Incentive Incompatibility of Logistic Regression

Abstract

We study the incentive compatibility problem of multi-class logistic regression. We provide a simple numerical example in which a strategic data provider has the incentive to misreport her private label to increase the classification probability of her true label. In particular, the model trained given her true label classifies her data point incorrectly, whereas the model trained given her misreported label classifies her data point correctly. We show that this incentive incompatibility problem disappears if the logistic model is estimated by minimizing the zero-one loss, and if a Bayes classifier estimated with maximum likelihood is used.

1 Introduction

There are n strategic agents each providing the label of one data point to the principal. The principal is the learner and builds a machine learning model based on the data points provided by the agents. An agent, i, has publicly known feature vector, x_i, and a private discrete label, y_i. The objective of the agent is to maximize the probability that her data point is labeled correctly by the principal's model, and the agent can choose to report $y_i^\dagger$ to achieve the objective, with the possibility of misreporting $y_i^\dagger \neq y_i$. We say a dataset is incentive incompatible with respect to the learner, described by a parametric model, if at least one of the n agents has the incentive to misreport.

The following is the diagram showing a dataset that is incentive incompatible with respect to the multi-class logistic regression model. In the dataset, each of the $n = 18$ agents, i, has a two dimensional feature vector and a private label can take on one of three values: "red", "green", or "blue".

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

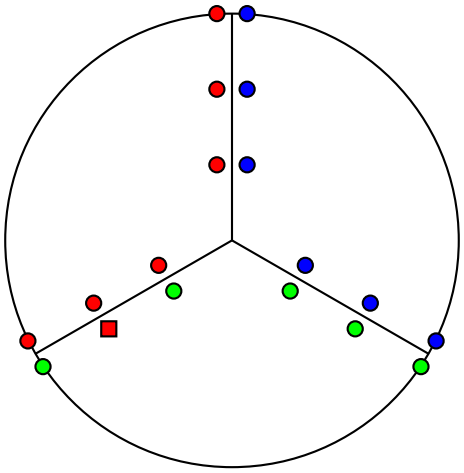

Figure 1: Incentive Incompatible Example

The 18 points are located inside a unit circle, and each point is 0.004 away from the three line segments through the origin that forms angles of 120 degrees between them. There is one point, labeled by a square in the plot, that is on the "incorrect" side of the boundary. Suppose the point corresponds to the feature vector of an agent i with private label "red", then truthfully reporting her label will lead to a multi-class logistic regression model that classifies her point as "green". The probability that this model classifies her point as "red" is 0.3290. However, if the agent misreports her label as "blue", the resulting model classifies her point as "red" with probability 0.4966. Therefore, by lying about her label, the agent can make the principal learn an incorrect model that classifies her point correctly and with a higher probability.

However, this dataset is incentive compatible if zero-one loss is minimized when estimating the logistic parameters, and if the Bayes classifier is used with maximum likelihood estimation. In general, misreporting will always lead to a lower classification probability of the agent's true label for these classifiers.

Previous work on mechanism design for machine learning with strategic data sources focus on designing robust algorithms to incentivize the data providers to report their private data truthfully. Their models mainly differ in the objective and the possible actions of the data providers (agents) and the machine learner (principal).

The first group of papers focuses on principal-agent problems in which each agent's private data point is the agent's type that the agent cannot change. The only action the agents can take is whether to report their private information truthfully.

1. Some models assume the agents' feature vectors are public, but their labels are private. **?**, **?**, and **?** focus on strategy-proof linear regression algorithms and introduced clockwise repeated median estimators, generalized resistant hyperplane estimators, and modified generalized linear squares estimators. **?** investigates the general regression problem with empirical risk minimization and absolute value loss. All the previously mentioned papers assume the labels are continuous variables (regression problems), and **?** assumes the labels are discrete variables (classification problems) and proposes a class of random dictator mechanisms.
2. Some models assume the agents' feature vectors are also private. **?** investigates such problems for linear regressions.
3. Other models do not distinguish between feature vectors and labels. Each agent has a private valuation. These problems are usually modeled as facility locations problems and the solution involves some variant of the Vickrey-Clarke-Groves or Meyerson auction. These include **?**, **?**, **?**, and **?**.

The second group papers focus on moral-hazard problems in which each agent does not have a type but they can choose an action (with a cost) that affects the probability of obtaining the correct label. **?**

focuses on the linear regression problem in this scenario, and **?** and **?** investigates the problem for more general machine learning problems. **?** also discusses a similar problem for general machine learning algorithms.

The last group of papers uses machine learning or robust statistics techniques without game-theoretic models. This group of papers include **?**, **?**.

2 Logistic Regression

2.1 Model and Example

In this section, we assume the principal is training a multi-class logistic (softmax) regression. There are n strategic agents each providing the label of one data point to the principal. An agent, i, with public feature vector, $x_i \in \mathbb{R}^m$, and private discrete label, $y_i \in \{1, 2, ..., k\}$, has the objective of maximizing the probability that her data point is labeled correctly by the principal's model, parameterized by the $m \times (k+1)$ weights (and bias) matrix w. The agent can choose to report $y_i^\dagger$ to achieve the objective, with possibly $y_i^\dagger \neq y_i$. Denoting the weights of the model resulting from the false report from agent i by $w^\star\left(y_i^\dagger\right)$, the agent's objective can be written as,

$$\max_{y^\dagger \in \{1,2,...,k\}} \mathbb{P}\left\{Y = y_i | w^\star\left(y_i^\dagger\right), x_i\right\},$$

where,

$$\mathbb{P}\left\{Y = c | w, x_i\right\} = \frac{e^{z_{i,c}}}{\sum_{c'=1}^{k} e^{z_{i,c'}}},$$

$$z_{i,c} = \sum_{j=1}^{m} w_{j,c} x_{i,j} + b_c, \text{ for } c \in \{1, 2, ..., k\}.$$

The principal is not strategic and he maximizes the likelihood of the data,

$$\max_{w} \sum_{i=1}^{n} \log\left(\mathbb{P}\left\{Y = y_i^\dagger | w, x_i\right\}\right).$$

We consider the case without a coalition of a group of agents, so only one agent is misreporting at a time, and use the following notations,

$$w^\star = \arg\max_{w} \sum_{i=1}^{n} \log\left(\mathbb{P}\left\{Y = y_i | w, x_i\right\}\right)$$

$$w^\star\left(y_i^\dagger\right) = \arg\max_{w} \log\left(\mathbb{P}\left\{Y = y_i^\dagger | w, x_i\right\} + \sum_{i'=0, i' \neq i}^{n} \log\left(\mathbb{P}\left\{Y = y_{i'} | w, x_{i'}\right\}\right),\right.$$

Definition 1. A dataset is incentive incompatible with respect to a learner if there exists at least one agent i, and some $y_i^\dagger \neq y_i$ such that,

$$\mathbb{P}\left\{Y = y_i | w^\star, x_i\right\} < \mathbb{P}\left\{Y = y_i | w^\star\left(y_i^\dagger\right), x_i\right\}.$$

A learner (algorithm) is incentive compatible if there does not exist a dataset that is incentive incompatible.

Proposition 1. *Multi-class logistic regression is not incentive compatible.*

Proof. The example described in Figure 1 is a dataset that is incentive incompatible.
In this example, agent i reports $x_i \in \mathbb{R}^2$ and y_i is one of 1 (red), 2 (green), or 3 (blue). Suppose the red square point correspond to agent 1 with $x_1 = (-1.63, -1.17)$ and $y_1 = 1$.

$$\mathbb{P}\left\{Y = 1 | w^\star, x_1\right\} = 0.3290,$$
$$\mathbb{P}\left\{Y = 1 | w^\star \left(y_1^\dagger = 3\right), x_1\right\} = 0.4966.$$

Here, parameter estimation is done using maximum likelihood estimation with BFGS, and $w^\star$ is given by, with class 1 weights normalized to 0,

Class	(Intercept)	x1	x2
2	-0.6053178	104.9925	-181.3391914
3	-0.2852057	209.4190	0.3656777

and $w^\star \left(y_1^\dagger = 3\right)$ is given by,

Class	(Intercept)	x1	x2
2	-0.1915645	3.473426	-5.507418
3	0.8273350	4.309293	-1.200060

2.2 Zero-One Loss Logistic Regression

It is, however, possible to change the loss function so that logistic regression is incentive compatible. Changing the loss function to absolute value L^1 loss is one possibility, due to **?**. Their result on incentive compatibility of empirical risk minimization in the regression setting is applicable in our model. In addition to absolute value loss, which is not a meaningful loss function for multi-class logistic regression, zero-one loss logistic regression with deterministic predictions is also incentive compatible.

□

Proposition 2. *Multi-class deterministic classifiers estimated by empirical risk minimization with zero-one loss is incentive compatible.*

Proof. For any dataset $\{(x_i, y_i)\}_{i=1}^n$, and the hypothesis class $\mathcal{H}$, let the optimal classifier be,

$$h^\star = \arg\min_{h \in \mathcal{H}} \sum_{i'=1}^{n} \mathbb{1}_{\{y_{i'} \neq h(x_{i'})\}}.$$

Fix an agent i, her feature vector x_i, and fix other agents' reports, (x_{-i}, y_{-i}), define the loss function given the classifier h and report of agent i, $y_i^\dagger$ as,

$$\mathcal{L}\left(h, y_i^\dagger\right) = \sum_{i' \neq i} \mathbb{1}_{\{y_{i'} \neq h(x_{i'})\}} + \mathbb{1}_{\left\{y_i^\dagger \neq h(x_i)\right\}}.$$

If $y_i = h^\star\left(x_i\right)$, then the classifier is already classifying x_i correctly, misreporting will not improve the outcome for i. Now let the prediction be $h^\star\left(x_i\right) = y^\star \neq y_i$, and suppose $h^\star$ is making k mistakes, meaning,

$$k = \min_{h \in \mathcal{H}} \mathcal{L}\left(h^\star, y_i\right).$$

Agent i can misreport in the following two ways: □

1. If agent i reports $y_i^\dagger = y^\star$, let the new classifier be $h^\dagger$, note that we must have,

$$\mathcal{L}\left(h^\dagger, y^\star\right) \leqslant k - 1,$$

because $\mathcal{L}\left(h^{\dagger}, y^{\star}\right) > k-1 = \mathcal{L}\left(h^{\star}, y^{\star}\right)$ contradicts the optimality of $h^{\dagger}$.
Now suppose that agent i could get her true label with $h^{\dagger}$, meaning $h^{\dagger}\left(x_i\right) = y_i$, then,

$$\begin{aligned}\mathcal{L}\left(h^{\dagger}, y_i\right) &= \mathcal{L}\left(h^{\dagger}, y^{\star}\right) - 1 \\ &\leqslant k-2 \\ &< \mathcal{L}\left(h^{\star}, y_i\right),\end{aligned}$$

which contradicts the optimality of $h^{\star}$. Therefore, agent i cannot improve the outcome by misreporting $y^{\star}$.

2. If agent i reports $y_i^{\dagger} = y' \neq y^{\star}$, let the new classifier be $h^{\dagger}$, note that we must have,

$$\mathcal{L}\left(h^{\dagger}, y'\right) \leqslant k,$$

because if $\mathcal{L}\left(h^{\dagger}, y'\right) > k = \mathcal{L}\left(h^{\star}, y'\right)$ contradicts the optimality of $h^{\dagger}$.
Now suppose that agent i could get her true label with $h^{\dagger}$, then,

$$\begin{aligned}\mathcal{L}\left(h^{\dagger}, y_i\right) &= \mathcal{L}\left(h^{\dagger}, y_i\right) - 1 \\ &\leqslant k-1 \\ &< \mathcal{L}\left(h^{\star}, y_i\right),\end{aligned}$$

which contradicts the optimality of $h^{\star}$. Therefore, agent i cannot improve the outcome by misreporting y'.
Therefore, no agent can improve the outcome and the dataset is incentive compatible.

3 Bayes Classifier

The example given previously is incentive compatible with respect to the Naive Bayes classifier. None of the agents have the incentive to misreport their labels. This is always true in general for any parametric Bayesian classifier estimated using maximum likelihood.

Proposition 3. *Bayesian classifiers are incentive compatible.*

Proof. Suppose the loglikelihood function of class y given the feature vector x and the parameter w is $\ell\left(x; w\right)$, and define the optimal parameter, $w^{\star}$ for class y_i, of the truthful model as,

$$\begin{aligned}w^{\star} &= \arg\max_{w} \sum_{i':y_{i'}=y_i} \ell\left(x_{i'}; w\right) \\ &= \arg\max_{w} \sum_{i'\neq i:y_{i'}=y_i} \ell\left(x_{i'}; w\right) + \ell\left(x_i; w\right).\end{aligned}$$

Let the optimal parameter when agent i reports $y_i^{\dagger} \neq y_i$ be $w^{\dagger,}$

$$w^{\dagger} = \arg\max_{w} \sum_{i'\neq i:y_{i'}=y_i} \ell\left(x_{i'}; w\right).$$

In particular, these implies the following optimality conditions,

$$\begin{aligned}\sum_{i'\neq i:y_{i'}=y_i} \ell\left(x_{i'}; w^{\star}\right) + \ell\left(x_i; w^{\star}\right) &\geqslant \sum_{i'\neq i:y_{i'}=y_i} \ell\left(x_{i'}; w^{\dagger}\right) + \ell\left(x_i; w^{\dagger}\right), \\ \sum_{i'\neq i:y_{i'}=y_i} \ell\left(x_{i'}; w^{\dagger}\right) &\geqslant \sum_{i'\neq i:y_{i'}=y_i} \ell\left(x_{i'}, w^{\star}\right).\end{aligned}$$

Taking the difference between the two inequalities, we have,

$$\ell\left(x_i; w^{\star}\right) \geqslant \ell\left(x_i; w^{\dagger}\right).$$

Note that the empirical prior probability for class y_i is decreased if the number of data with label y_i is decreased by 1. Therefore, the posterior probabilities satisfy,

$$\mathbb{P}\left\{y_i|x_i, w^\star\right\} \geqslant \mathbb{P}\left\{y_i|x_i, w^\dagger\right\}.$$

Therefore, no agent can improve the outcome and the dataset is incentive compatible.

□