM), even though the inputs were identical and the machine was the same, the output equations of the models were not exactly the same. Therefore, the predicted probabilities of the problematic test samples were sometimes above 0.5 and sometimes below, thus producing opposite predicted categories.

So how should we report the predicted categories of the problematic three test samples? The obvious answer would be

 doi:10.20944/preprints202108.0521.v1

to say that we cannot predict them. However, how would one then calculate the misclassification error, sensitivity and specificity of the test set? We finished the project by repeating the process of creating the model and predicting the test set 101 times. For the problematic three test samples we have reported the predicted category which had highest number of votes amongst 101 votes.

Even though we have completed our project and reported the predicted categories for 100 test samples, we were not satisfied with our predictions. If we were a patient and someone used a binary classifier to predict whether we have a cancer, we would like to have been informed if the binary classifier cannot predict an outcome for us.

2. Introduction

A common approach to describe the quality of a single predicted classification is to assign a probability to it. It is a measure of the extent to which a predictive model is confident in its answer. Why then introduce the "I don't know" answer when we can have the probability as the measure of its confidence? One may argue that it is up to the user to decide how to act upon that information. We provide two cases to demonstrate why that might be problematic. We also specify the precondition for introducing the "I don't know" answer and provide a case for referring to it as NotAvailable.

2.1. Are probabilities well calibrated?

Prior to accepting that probability values may be used, one ought to make sure that the predicted probabilities are well calibrated [2]. Generally speaking, well-calibrated probabilities mean that for 100 samples with predicted probability W, the proportion of correct predictions amongst them is expected to be W. From our experience in creating binary classification models, we have found it difficult to generate a well calibrated model. We are not aware of any model selection process that optimises the goal of achieving well-calibrated binary classification models. Consequently, the values of predicted probabilities might on average be misleading.

2.2. Do we have the same understanding of a probability value?

The following argument is based on psychological research related to the way individuals interpret or make use of probability estimates. Tversky and Kahneman [3][4] found that most people could not maintain a consistent view of what different numerical probabilities meant, and therefore "anchor", i.e. rely too much on the first piece of information offered to them when making decisions. The best they could find was that people could keep a consistent sense of the meaning of 50:50 and the meaning of "almost certain". This means that a decision maker who tries to distinguish between a 0.85 probability and a 0.7 probability cannot really tell the difference, and in practice the human decision maker is unlikely to assign any significance to the difference.

 doi:10.20944/preprints202108.0521.v1

So if a model predicts that a chemical compound is toxic with probability 0.89, then does it make any difference to a toxicologist if it is 0.83? Similarly, if the model predicts that a patient has an aggressive cancer with probability 0.29, then does it make any difference to an oncologist if it is 0.31? According to Tversky and Kahneman [3][4], anchoring may effect both the toxicologist and the oncologist. However, it appears that it would not be the same if the predicted probability is close to 0.5. We would like to emphasise that we are here only concerned with individual probability predictions. We are not underestimating in any way the importance of probability predictions when comparing different potential toxic compounds or different patients, especially when they are well calibrated.

Tversky and Kahneman's [3][4] findings lead Patrick Suppes to propose a simple probability model with only five probabilities [5]:

1. surely true
2. more probable than not
3. as probable as not
4. less probable than not
5. surely false

As we are dealing with predicting unknown test samples, "surely true" and "surely false" answers cannot be expected.

In our opinion, Tversky and Kahneman [3][4] provide the case for the introduction of the third category when a human needs to make a decision based on the answer from a binary classifier.

2.3. Precondition

There is a basic precondition for introducing the "I don't know" answer. If we compare the accuracy, or quality, of predictions prior to and after its introduction, then answers that are not "I don't know" need to be on average more accurate than all answers prior to its introduction. We expect to gain in accuracy and lose in productivity.

There are costs and benefits associated with the introduction of an "I don't know" answer. One benefit is the improvement in accuracy, or quality, of non "I don't know" answers, while a cost is the proportion of "I don't know" answers. There might also be other benefits and costs. When the benefits outweigh the costs, then arguments for its introduction are self-evident. However, we shall presume throughout the text that we are dealing with a zero sum game, that is to say that the gains in accuracy are the same as the losses in productivity.

2.4. NotAvailable answer instead of "I don't know"

If we introduce an "I don't know" answer in binary classification, what does it say about answers that are not "I don't know"? We are concerned that it might imply that a binary classification model knows them, which is not true.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 August 2021 doi:10.20944/preprints202108.0521.v1

Krstajic *et al.* [6] proposed the introduction of a third prediction category in binary classifications, which they referred to it is as Uncertain. The author [7] has recently put the case in favour of naming the term as NotAvailable. By referring to the term as NotAvailable we are not implying anything regarding the quality of answers that are not NotAvailable, except that they are – available.

Colloquially, we have used the term "I don't know", but formally we will refer to it as NotAvailable. We will use the term NotAvailable when referencing results from Krstajic *et al.* [6] even though they called it Uncertain.

3. Methods

The goal is to create a predictive model F() which predicts a variable Y using values of variables X1,..,Xm. It can be viewed as the relationship Y=F(X1,,..,Xm). If our predictive model is a statistical model, then it is created using previously known values (Yi,Xi1,…Xim) i=1..N, which we refer to as the learning data. Binary classification models are predictive models where Y has only two values, e.g. aggressive cancer and non-aggresive cancer, which we will refer to as positive and negative. The common measure of quality for a binary classifier is the misclassification error, i.e. the proportion of wrong predictions. Furthermore, in binary classification settings it is common for a predictive model F() not only to predict whether something is positive or negative, but also to estimate its probability.

3.1. Using an uncertainty interval

Krstajic *et al.* [6] examined the introduction of NotAvailable predictions in binary classifications using an uncertainty interval. If a predicted probability is inside the uncertainty interval then it would be classified as NotAvailable. For example, if the uncertainty interval is [0.47, 0.55] then all samples with predicted probabilities within the interval will be categorised as NotAvailable.

3.2. All Leave One Out Models (ALOOM)

Krstajic *et al.* [6] suggested the following rule for defining NotAvailable predictions. If a learning dataset consists of N samples then one would create N binary classification models on (N-1) samples in exactly the same way as the original model F() was created. This would mean that when predicting an unseen sample we would have N predicted categories from the N models. The sample would be categorised as NotAvailable if and only if it has opposite predicted categories among the N predictions. Otherwise all N models would predict unanimously, either all positive or all negative, and that would be the predicted category. Krstajic *et al.* [6] refer to the above process as All Leave One Out Models (ALOOM).

During the ALOOM process the predicted probabilities of a validation sample from all N models are recorded. Therefore, at the end of ALOOM we may calculate the minimum and maximum of N predicted probabilities for each validation

 doi:10.20944/preprints202108.0521.v1

sample, which we refer to as the *ALOOM individual prediction interval*.

4. Results

Here we summarise the main results from Krstajic *et al.* [6] and Damjan Krstajic [7] that are important for our discussion later.

We shall here present their findings in the case of mutagenicity predictions. They used a publicly available dataset of 4335 chemical compounds, 2400 categorised as "mutagen" (positive) and the remaining 1935 compounds as "nonmutagen" (negative) [8]. Chemical descriptors were calculated for each compound, and these were inputs (X1,..,Xm) for creating predictive models [9].

They applied the following two model building techniques: ridge logistic regression and random forest. They were chosen as representatives of the linear and the nonlinear approach and because a variable selection process is not required.

4.1. Using an uncertainty interval

Krstajic *et al.* [6] repeated the following process 100 times. The dataset was split into equally sized halves ensuring that each half contains the same proportion of "mutagen" and "nonmutagen" samples. One half was used as the learning dataset to create a predictive model and the other half to predict an outcome, i.e. as the validation dataset. For each validation sample the model predicted a category ("mutagen" or "nonmutagen") as well as the probability of it being "mutagen".

They examined what would happen if all samples with probabilities in the range [0.49, 0.51] were to be categorised as NotAvailable, and then for the range [0.48, 0.52], and so on until [0.3, 0.7]. In Figures 1 and 2 the relationship between the intervals of uncertainty and average misclassification errors (from 100 repeats) as well as the average proportion of NotAvailable is shown for each model building technique.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 August 2021 doi:10.20944/preprints202108.0521.v1

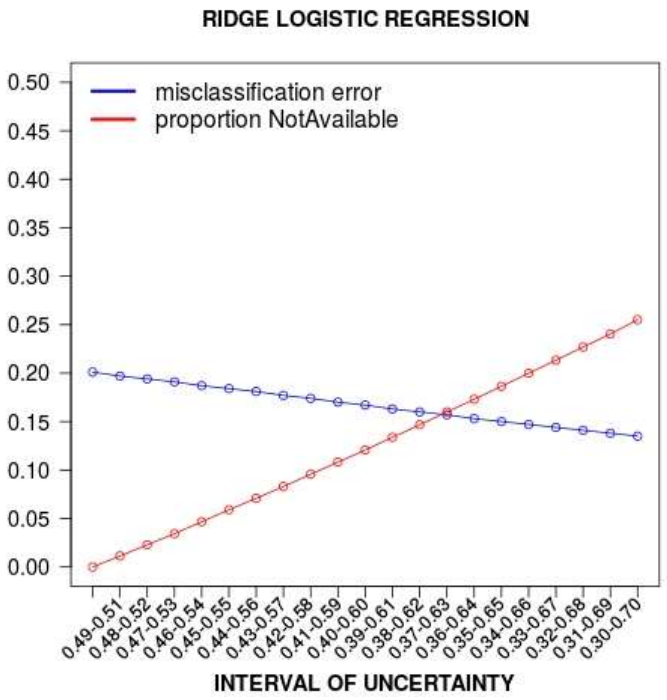

Figure 1. – Relationship between the intervals of uncertainty and average misclassification errors (from 100 repeats) as well as the average proportion of NotAvailable for ridge regression.

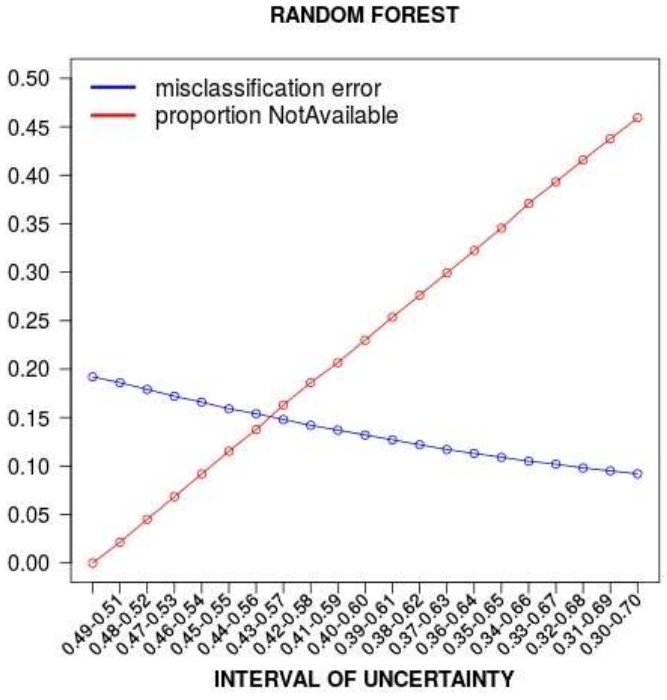

Figure 2. – Relationship between the intervals of uncertainty and average misclassification errors (from 100 repeats) as well as the average proportion of NotAvailable for random forest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 August 2021 doi:10.20944/preprints202108.0521.v1

In both cases (ridge regression and random forest) the introduction of NotAvailable predictions caused other predictions to be on average more accurate. We would like to point out that the relationship between the width of the uncertainty interval and average misclassification errors appears linear, as does the relationship between the width of the uncertainty interval and the proportion of NotAvailable predictions. This means that there is not an obvious interval of uncertainty to be selected.

4.2. ALOOM

Krstajic *et al.* [6] applied the ALOOM on the same mutagenicity dataset using the same 100 splits into halves and applying the same two types of models: ridge logistic regression and random forest. In Table 1 we show the mean misclassification error and mean percentage of NotAvailable predictions generated with the ALOOM for each model type as well as the results of the original model F().

MODEL TYPE	ALOOM MISCLASSIFICATION ERROR	% NotAvailable	ORIGINAL MODEL'S MISCLASSIFICATION ERROR
random forest	0.112	19.73	0.192
ridge logistic regression	0.174	6.59	0.201

Table 1. – Mean misclassification error and mean percentage of NotAvailable predictions generated with ALOOM for each model type as well as the results of the original model F().

It may be seen from Table 1, and demonstrated by Krstajic *et al.* [6] on other datasets as well, that the ALOOM approach generates a different percentage of NotAvailable predictions, depending on the type of model.

Krstajic *et al.* [6] have found that the widths of ALOOM individual prediction intervals differ between validation samples. This means that the ALOOM individual prediction intervals may provide a way of assessing the stability for each prediction.

In Table 2 we show examples of pairs of compounds where a compound with original predicted probability closer to 0.5 was not categorised as NotAvailable, while another compound with original predicted probability further away from 0.5 was categorised as such.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 August 2021 doi:10.20944/preprints202108.0521.v1

MODEL TYPE	COMPOUND NAME	ORIGINAL MODEL'S PREDICTED PROBABILITY	ALOOM PREDICTION	ALOOM MINIMUM PROBABILITY	ALOOM MAXIMUM PROBABILITY
random forest	64	0.572	positive	0.510	0.664
random forest	1	0.690	NotAvailable	0.494	0.706
ridge logistic regression	214670	0.454	negative	0.398	0.498
ridge logistic regression	214959	0.383	NotAvailable	0.249	0.559

Table 2. – Examples of pairs of compounds where a compound with an originally predicted probability closer to 0.5 was not categorised as NotAvailable, while another compound with an original predicted probability further away from 0.5 was categorised as such. ALOOM's minimum and maximum predicted probabilities are also shown.

4.3. Active learning

The author [7] has simulated the process of updating current binary classification models with samples currently predicted as NotAvailable, and his initial results show that on average it is better to update the model with samples that are currently NotAvailable rather than with randomly selected samples.

5. Discussion

There are benefits and costs of introducing the third category in binary classification settings. In our opinion, the costs are obvious, while the benefits need to be considered case by case. At the moment, we are only aware of potential benefits of introducing NotAvailable in situations where a human needs to make a decision based on the predicted probability. Our understanding of Tversky and Kahneman [3][4] is that instead of providing the probability to a human decision maker, we ought to consider that it might be sufficient to inform her/him of the following three possible answers:

1) positive - "more probable than not"

2) NotAvailable - "as probable as not"

3) negative - "less probable than not"

Furthermore, our understanding of Tversky and Kahneman [3][4] is that they have demonstrated that the issue of predicted probabilities being well calibrated might be an overkill in situations where a human needs to make decision. It is worth noting that Patrick Suppes's simple probability model ("surely true", "more probable than not", "as probable as not", "less probable than not", "surely false") satisfies the Kolomogorov axioms [5][10].

We consider that there are potential benefits of using the ALOOM approach and we shall discuss them separately.

5.1. Costs

5.1.1. Less useful

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 August 2021 doi:10.20944/preprints202108.0521.v1

In our view, the main cost associated with the introduction of NotAvailable as an answer is that a binary classification model might become less useful. It is very difficult to justify the existence of a predictive model if the majority of its answers are NotAvailable. Furthermore, there is the issue of measuring the ignorance of the model, which in our opinion means calculating the percentage of NotAvailable answers.

In machine learning practice there is a procedure called nested cross-validation which may be used for assessing the quality of predictive models prior to their testing [1]. One may foresee a similar procedure for estimating the proportion of NotAvailable answers, but we can imagine that there might be additional complications.

5.1.2. How to compare models

As we have shown, there are models (such as random forest) that generate more NotAvailable predictions than others (such as linear models). We currently do not have an answer as to how to compare them. We do, however, think that the NotAvailable predictions ought to be introduced after the original model is selected, and that the performance of the original model on the validation set ought to be reported.

5.2. Potential benefits

5.2.1. Binary predictions but diverse set of actions

In our view, the main benefit accrued from introducing NotAvailable is that we can identify some answers that should not be taken into consideration. There are real life situations where decisions may depend on answers from a binary classification model, and we have several courses of actions at our disposal.

In medical diagnostics a binary classification model may predict whether a person has an aggressive cancer or not. In toxicology a binary classification model may predict whether a compound is toxic or not. Those are situations where we have true or false predictions. However, in situations where the resulting actions are not strictly binary (to pull a trigger or not), where we have various options to execute, then having a NotAvailable answer may be very beneficial. In the case of toxicology, all chemical compounds predicted as NotAvailable would automatically be sent to the laboratory for in vitro screening. In the case of cancer predictions, a patient with a NotAvailable prediction, regarding the existence of an aggressive cancer, may provide the basis for an oncologist to seek additional tests. Furthermore, even in situations where we have strictly binary actions, a NotAvailable answer ought to be useful to a decision maker, so that he/she would not take it into consideration.

5.2.2. Active learning

If we plan to continuously update and improve a binary classification model, then we consider that the introduction of the NotAvailable answer may be helpful in the long run. We

 doi:10.20944/preprints202108.0521.v1

consider that, figuratively speaking, we should supply it with answers for questions it does not know.

5.3. Potentical benefits of ALOOM

The ALOOM is a non-parametric approach where the model, so to speak, defines its own "blind spot". It provides us with the ALOOM individual prediction interval for each predicted sample. Currently there are no theories which would support the use of ALOOM.

5.3.1. ALOOM individual prediction intervals

Even though we do not have any theory which would explain the meaning of the ALOOM individual predictions intervals, we consider them to be very informative. In Table 2 we have shown the compound 214959 with the ALOOM individual prediction interval to be (0.249, 0.559). This means that by removing one training sample from over 2000 training samples we created a model which predicted the probability of compound 214959 being mutagen to be 0.249, but if we were instead to remove another training sample then the predicted probability of the compound 214959 being mutagen would be 0.559. We think that this is very informative in the investigation of this situation. What is so specific about the compound 214959? The removal of which two compounds affected the difference in the predicted probabilities of 214959? Looking into these details might provide us with some insights as to how the model is working in practice.

5.3.2. Investigate the effect of the removal of each training sample

We consider that ALOOM might provide an insight into which training samples we should consider removing from our training data set. It might be useful to understand how the removal of certain training samples in ALOOM affected the correct predictions to become NotAvailable, and vice-versa the incorrect predictions to become NotAvailable. The analysis might provide us with an insight into potential influencers in the training data set.

5.3.3. Active learning

The author [7] has shown the potential benefit of using ALOOM in selecting future samples to update the current binary classification model.

6. Conclusion

The introduction of an "I don't know" answer in binary classification settings will inevitably make the process of creating and assessing classifiers more complicated. We consider that there are situations where the price of the complication is worthwhile. As far as we are concerned, an obvious example would be the use of medical diagnostics tools that predict whether a person has an aggressive cancer or not.

We consider that it is the responsibility of the creator of a binary classifier to make sure that the user is informed when the predictive model should not be taken into consideration. The creator of the predictive model ought to understand better than a user when that might happen. In our opinion, by merely providing a predicted probability, creators are to some extent relieving themselves of their responsibility.

The ALOOM approach, which we suggested and analysed, is a simple example of how to create a predictive model that, loosely speaking, knows when it does not know. We would like to emphasise that the ALOOM approach is not a final solution, but a proof of concept that it is possible to create a binary classification model with a meaningful "I don't know" answer.

Acknowledgments: The author would like to thank his mother, Linda Louise Woodall Krstajic, for correcting English typos and language improvements in the text.

Competing interests: The author declares that he has no competing interests.

Funding: The research was conducted in the author's spare time. No funding received.

References

1. Krstajic, Damjan, et al. "Cross-validation pitfalls when selecting and assessing regression and classification models." *Journal of cheminformatics* 6.1 (2014): 1-15.
2. Dawid, A. Philip. "Calibration-based empirical probability." *The Annals of Statistics* (1985): 1251-1274.
3. Tversky, Amos, and Daniel Kahneman. "Judgment under uncertainty: Heuristics and biases." *science* 185.4157 (1974): 1124-1131.
4. Tversky, Amos, and Daniel Kahneman. "Advances in prospect theory: Cumulative representation of uncertainty." *Journal of Risk and uncertainty* 5.4 (1992): 297-323.
5. Salsburg, David. *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Macmillan, 2001.
6. Krstajic, Damjan, et al. "Binary classification models with" Uncertain" predictions." *arXiv preprint arXiv:1711.09677* (2017).
7. Krstajic, Damjan. "Non-applicability Domain. The Benefits of Defining "I Don't Know" in Artificial Intelligence." *Artificial Intelligence in Drug Discovery* 75 (2020): 102.
8. Kazius J, McGuire R, Bursi R. Derivation and validation of toxicophores for mutagenicity prediction. Journal of medicinal chemistry. 2005 Jan 13;48(1):312-20.
9. QSARdata. https://cran.r-project.org/package=QSARdata
10. Kolmogorov, Andreĭ Nikolaevich, and Albert T. Bharucha-Reid. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.