Towards a Theory of Incentives in Machine Learning

ARIEL D. PROCACCIA
School of Computer Science and Engineering, The Hebrew University of Jerusalem

1. INTRODUCTION

The connection between machine learning and economics is, I feel, quite natural. There is a growing body of work that lies at the intersection of the two fields, but most of this work focuses on applying machine learning paradigms to economic problems. Examples include prediction of consumer behavior [Kalai 2003; Beigman and Vohra 2006], automated design of voting rules [Procaccia et al. 2007; Procaccia et al. 2008], and reduction of mechanism design problems to standard algorithmic questions [Balcan et al. 2005].

Nevertheless, there are preciously few papers investigating the incentives that, in some settings, govern the learning process itself (see, e.g., Perote and Perote-Peña [2004], Dalvi et al. [2004]); none of them do so in a general machine learning framework. Where, indeed, do strategic considerations come into play in the learning world? In general, a machine learning algorithm receives a (small but hopefully representative) training set consisting of points sampled from an input space and labeled according to some target function; the algorithm outputs a hypothesis that is presumably close to the target function. For simplicity, consider a basic setup where n selfish agents control n disjoint subsets of the input space. The label of each point in the training set is reported by the agent that controls it (whereas the identities of the points controlled by an agent are common knowledge). Crucially, each agent is interested only in the accuracy of the generated hypothesis on its own part of the input space. An agent can influence the outcome of the learning process by misreporting the labels of the points under its control.

The above strategic setup seems relevant, for instance, in the context of decisions taken by a central bank, such as the European Central Bank (ECB). The governing council of the central bank collects information from national bankers (the agents), who in turn gather data on different economic parameters by means of their own institutions. The central bank decides on an economic policy (hypothesis) by using, say, regression learning on the examples provided by the national bankers. The national bankers may thus be motivated to manipulate their portion of the data

Author's address: arielpro@gmail.com. The author is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

in a way that achieves a central policy that is more accurately aligned with the interests of their nation.

Strategic behavior introduces an undesirable bias to the training set, and contaminates the entire learning process. The goal is therefore to design learning algorithms—mechanisms, that is—that are strategyproof, while approximately maximizing social welfare. I believe that this agenda, more than ever before, calls for a synthesis of the two fields, namely Machine Learning and Mechanism Design, and that the results would truly be of interest to researchers from both communities.

2. SOME RESULTS

In recent work [Dekel et al. 2008; Meir et al. 2008] we pursued the above agenda under a more general mathematical model than the one hinted at in the introduction. Each agent holds a probability distribution over the input space, reflecting the importance it attributes to different issues. The agents also hold private functions, defined on the input space, that reflect the ideal outcome from their point of view.

The *risk* of an agent—its disutility, or cost—is the inaccuracy of the hypothesis returned by the mechanism with respect to the private function and distribution of the agent.[1] The definition of risk depends on the output space, among other things, and hence it is defined differently in different settings.

As it turns out, when it comes to strategyproof mechanisms, this learning theoretic setting can be reduced to a much simpler one, in a way that the transformation is valid with high confidence and accuracy given enough examples from the distribution of each agent; we omit the details here. In the simple setting, each agent controls a subset of the training set, and its *empirical risk* is defined only with respect to its own subset. In the following I make this model slightly more concrete, and sketch some results.

2.1 Regression Learning

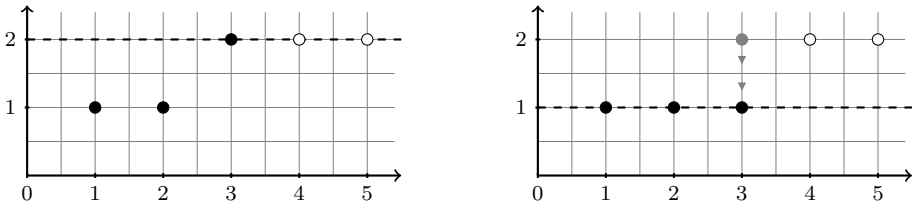
In joint work with Dekel and Fischer [Dekel et al. 2008] we obtained some encouraging results with respect to strategyproof regression learning. In the regression model, the output space is the real line $\mathbb{R}$. The accuracy of a hypothesis is evaluated according to a *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. Common choices of ℓ are the *squared loss*, $\ell(\alpha, \beta) = (\alpha - \beta)^2$, and the *absolute loss*, $\ell(\alpha, \beta) = |\alpha - \beta|$. Given a hypothesis $f \in \mathcal{F}$ returned by the mechanism on a training set $S = \biguplus_i S_i$, where $\mathcal{F}$ is the hypothesis class and S_i is the subset of the training set controlled by agent i, the empirical risk of agent i is

$$\hat{R}(f, S_i) = \frac{1}{|S_i|} \sum_{(\mathbf{x},y) \in S_i} \ell(f(\mathbf{x}), y) \ .$$

We define the global empirical risk as

$$\hat{R}(f, S) = \frac{1}{|S|} \sum_{(\mathbf{x},y) \in S} \ell(f(\mathbf{x}), y) = \frac{1}{|S|} \sum_{i=1}^{n} |S_i| \cdot \hat{R}(f, S_i) \ .$$

[1] "Risk", rather than "cost", is the prevalent term in the machine learning literature.



(a) Truthful dataset. (b) Manipulated dataset.

Fig. 1. ERM can be manipulated in the general regression model even under the absolute loss function. In this example, $\mathcal{X} = \mathbb{R}$ (the x axis), and $\mathcal{F}$ is the class of constant functions over $\mathbb{R}$. Agent 1 controls the black examples, whereas agent 2 controls the white examples. Given the truthful dataset (a), ERM would return the constant function 2, causing agent 1 to suffer a risk of 2 (on its examples at $x = 1$ and $x = 2$). However, the agents may misreport the labels (y values) of their examples. If agent 1 lied about the label of its example at 3 (see (b)), then ERM would return the constant function 1, with agent 1 incurring a risk of only 1 (with respect to the truthful label of its example at 3).

So, the empirical risk is in fact the weighted social welfare. Our mathematically cleanest results are obtained in the very simple, but still nontrivial, setting where each agent only controls one point, that is, $|S_i| = 1$ for all agents i. In this setting, it is possible to show that under the absolute loss function, Empirical Risk Minimization (ERM)—that is, simply choosing a hypothesis that minimizes the global empirical risk, or maximizes social welfare—is group strategyproof (i.e. no coalition of agents can benefit from lying).

THEOREM 2.1. *[Dekel et al. 2008] Assume that $|S_i| = 1$ for all i, and that $\mathcal{F}$ is a convex hypothesis class.*

(1) *Under the absolute loss function, ERM is group strategyproof.*

(2) *Under any superlinear loss function (including the squared loss) and additional very mild technical assumptions on $\mathcal{F}$, ERM is not (even individually) strategyproof.*

Unfortunately, the positive part of Theorem 2.1 does not hold when agents control multiple points; see Figure 1 for an illustration. It is important to note that even in this more general setting, it is possible to make ERM truthful by augmenting it with VCG payments. However, we are interested in obtaining strategyproofness without assuming quasi-linear preferences.

A very intuitive mechanism performs ERM on the dataset of each agent separately, relabels the examples of agent i according to the hypothesis returned by ERM on the dataset of i, and then performs ERM globally on the relabeled dataset. We are able to show the following:

THEOREM 2.2. *[Dekel et al. 2008] Let $\mathcal{F}$ be the class of constant functions over $\mathbb{R}^d$, $d \geq 1$, or the class of homogeneous linear functions over $\mathbb{R}$, and assume ℓ is the absolute loss. Then the above mechanism is group strategyproof and gives a 3-approximation of the global risk. Moreover, no strategyproof mechanism can yield a better approximation ratio.*

Extending this result to more complex hypothesis classes currently seems out of our reach.

2.2 Classification

In joint work with Meir and Rosenschein [Meir et al. 2008] we extended the investigation to the world of classification. In this model, the output space is simply the set $\{+, -\}$. The common loss function used for classification is the 0-1 loss function, which is 0 if the label of a point matches its image under a hypothesis, and 1 otherwise. In other words, the empirical risk of an agent is simply the number of examples under its control that the hypothesis mislabels. Crucially, in this setting it is also possible to obtain a reduction from a general learning theoretic setting—where the private distribution of each agent is sampled—to a simple setting involving only the strategyproof minimization of empirical risk.

On the face of it, obtaining strategyproofness in the classification model seems simpler than in the regression model. For instance, it is quite obvious that an agent that only controls one point does not have an incentive to lie. However, appearances can be deceiving.

Our early results are concerned with a classification setting where $\mathcal{F}$ contains only two hypotheses: the constant positive hypothesis (that labels all the points positively), and the constant negative hypothesis. This (surprisingly nontrivial) setting is motivated in its own right; for instance, consider the central bank example given above, and suppose the bank simply has to make a positive or negative decision about a given issue. We prove:

THEOREM 2.3. *[Meir et al. 2008] Let $\mathcal{F}$ be the class of constant hypotheses, and let ℓ be the 0-1 loss. Then:*

(1) *There exists a (trivial) group strategyproof deterministic mechanism that yields a 3-approximation of the global risk. Moreover, no strategyproof deterministic mechanism can give a better approximation ratio.*
(2) *There exists a (nontrivial) group strategyproof randomized mechanism that yields a 2-approximation of the global risk. Moreover, no strategyproof randomized mechanism can give a better approximation ratio.*

Once again, it seems difficult to obtain similar results under more complex hypothesis classes. We are currently able to put forward a randomized mechanism that, under the assumption that $\mathcal{F}$ is the class of linear functions over $\mathbb{R}$, is strategyproof and yields an approximation ratio of $O(k^2)$, where k is an upper bound on the number of points controlled by any agent. In brief, the mechanism iteratively chooses a random dictator, breaking ties according to the next randomly chosen dictator. We can also demonstrate a lower bound of $\Omega(k)$ for deterministic mechanisms. We conjecture that the optimal truthful approximation ratio, under the above assumption, is $\Theta(k)$.

3. CONCLUSIONS

I believe that the results discussed above are merely the tip of the iceberg, a first step towards a theory of incentives in machine learning. First, we are very far from fully understanding the models described above. Second, a host of other

machine learning models are begging to be explored. Third, conceptually it is not clear that strategyproofness in dominant strategies (while only approximately minimizing risk) is the correct solution concept that one should look at, rather than, say, ex-post Nash equilibrium, Bayes-Nash equilibrium, or regret minimization.

For more information, including a comprehensive presentation of our models and results, the reader is referred to our published papers [Dekel et al. 2008; Meir et al. 2008].

4. ACKNOWLEDGMENTS

I thank Felix Fischer and Reshef Meir for commenting on an earlier version of this letter.

REFERENCES

Balcan, M.-F., Blum, A., Hartline, J. D., and Mansour, Y. 2005. Mechanism design via machine learning. In *Proceedings of the 46th Symposium on Foundations of Computer Science (FOCS)*. 605–614.

Beigman, E. and Vohra, R. 2006. Learning from revealed preference. In *Proceedings of the 7th ACM Conference on Electronic Commerce (ACM-EC)*. 36–42.

Dalvi, N., Domingos, P., Mausam, Sanghai, S., and Verma, D. 2004. Adversarial classification. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD)*. 99–108.

Dekel, O., Fischer, F., and Procaccia, A. D. 2008. Incentive compatible regression learning. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 277–286.

Kalai, G. 2003. Learnability and rationality of choice. *Journal of Economic Theory 113,* 1, 104–117.

Meir, R., Procaccia, A. D., and Rosenschein, J. S. 2008. Strategyproof classification under constant hypotheses: A tale of two functions. In *Proceedings of the 23rd Conference on Artificial Intelligence (AAAI)*. To appear.

Perote, J. and Perote-Peña, J. 2004. Strategy-proof estimators for simple regression. *Mathematical Social Sciences 47*, 153–176.

Procaccia, A. D., Zohar, A., Peleg, Y., and Rosenschein, J. S. 2007. Learning voting trees. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI)*. 110–115.

Procaccia, A. D., Zohar, A., and Rosenschein, J. S. 2008. Automated design of scoring rules by learning from examples. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. To appear.