Imprecise oracles impose limits to predictability in supervised learning

Anjali Sifar[1] , **Nisheeth Srivastava**[1]
[1]Indian Institute of Technology, Kanpur
{sanjali, nsrivast}@iitk.ac.in

Abstract

Supervised learning operates on the premise that labels unambiguously represent ground truth. This premise is reasonable in domains wherein a high degree of consensus is easily possible for any given data record, e.g. in agreeing on whether an image contains an elephant or not. However, there are several domains wherein people disagree with each other on the appropriate label to assign to a record, e.g. whether a tweet is toxic. We argue that data labeling must be understood as a process with some degree of domain-dependent noise and that any claims of predictive prowess must be sensitive to the degree of this noise. We present a method for quantifying labeling noise in a particular domain wherein people are seen to disagree with their own past selves on the appropriate label to assign to a record: choices under prospect uncertainty. Our results indicate that 'state-of-the-art' choice models of decisions from description, by failing to consider the intrinsic variability of human choice behavior, find themselves in the odd position of predicting humans' choices better than the same humans' own previous choices for the same problem. We conclude with observations on how the predicament we empirically demonstrate in our work could be handled in the practice of supervised learning.

1 Introduction

As machine learning models are applied to an increasing proliferation of prediction tasks it is puzzling to find high predictive power in domains wherein even human experts are known to be highly imprecise. For example, it is well known that the interpretation of chest X rays for diagnosing pneumonia is a highly variable exercise, with inter-observer agreement between radiologists measured to be around Cohen's $\kappa = 0.4$ across multiple studies. In other words, given binary class labels (pneumonia or not), it is approximately as likely to expect any two radiologists to converge to a positive diagnosis of pneumonia from an X-ray as to flip a coin [Wootton and Feldman, 2014].

How are we then to interpret state-of-the-art supervised learning models of pneumonia detection, which report extremely high validation set accuracy in excess of 90% [Stephen *et al.*, 2019]? Similarly, how do we parse claims of algorithms demonstrating superior agreement to chest X ray dataset labels than a panel of radiologists [Rajpurkar *et al.*, 2017]? More generally, how do we make sense of situations wherein (a) the only source of ground truth is human judgment, and (b) human predictive ability for other humans' choices is exceeded by algorithms' ability to predict human choices? This is the problem we discuss in this paper[1].

To characterize the problem mathematically, assume a fractional confusion matrix emerging from a binary classification task

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix},\ a+b+c+d=1.$$

With some algebra, it is possible to specify the F1-score corresponding to this confusion matrix as

$$F = \frac{2a}{2a+b+c}, \tag{1}$$

and it's Cohen's κ as

$$\kappa = \frac{2(ad-bc)}{2(ad-bc)+b+c}. \tag{2}$$

The term $b+c$ is simply the error fraction in the classification. When a classifier is perfect, $b=c=0$, and it is trivial to see that $F=\kappa=1$, independent of the values of $\{a,d\}$. However, notice that a simple rearrangement of Equation 2 gives us,

$$b+c = 2(ad-bc)(\kappa^{-1}-1). \tag{3}$$

Since the determinant of the confusion matrix $(ad-bc)$ is essentially a measure of how much discriminability is possible in the dataset for human annotators, it is not possible to drive it arbitrarily low without compromising on annotation quality. Therefore, Equation 3 places an interesting limit on the minimal (or Bayes) error expected out of a binary classification problem [Hastie *et al.*, 2009], given empirical estimates of κ. In particular, the error fraction rises hyperbolically with respect to κ. The error fraction magnitude, in turn, places an upper bound on the maximum possible F1 score. That is, for a particular value of κ, no matter what values of

[1]This work is primarily based on our paper [Sifar and Srivastava, 2020] which won the Marr prize at CogSci 2020.

$\{a,d\}$ we use, the highest achievable F1 score will be a value lower than 1.

As we point out above, we increasingly see extremely good predictions emerging from prediction algorithms for domains wherein κ is known to be low, suggesting that the corresponding ground truth predictive ability of annotators is also low. Thus, prediction algorithms appear to show a negative performance gap with respect to annotators.

We recently characterized the presence of such a negative performance gap between human and model judgments for risky choices in economic settings [Sifar and Srivastava, 2020]. Beginning with Erev et al. [Erev *et al.*, 2010], tournaments have been conducted by allowing teams to fit choice models to human choices made on some certainty equivalence problems [Farquhar, 1984], and winning models are identified as the ones that "most accurately" predict human choices for a different set of problems. The empirical success of this research program, given its pure predictive emphasis, is measured in terms of the correlation of model predictions with human choices. Choice models developed through these tournaments have gone from explaining about 70% of the variance in human choices, as in the baseline models used in Erev et al. [Erev *et al.*, 2010] to explaining more than 90% of the variance in human choices, as in the BEAST model presented in [Erev *et al.*, 2017]. Machine learning models built using features identified as important by BEAST are able to approach test set values even more closely [Bourgin *et al.*, 2019].

As with the pneumonia example above, this empirical success is more than a little surprising, given the irreducibly stochastic nature of risky choices [Bhatia and Loomes, 2017]. If someone asks you to either pick 20 tokens of cash for certain or a gamble that will pay 100 tokens 20% of the time, it is very likely that your response may vary across multiple elicitations [Luce *et al.*, 1965]. Thus, if someone uses one of these elicitations to construct a dataset to fit a model of decisions under risk, exceptionally high agreement between the model and the dataset labels would actually mean that the model offers a poor explanation for the behavior, since it shows less variability in choices than humans actually do. This leads to an obvious question: what is the highest model-human agreement one could expect in a model of risky economic choice that would be consistent with real human judgments? In [Sifar and Srivastava, 2020], we answered this question via a set of behavioral choice experiments to calculate within-subject consistency on risky choice problems, as we summarize below.

2 Finding limits to predictability

2.1 Design

A set of expectation-matched risky choice problems were presented to each participant, at a gap of at least a week over the course of three weeks Two experiments were conducted, testing for choice consistency in decisions from description and experience respectively.

In each experiment, each participant was presented with 30 problems per week. Half of the problems presented in Week 1 were repeated during Week 2, and the other half repeated in Week 3. Problem order and selection was randomized across participants, and within participants for repeat presentations. Here, we refer to the first instances of problem presentation for a given participant as fresh problems, while second presentations will be called repeated problems.

The problem space used in both experiments is the Estimation set in the first Technion Choice Prediction Tournament consisting of 60 problems [Erev *et al.*, 2010]. University students were recruited to run the experiment which was hosted online and participants were able to participate at their convenience. Email reminders were sent every week to all participants. The study protocol was reviewed and approved by an IRB.

2.2 Decisions from Description

In decisions from description (DFD), for each problem, participants were asked to choose between risky and safe choices, given explicit payoff and probability descriptions using description paradigm. A total of 58 (19 female, 39 male) participants completed the experiment, without compensation. Feedback was only provided for one of the randomly selected problems at the end of each day of the experiment as a notional payoff.

2.3 Decisions from Experience

25 male, 22 female participants completed the 30 problems in Decisions from Experience (DFE) experiment, specified using the sampling paradigm given in [Erev *et al.*, 2010]. Instructions were followed by two practice games. Participants had the opportunity to learn the payoff distribution of each game in "sampling stage" after which they indicated their final preference in the "choice stage". At the end of the experiment, one of the randomly chosen game's outcome (as recorded in choice stage) was scaled between -100 to 100 to make the net payoff between INR 0-200 (inclusive of a participation fee of INR 100).

2.4 Response Variables

For every problem in each experiment, the proportion of people choosing the risky prospect is recorded and hereby we call it risky choice rate (R-rate).

Decisions from experience, however, involve another latent decision of when to stop sampling. We measure the consistency of this decision by recording the number of samples the decision maker takes before committing to the final choice, henceforth called sampling duration (SD).

2.5 Measuring choice consistency

We measured choice consistency by comparing the response variable measures across repeated elicitations of each problem, where repeated elicitation act as the prediction for the first week's observations. At the cohort level, we quantify this by computing the correlation between the R-rate of fresh and repeated elicitations. This measure is additionally attractive for offering a direct interpretation in terms of percentage of variance explained [Erev *et al.*, 2010].

We also report the proportion of agreement P_{Agree}, as calculated in Erev [Erev *et al.*, 2010], as an additional cohort-level measurement of consensus in choices. This is set to 1 for a

problem if both predicted and observed R-rates are greater than or less than 0.5; otherwise it is set to zero. We report this value, averaged across all tested problems, in percentage terms, following convention [Erev *et al.*, 2010].

To understand internal stochasticity at the individual level, we measure the intra-rater reliability using Cohen's κ [Landis and Koch, 1977]. We report median intra-rater reliability across participants, unless stated otherwise.

3 Summary of results

In all analyses reported below, we refer to our data sources as follows. Data collected in our experiments are denoted as coming from RR (Repeated Risk). Data from the training datasets from the first Technion tournament will be referred to TE (Technion Estimation), and from the competition datasets from the same tournament as TC (Technion Competition). We use both DFD and DFE datasets from all these data sources. Data from the mixed paradigm used in Plonsky et.al [Erev *et al.*, 2017] will be referred to using EEP.

Table 1 summarizes the human-human consistency measures for both experiments in the first row. These results are presented alongside human-model consistency measures on our data (second row) as well as empirical measures reported in [Erev *et al.*, 2010; Erev *et al.*, 2017] for comparison.

3.1 Decisions from description

Three important observations stand out. First, when we use the cohort's R-rate calculated during repeated presentations of problems to predict their own R-rate during the first presentation of the same problems, we obtain correlations ranging between 0.67-0.78 and proportions of agreement ranging between 78%-85%. Notably, these values are much lower than corresponding values previously found for model-data comparisons, even for baseline models like cumulative prospect theory [Erev *et al.*, 2010].

Second, the baseline model does not perform expectedly well when we tested it on our data, yielding correlation and agreement proportions of 0.69 and 81% respectively, suggestive of some over-fitting of the original models.

Third, very little intra-level agreement at individual level is seen across participants [Landis and Koch, 1977], with $\kappa = 0.25$ in both repeated sets, consistent with our expectation of high intrinsic stochasticity of human choice behavior [Luce *et al.*, 1965]. The baseline model agrees with the human choice data consistently more while predicting the Technion data sets $\kappa \in \{0.43, 0.51\}$ than human choices in our dataset agrees with itself. To add to the puzzle, when applied to our dataset, the model agrees with the data about as much ($\kappa = 0.23$) as expected by the empirical measurements. This observation is again consistent with over-fitting of the original models.

3.2 Decisions from experience

The main observation here is that the range of human-human correlations and agreement proportions seen in our data includes the corresponding model-human measurements reported on the competition set in Erev [Erev *et al.*, 2010], though not similar measurements seen on the estimation set. Since the winning ensemble model in this tournament was not significantly better than the simple primed sampler baseline, our observation is consistent with the possibility that a simple primed sampler model might be the best possible model for predicting R-rates in the decision from experience task.

Secondarily, as noted above, the DFE paradigm actually involves two decisions per problem presentation - an overt risk preference, and a latent information search stopping decision governing when to stop sampling and make a final choice. As shown in last column of Table 1, human-human correlations for sampling duration in repeated problems for all observations is 0.64, dropping to 0.54 when only observations with sampling duration greater than two are considered. These values indicate reasonable upper bounds on the predictability of sampling duration in decisions from experience.

4 Discussion

The principal contribution of [Sifar and Srivastava, 2020] was a demonstration, using a test-retest experimental paradigm, that state-of-the-art prediction models of risky choice predict individuals' choices in economic risky choice experiments better than the individuals' own previous history of responding to the exact same choice problem. It is possible that subtly over-fitting to the validation set in a tournament setting accounts for the negative performance gap seen in this setting [Dwork *et al.*, 2015]. In this condition, the large number of degrees of freedom in model architecture design means that a process of iterative development of models that yield improvement in tournament leader-board ranks is effectively equivalent to using validation set error to search the space of available models. Restricting access to validation sets during the process of model development is a simple technical solution for this problem.

However, the point of genuine interest in this work lies in its amplification of a really simple point: the process of attaching labels to observations is stochastic [Anderson, 1991], and the degree of variability in this process in different domains places fundamental limits on the degree of predictability possible in those domains [Hastie *et al.*, 2009]. For domains with low or zero variability, e.g. object recognition in images, domains where access to some deterministic ground truth is available etc. it is possible to ignore the variability and interpret supervised learning results reasonably. However, for domains with high variability in the label assignment process, ignoring this variability is problematic, since it becomes difficult to claim that a model that makes more accurate predictions on our dataset's labels is a better model of the underlying reality. If my model is trained using the tuple $X, y, y \in \{0,1\}$ based on one elicitation of y, can I really say I have a useful model if reality approaches $y \sim Bern(0.5)$?

4.1 Negative performance gaps are ubiquitous

Ignoring the stochasticity of data labels has led to a proliferation of highly accurate machine learning predictions for problems, surpassing empirically estimated Bayes error estimates. In parallel with the pneumonia example cited above, multiple authors have proposed high accuracy depression detection models that learn to labels arising from DSM or PHQ based diagnoses of patients with respondents' EEG signals [Li *et*

Table 1: **Choice consistency analysis** Choice agreement indicators to check the test-retest reliability between the repeated problems in the first row. Remaining rows indicate similar analysis using different models as predictors on the observed data sources.
BM, WM = Baseline, Winning model for respective paradigms as given in [Erev et.al 2010]
' fresh elicitations, * as claimed in [Erev et. al, 2010], ** as claimed in [Erev et. al, 2017]

				DFD			**DFE**			
Data Sources	Subset	Observed	Predictor	Correlation	κ	P_{Agree}	Correlation	κ	P_{Agree}	SD
RR	A	Week 1	Week 2	**0.78**	0.25	85%	**0.83**	0.40	90%	0.64
	B	Week 1	Week 3	**0.67**	0.25	78%	**0.73**	0.32	88%	0.65
RR	Fresh	A′, B′, C, D	BM	0.69	0.23	81%	0.84	0.23	92%	-
TE	-	TE	BM	0.85*	0.43	95%*	0.88*	0.25	95%*	-
			WM	0.92*	-	88%*	0.92*	-	95%*	-
TC	-	TC	BM	0.86*	0.51	93%*	0.80*	0.19	82%*	-
			WM	0.94*	-	90%*	0.80*	-	83%*	-
EEP	-	EEP	BEAST	0.95**	-	-	-	-	-	-

al., 2019], speech patterns [Low *et al.*, 2020], social media usage dynamics [De Choudhury *et al.*, 2013] and a variety of other sources of information. State-of-the-art predictions attain F-scores of close to 0.9, whereas test-retest κ estimates for the most recent DSM-5 instrument are estimated at a much more modest 0.47 [Chmielewski *et al.*, 2015].

Beyond domains for which consistency or test-retest reliability estimates are empirically available, it is deeply problematic to observe supervised learning methods being used in criminological settings, e.g. trying to predict the type of crime that someone may potentially commit, based on a retrospective analysis of crimes committed by people with similar psychological profiles [Watts *et al.*, 2021]. Such projects are fundamentally flawed because the math underlying the prediction model only makes sense if we assume that a person in the training dataset with a psychological profile X would have committed crime y as opposed to some other y' in any possible alternative circumstances, or even if given the chance again in the same circumstances.

This last example further accentuates our basic point about the epistemic limitations of supervised learning as currently practised: the deterministic mapping from features to labels, in conjunction with limited sampling from the labeling process, restricts Bayes error calculations [Hastie *et al.*, 2009]. It also points to an alarming prospect: supervised learning systems become superficially more attractive precisely when they present large negative performance gaps, i.e. when the machine appears to make clear and confident predictions about things that humans only have muddled ideas about. Thus, for example, several units of the US judicial system are using COMPAS, a putatively intelligent supervised learning algorithm built using 137 demographic factors that predicts offenders' recidivism risk with about 65% accuracy, as opposed to a simple two factor linear regression model that offers the same degree of accuracy [Dressel and Farid, 2018]. The opacity and complexity of the model furnishes its predictions with an undeserved patina of competence. To avoid such false confidence in AI systems operating in high stakes domains, it is vital that the stochasticity of dataset labels be examined rigorously and openly.

4.2 Ways forward

The reproducibility crisis in artificial intelligence is widely acknowledged [Hutson, 2018], and several institutional reforms have been set in motion to proactively address it, including code and data sharing initiatives [Pineau *et al.*, 2020]. Our work suggests an additional aspect that could assist in this exercise - designing mechanisms to incentivize and promote replica data annotation efforts, particularly for data domains where ground truth is not objectively available. At a minimum, such efforts will allow us to estimate limits to predictability in such domains, and allow more reasonable interpretations of machine learning models' performance.

Some variability in label assignment can be reduced by averaging inputs from multiple annotators, thereby reducing *measurement* noise [Karimi *et al.*, 2020]. However, as our work demonstrates, in several important domains, such aggregation will still not remove *process* noise intrinsic to the labelling process.

Acquiring an empirically informed appreciation of limits to predictability in important domains will permit considerations of explainability and interpretability to drive the discourse on model selection [Rudin, 2019]. If notional predictive ability claimed by complex, uninterpretable models exceeds reasonable estimates of predictability upper bounds in such domains, it will be much easier to justify shifts to simpler models that are interpretable by design [Rudin, 2019].

So long as supervised learning remained a mathematical curiosity, it was possible to remain agnostic about the reliability of data annotations. Today, when supervised learning is increasingly being used in consequential real-world applications, practitioners and consumers can no longer afford this luxury [O'neil, 2016].

References

[Anderson, 1991] John R Anderson. The adaptive nature of human categorization. *Psychological review*, 98(3):409, 1991.

[Bhatia and Loomes, 2017] Sudeep Bhatia and Graham Loomes. Noisy preferences in risky choice: A cautionary note. *Psychological review*, 124(5):678, 2017.

[Bourgin *et al.*, 2019] David D Bourgin, Joshua C Peterson, Daniel Reichman, Thomas L Griffiths, and Stuart J Russell. Cognitive model priors for predicting human decisions. *arXiv preprint arXiv:1905.09397*, 2019.

[Chmielewski *et al.*, 2015] Michael Chmielewski, Lee Anna Clark, R Michael Bagby, and David Watson. Method matters: Understanding diagnostic reliability in dsm-iv and dsm-5. *Journal of abnormal psychology*, 124(3):764, 2015.

[De Choudhury *et al.*, 2013] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 2013.

[Dressel and Farid, 2018] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[Dwork *et al.*, 2015] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.

[Erev *et al.*, 2010] Ido Erev, Eyal Ert, Alvin E Roth, Ernan Haruvy, Stefan M Herzog, Robin Hau, Ralph Hertwig, Terrence Stewart, Robert West, and Christian Lebiere. A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, 23(1):15–47, 2010.

[Erev *et al.*, 2017] Ido Erev, Eyal Ert, Ori Plonsky, Doron Cohen, and Oded Cohen. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, 124(4), 2017.

[Farquhar, 1984] P. H. Farquhar. State of the art—utility assessment methods. management science. *Management Science*, pages 1283–1300, 1984.

[Hastie *et al.*, 2009] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[Hutson, 2018] Matthew Hutson. Artificial intelligence faces reproducibility crisis, 2018.

[Karimi *et al.*, 2020] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

[Landis and Koch, 1977] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[Li *et al.*, 2019] Xiaowei Li, Xin Zhang, Jing Zhu, Wandeng Mao, Shuting Sun, Zihan Wang, Chen Xia, and Bin Hu. Depression recognition using machine learning methods with different feature generation strategies. *Artificial intelligence in medicine*, 99:101696, 2019.

[Low *et al.*, 2020] Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116, 2020.

[Luce *et al.*, 1965] Robert Duncan Luce, Patrick Suppes, et al. *Preference, utility, and subjective probability*. Wiley New York, 1965.

[O'neil, 2016] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

[Pineau *et al.*, 2020] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *arXiv preprint arXiv:2003.12206*, 2020.

[Rajpurkar *et al.*, 2017] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[Sifar and Srivastava, 2020] Anjali Sifar and Nisheeth Srivastava. Limits on predictability of risky choice behavior. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Cognitive Science Society, 2020.

[Stephen *et al.*, 2019] Okeke Stephen, Mangal Sain, Uchenna Joseph Maduh, and Do-Un Jeong. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of Healthcare Engineering*, 2019:4180949, Mar 2019.

[Watts *et al.*, 2021] Devon Watts, Heather Moulden, Mini Mamak, Casey Upfold, Gary Chaimowitz, and Flávio Kapczinski. Predicting offenses among individuals with psychiatric disorders-a machine learning approach. *Journal of Psychiatric Research*, 138:146–154, 2021.

[Wootton and Feldman, 2014] Dan Wootton and Charles Feldman. The diagnosis of pneumonia requires a chest radiograph (x-ray)—yes, no or sometimes? *Pneumonia*, 5(1):1–7, Dec 2014.