Trustworthy Algorithmic Decision-Making via Transparent Machine Learning

by
Emrah Akyol, SUNY Binghamton
eakyol@binghamton.edu

Keywords: Trustworthy algorithmic decision making; Security, privacy and abuse prevention; human-in-the-loop machine learning; game theory; information theory

1 Motivation

Classical Machine Learning (ML) algorithms operate under several important yet often unstated assumptions: that algorithm designer and data providers share a known common goal, and that they are both acting truthfully towards the goal (i.e., providers transmit truthful information, and the designer only uses it towards the stated, jointly agreed upon, goal). However, recent technical investigations have shown the limitations of such a paradigm, and revealed that many practical uses of ML involve, e.g., "baked-in" discrimination and/or hidden motives [1]. One approach (but not necessarily the only one) to limit such occurrences of invidious ML is to enforce some kind of "transparency", whereby the goal and inner-workings of the algorithm would be made public [2, 3]. Such transparency is desirable for the data providers, since it prevents, in principle, designers from acting untruthfully. However, making an algorithm public could also open the door to *gaming*, and adversely affect the ML outcome. In other words, there are strong incentives against transparency, even for a designer fully intent on working truthfully, and it is thus essential to quantify the possible costs for system designers, if algorithmic transparency is truly to become a widespread requirement mandated by regulation.

Gaming, broadly defined, is strategic use of methods that, while not against the rules, give the individual an unintended advantage. Gaming behavior has been well-known in finance and it has lead to a classical principle of financial policy making, known as the Goodhart's law: *If a measure becomes the public's goal, it is no longer a good measure.*[1] Gaming poses a formidable threat to the extensive utilization of ML to make accurate decisions about individuals in areas such as employment, health, commerce and education. For example, concerns of gaming are often used as a reason for keeping the algorithms used credit scoring, such as FICO scores, a secret. Secrecy is obviously not a robust solution, and hence the scoring algorithms change in time to mitigate the impact of gaming, which in turn results in inconsistencies in financial monitoring (cf. [2]).

Note that such strategic intent, such as the one in gaming, can also be inferred by the algorithm designer, i.e., not only the data provider can learn the details of the algorithm, the algorithm designer can also learn the strategic objective of the data provider and his bias (in statistical sense). This point of view enabled us to model the problem in a Bayesian framework, as a non-cooperative game. **The main objective here is to develop and analyze gaming-aware (robust, in the sense of game-theoretic equilibria) fully transparent ML algorithms that take strategic intent (modeled in a Bayesian framework) into account, using tools from optimization, game theory and information theory.** A byproduct of our approach is to characterize and evaluate the "price of transparency" (PoT) or specific classes of ML algorithms, by treating the ML task as a non-cooperative game between data providers and the designer, and considering "transparency" as a widening of the common knowledge available to both players, with respect to the non-transparent version[5].

Such problems were also considered in a recent study [6], albeit with substantially different approach: In the setting considered in [6], the bias incurs an explicit cost to the users (altering the true data has a penalty to the data providing agent). Here, we do not assume such a cost, but instead we assume both agents' strategies are transparent, i.e., the classifier designer knows, in a statistical sense, how much the users are misreporting. Our approach enables robust ML algorithms that would work in the wild, without any assumptions on the structure of the payoff functions and a penalty for misreported data[7].

[1] A good example to gaming is related to the well-known social study[4] which shows that a student's success can be very well predicted from the number of books in the parents' household. However, this feature cannot be used in actual school admissions simply because when it is publicly known that this is a factor in admission (if the algorithm is transparent), the parents will simply cheaply obtain several (unread) books to increase chances of their children admitted to top schools.

2 Technical Approach & Preliminary Results

2.1 Preliminary Model

Our approach is to leverage the recent results on strategic communication in [8] to devise transparent gaming-aware ML algorithms. Here, we consider the Stackelberg equilibrium of a nonzero-sum signaling game: the machinery of the algorithm (the strategy of the receiver in the strategic communication game) is known to the strategic individuals (the senders of the communication game) who provide the training and test data. In return, the strategic intent and the functional form of the data generation (how the bias is statistically introduced into the data), are known to the ML algorithm designer. In game theory terms, this setting can be viewed as a signaling game between two agents: a sender (strategic data provider, the leader of the game) and a receiver (the ML algorithm designer, the follower). The objective of the receiver is to design the ML algorithm (say a classifier). The objective of the sender is deceive the receiver to render it's output to be close to $X + \theta$ (e.g., in the mean-squared error sense), where θ is a bias random variable, correlated with X. The statistics of X and θ as well as the strategies of both agents are common knowledge (known to both agents). The data provider (sender) is restricted to pure strategies and plays first as the leader, and the algorithm designer (receiver) is the follower, who observes the sender's choice of pure strategies and plays accordingly to minimize its own distortion with the knowledge of the mappings of the sender.

The value of this game, in conjunction with optimal strategies for both agents, is used in the proposed metric to measure the price of transparency (PoT). We define PoT as the ratio of the designer costs at the Stackelberg equilibrium (i.e., the algorithm is transparent) and the non-strategic setting (i.e., the algorithm is not transparent).

2.2 Strategic Communication

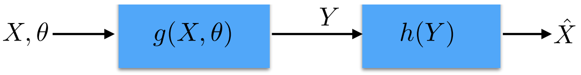

Figure 1: The basic model of strategic communication.

In its simplest form[2], the problem of strategic communication is captured by Figure 1. A "receiver" wants to estimate a random variable $X \in \mathbb{R}^{n_x}$ that she cannot observe directly, although its distribution is known. A "sender" has access to this variable and can transmit a message $Y \in \mathbb{R}^{n_y}$ to the receiver. Based on this available information, the receiver computes the optimal estimator as $\hat{X} = \mathbb{E}(X|Y)$. *Knowing that this is the goal of the receiver*, the sender wants to choose his message Y, so as to minimize $\mathbb{E}$ where $\theta \in \mathbb{R}^{n_x}$ is a privately known bias. In other words, the goal of the sender is to provide a message to the receiver that leads her to believe (based on its estimate $\hat{X}$) that the state is $x + \theta$ rather than x. The question then is to determine which message is sent by the sender and characterize the corresponding estimation error.

One interesting result pertains to the case of jointly Gaussian variables. We have shown in [8] that optimal strategies are linear, i.e., in the form of $g(X, \theta) = X + \alpha^T \theta$ where α is a column vector that can be computed using the problem parameters. One particularly observation derived from this result is that, for the scalar case, even though the sender wants to mislead the receiver into believing that the state is $X + \theta$, it is not optimal for him to "flat-out lie" and report $x + \theta$ in lieu of x.

2.3 Proposed Research

Although admittedly simple, the basic model and classification example for the specific quadratic-Gaussian case allowed us derive several structural insights–some rather non-intuitive which can be summarized as: i) There exists a unique equilibrium. At the equilibrium, the strategic classification problem can be decomposed into two separate problems: strategic communication followed by a classical (non-strategic) classification problem. Hence, the optimal strategy for the data provider is the same as that of the sender in strategic communication game; and the structure of the optimal strategic classifier is tandem concatenation of the receiver strategy in strategic communication and off-the-shelf classifier applied to the estimate.ii) Equilibrium achieving data providing (sender) strategy is linear in state and bias.

The proposed research is two-fold: i) We will explore these findings can be generalized to more practical settings. Particularly, we will analyze which (if any) of the properties of the equilibrium continue to hold when the Gaussianity and quadratic cost function assumptions are relaxed. We will numerically design optimal strategies when they are non-linear. ii) We will design experiments to learn the statistics and bias introducing mechanisms for different practical scenarios.

[2] There exists a substantial amount of literature in information economics on communicating information in sender-receiver games: disclosure of verifiable information [9], or cheap talk [10], or information disclosure[11, 12]. Here, we adopt the information disclosure model in [11] and [12] since it fits best to our setting.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks," *ProPublica, May*, vol. 23, 2016.

[2] D. K. Citron and F. Pasquale, "The scored society: due process for automated predictions," *Washington Law Review*, vol. 89, 2014.

[3] F. Pasquale, *The black box society: The secret algorithms that control money and information*, Harvard University Press, 2015.

[4] M. Evans, J. Kelley, J. Sikora, and D. Treiman, "Family scholarly culture and educational success: Books and schooling in 27 nations," *Research in social stratification and mobility*, vol. 28, no. 2, pp. 171–197, 2010.

[5] E. Akyol, C. Langbort, and T. Başar, "Price of transparency in strategic machine learning," *arXiv preprint arXiv:1610.08210, presented at the Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT ML 2016, November 18, 2016; New York University, NY)*, 2016.

[6] M. Hardt, N Megiddo, C. Papadimitriou, and M. Wootters, "Strategic classification," in *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. ACM, 2016, pp. 111–122.

[7] E. Akyol, "The game of misinformation," in *55th Annual Allerton Conference on Communication, Control, and Computing*, Oct 2017.

[8] E. Akyol, C. Langbort, and T. Başar, "Information-theoretic approach to strategic communication as a hierarchical game," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 205–218, 2017.

[9] P. Milgrom, "What the seller won't tell you: Persuasion and disclosure in markets," *The Journal of Economic Perspectives*, vol. 22, no. 2, pp. 115–131, 2008.

[10] V. Crawford and J. Sobel, "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, pp. 1431–1451, 1982.

[11] L. Rayo and I. Segal, "Optimal information disclosure," *Journal of Political Economy*, vol. 118, no. 5, pp. 949–987, 2010.

[12] M. Gentzkow and E. Kamenica, "Bayesian persuasion," *American Economic Review*, vol. 101, no. 6, pp. 2590–2615, 2011.