Strategy Free Machine Learning

-

April 10, 2020

1 Short Summary

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".

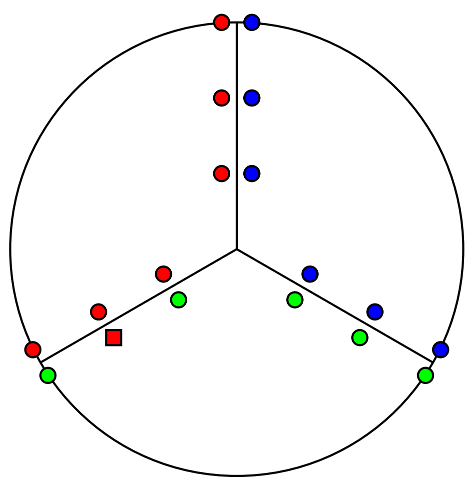

Figure 1: Incentive Incompatible Example

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.

The same dataset is also incentive incompatible with respect to the one-vs-rest linear support vector machine if the margin is used as the class "probabilities". However, in this case, the agent, with feature vector corresponding to the blue point close to the center and close to the green point, can only improve the margin slightly without making the model switch from classifying her point incorrectly to classifying her point correctly.

However, this dataset is incentive compatible with respect to the Naive Bayes classifiers, and in general, there does not exist any dataset that is incentive incompatible with respect to discrete-valued Naive Bayes classifiers. Misreporting will always lead to a lower posterior probability of the agent's true label. In addition, no dataset is incentive incompatible with respect to classifiers that minimize empirical risk with zero-one loss.

2 Literature Review

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. Perote and Perote-Pena (2004), Chen, Podimata, Procaccia, and Shah (2018), and Gast, Ioannidis, Loiseau, and Roussillon (2013) focus on strategy-proof linear regression algorithms and introduced clockwise repeated median estimators, generalized resistant hyperplane estimators, and modified generalized linear squares estimators. Dekel, Fischer, and Procaccia (2010) 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 Meir, Procaccia, and Rosenschein (2012) 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. Chen, Liu, and Podimata (2019) 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 Dütting, Feng, Narasimhan,

Parkes, and Ravindranath (2017), Golowich, Narasimhan, and Parkes (2018), Epasto, Mahdian, Mirrokni, and Zuo (2018), and Procaccia and Tennenholtz (2009).

- 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. Richardson, Rokvic, Filos-Ratsikas, and Faltings (2019) focuses on the linear regression problem in this scenario, and Cai, Daskalakis, and Papadimitriou (2015) and Shah and Zhou (2016) investigates the problem for more general machine learning problems. Mihailescu and Teo (2010) 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 Dekel and Shamir (2009b), Dekel and Shamir (2009a).

3 Logistic Regression

3.1 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,

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

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

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,

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

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

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 given previously is a dataset that is incentive incompatible.

□

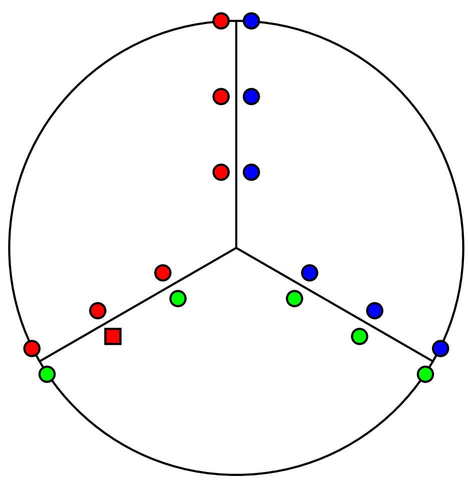

Figure 2: Incentive Incompatible Example

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

3.2 Incentive Incompatibility

To characterize the set of incentive incompetible datasets, we rewrite the principal's choice of optimal weights by,

$$w^{\star} = \arg\max_{w} \log\left(\mathbb{P}\left\{Y = y_i | w, x_i\right\}\right) + C_{-i}\left(w\right),$$

where the function $C_{-i}\left(w\right)$ summarizes the loss due to agents other than i, assuming they are reporting labels truthfully,

$$C_{-i}\left(w\right) = \sum_{i'=0, i' \neq i}^{n} \log\left(\mathbb{P}\left\{Y = y_{i'} | w, x_{i'}\right\}\right).$$

Since the objective is globally convex and differentiable, as shown in , the problem translates to the first derivative condition,

$$\frac{\nabla_w \mathbb{P}\left\{Y = y_i | w^{\star}, x_i\right\}}{\mathbb{P}\left\{Y = y_i | w^{\star}, x_i\right\}} + \nabla_w\left(C_{-i}\left(w^{\star}\right)\right) = 0,$$
$$\frac{\nabla_w \mathbb{P}\left\{Y = y_i^{\dagger} | w^{\star}\left(y_i^{\dagger}\right), x_i\right\}}{\mathbb{P}\left\{Y = y_i^{\dagger} | w^{\star}\left(y_i^{\dagger}\right), x_i\right\}} + \nabla_w\left(C_{-i}\left(w^{\star}\left(y_i^{\dagger}\right)\right)\right) = 0.$$

For logistic regression with weights $w_c, c = 1, 2, ..., k$, without normalization,

$$\mathbb{P}\left\{Y = c | w, x\right\} = \frac{e^{w_c^T x + b_c}}{\sum_{c'} e^{w_{c'}^T x + b_{c'}}},$$
$$\nabla_{w_c} \mathbb{P}\left\{Y = c | w, x\right\} = \frac{e^{w_c^T x + b_c} \sum_{c' \neq c} e^{w_c^T x + b_c}}{\left(\sum_{c'} e^{w_{c'}^T x + b_{c'}}\right)^2} x.$$
$$\nabla_{w_c} \mathbb{P}\left\{Y = \hat{c}, \hat{c} \neq c | w, x\right\} = \frac{e^{w_c^T x + b_c} e^{w_{\hat{c}}^T x + b_{\hat{c}}}}{\left(\sum_{c'} e^{w_{c'}^T x + b_{c'}}\right)^2} x.$$

The derivative conditions implies,

$$\left(1 - \mathbb{P}\left\{Y = c | w^{\star}, x_i\right\}\right) x_i + \nabla_{w_c}\left(C_{-i}\left(w^{\star}\right)\right) = 0, c = y_i,$$
$$\left(\mathbb{P}\left\{Y = c | w^{\star}, x_i\right\}\right) x_i + \nabla_{w_c}\left(C_{-i}\left(w^{\star}\right)\right) = 0, c \neq y_i,$$

same for the expression with $w^\star\left(y_i^\dagger\right)$.
Substitute into the incentive incompatibility condition,

$$\nabla_{w_{y_i,j}}\left(C_{-i}\left(w^\star\right)\right)x_{i,j} \leqslant \nabla_{w_{y_i,j}}\left(C_{-i}\left(w^\star\left(y_i^\dagger\right)\right)\right)x_{i,j}, j = 1, 2, ..., m.$$

3.3 Continuous Label

The previous formulation does not permit $y_i^\dagger$ to be a continuous variable, but if we rewrite the optimization as the maximization of the cross-entropy, then we could treat $y_i^\dagger \in \Delta^{K-1}$ as a continuous multinomial distribution where $y_{i,c}^\dagger \in [0, 1]$ denotes the probability of agent i reporting label $c \in \{1, 2, ..., K\}$. The principal's problem can be rewritten as,

$$\min_w \sum_{i=1}^n \sum_{c=1}^K -y_{i,c}^\dagger \log\left(\mathbb{P}\left\{Y = c|w, x_i\right\}\right).$$

Assuming $w^\star\left(y_i^\dagger\right)$ is the optimal weights, the objective function becomes,

$$\mathcal{L}\left(w, y_i^\dagger\right) = \sum_{c=1}^k -\hat{y}_{i,c}\log\left(\mathbb{P}\left\{Y = c|w, x_i\right\}\right) - \sum_{i'=1, i'\neq i}^n \log\left(\mathbb{P}\left\{Y = y_{i'}|w, x_{i'}\right\}\right),$$

and the value function is,

$$\mathcal{L}^\star\left(y_i^\dagger\right) = \sum_{c=1}^k -\hat{y}_{i,c}\log\left(\mathbb{P}\left\{Y = c|w^\star\left(y_i^\dagger\right), x_i\right\}\right) - \sum_{i'=1, i'\neq i}^n \log\left(\mathbb{P}\left\{Y = y_{i'}|w^\star\left(y_i^\dagger\right), x_{i'}\right\}\right),$$

and apply the envolope theorem,

$$\begin{aligned}\frac{\partial \mathcal{L}^\star\left(y_i^\dagger\right)}{\partial y_i^\dagger} &= -\log\left(\mathbb{P}\left\{Y = y_i^\dagger|w^\star\left(y_i^\dagger\right), x_i\right\}\right) \\ &> 0.\end{aligned}$$

Alternatively, if gradient descent is used in the optimization process, one iteration of the gradient descent with learning rate η is given by,

$$w_{j,c}' = w_{j,c} - \eta x_{i,j}\left(\mathbb{P}\left\{Y = c|x_i\right\} - \mathbb{1}_{y_i^\dagger}\right).$$

Now fix instance i and define $o_c = \mathbb{P}\left\{Y = c|x_i\right\}$, then,

$$\begin{aligned}\frac{\partial o_c}{\partial y_c} &= \frac{\partial o_c}{\partial z_c}\sum_{j=1}^m \frac{\partial z_c}{\partial w_{j,c}}\frac{\partial w_{j,c}}{\partial y_c} \\ &= o_c\left(1 - o_c\right)\sum_{j=1}^m x_j^{(i)} x_j^{(i)}\eta \\ &= \eta o_c\left(1 - o_c\right)\sum_{j=1}^m \left(x_j^{(i)}\right)^2 \\ &\geqslant 0.\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{\partial o_c}{\partial y_{c'}} &= \frac{\partial o_c}{\partial z_{c'}} \sum_{j=1}^{m} \frac{\partial z_{c'}}{\partial w_{j,c'}} \frac{\partial w_{j,c'}}{\partial y_{c'}} \\ &= -o_c o_{c'} \sum_{j=1}^{m} x_j^{(i)} x_j^{(i)} \eta \\ &= -\eta o_c o_{c'} \sum_{j=1}^{m} \left(x_j^{(i)}\right)^2 \\ &\leqslant 0.\end{aligned}$$

This implies decreasing $y_{i,c}$ and thus increasing $y_{i,c'}$ for some c' will always increase o_c. Therefore, there should be no incentive to misreport by changing $y_i^{\dagger}$ slightly from $y_{i.}$

3.4 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 Dekel, Fischer, and Procaccia (2010). 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 is also incentive compatible.

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

Proof. For any dataset $\{(x_i, y_i)\}_{i=1}^n$, suppose the model is parameterized by w,

$$w^\star = \arg\min_w \sum_{i=1}^{n} \mathbb{1}_{\left\{y_i \neq \arg\max_c \mathbb{P}\{Y=c|w,x_i\}\right\}}.$$

If $y_i = \arg\max_c \mathbb{P}\{Y = c|w, x_i\}$, then the classifier is already classifying x_i correctly. Now suppose,

$$y^\star = \arg\max_c \mathbb{P}\{Y = c|w, x_i\}, y^\star \neq y_{i.}$$

If agent i reports $y^\star$, then the total loss is decreased by 1. It is impossible to achieve a lower loss because the converse would contradict the optimality of $w^\star$. Therefore $w^\star\left(y_i^\dagger = y^\star\right) = w^\star$ and the classification will not change.

If agent i reports $y' \neq y^\star$, then the total loss stays the same. The only way to possibly achieve a lower loss is having $w^\star\left(y_i^\dagger = y'\right)$ classifies x_i as y'. Therefore, agent i is not better off.

$\square$

4 Naive Bayes Model

The example given previously is incentive compatible with respect to Gaussian Naive Bayes. None of the agents have the incentive to misreport their labels. This is always true in general for multinomial Naive Bayes.

Proposition 3. *Multinomial Naive Bayes classifier is incentive compatible.*

Proof. For any dataset $\{(x_i, y_i)\}_{i=1}^{n}$, denote the total count of feature j with label c as $n_{j,c}$, and the number of points with label c as n_c.

$$n_{j,c} = \sum_{i=1}^{n} x_{i,j} \mathbb{1}_{\{y_i=c\}}),$$
$$n_c = \sum_{i=1}^{n} \mathbb{1}_{\{y_i=c\}}.$$

The probability that x_i is classified as y_i, if agent i reports truthfully is,

$$\mathbb{P}\{Y = y_i | w^\star, x_i\} = \frac{\mathbb{P}\{Y = y_i, w^\star\} \mathbb{P}\{x_i | y_i, w^\star\}}{\mathbb{P}\{x_i, w^\star\}} = \frac{\frac{n_{y_i}}{n} \prod_{j=1}^{m} \left(\frac{n_{j,y_i}}{n_{y_i}}\right)^{x_i}}{C},$$

where C is a scaling constant.
If agent i misreports $y_i^\dagger \neq y_i$,

$$\mathbb{P}\left\{Y = y_i | w^\star\left(y_i^\dagger\right), x_i\right\} = \frac{\mathbb{P}\left\{Y = y_i, w^\star\left(y_i^\dagger\right)\right\} \mathbb{P}\left\{x_i | y_i, w^\star\left(y_i^\dagger\right)\right\}}{\mathbb{P}\left\{x_i, w^\star\left(y_i^\dagger\right)\right\}} = \frac{\frac{n_{y_i} - 1}{n} \prod_{j=1}^{m} \left(\frac{n_{j,y_i} - 1}{n_{y_i} - 1}\right)^{x_i}}{C},$$

note that the constants are identical, so we have,

$$\frac{n_{y_i} - 1}{n} < \frac{n_{y_i}}{n},$$
$$\frac{n_{j,y_i} - 1}{n_{y_i} - 1} < \frac{n_{j,y_i}}{n_{y_i}}, \text{ for each } j \in \{1, 2, ..., m\},$$

which implies,

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

Therefore, the classifier is not incentive incompatible on this dataset.

□

On the other hand, the parameters of the Gaussian Naive Bayes $w = (\mu, \Sigma, \pi)$ (mean, variance, prior) are estimated by,

$$\begin{aligned}
\mu_c^\star &= \frac{\sum_{i'=1}^{n} x_{i'} \mathbb{1}_{y_{i'}=c}}{n_c}, \\
\Sigma_c^\star &= \frac{\sum_{i'=1}^{n} \left(x_{i'} - \mu_c^\star\right) \left(x_{i'} - \mu_c^\star\right)^T \mathbb{1}_{y_{i'}=c}}{n_c}, \\
\pi_c^\star &= \frac{n_c}{n}.
\end{aligned}$$

If agent i reports truthfully, the classification probability is,

$$\mathbb{P}\left\{Y = y_i | w^\star, x_i\right\} \propto \pi_{y_i}^\star \frac{1}{\left|\Sigma_{y_i}^\star\right|} \exp\left(-\frac{1}{2}\left(x_i - \mu_{y_i}^\star\right)^T \left(\Sigma_{y_i}^\star\right)^{-1} \left(x_i - \mu_{y_i}^\star\right)\right).$$

If agent i misreports $y_i^\dagger \neq y_i$,

$$\begin{aligned}
\mu_c^\star\left(y_i^\dagger\right) &= \frac{\sum_{i' \neq i} x_{i'} \mathbb{1}_{y_{i'}=c}}{n_c - 1}, \\
\Sigma_c^\star\left(y_i^\dagger\right) &= \frac{\sum_{i' \neq i} \left(x_{i'} - \mu_c^\star\left(y_i^\dagger\right)\right) \left(x_{i'} - \mu_c^\star\left(y_i^\dagger\right)\right)^T \mathbb{1}_{y_{i'}=c}}{n_c - 1}, \\
\pi_c^\star\left(y_i^\dagger\right) &= \frac{n_c - 1}{n}.
\end{aligned}$$

The classification with the parameters $w^\star\left(y_i^\dagger\right)$ is,

$$\mathbb{P}\left\{Y = y_i | w^\star\left(y_i^\dagger\right), x_i\right\} \propto \pi_{y_i}^\star\left(y_i^\dagger\right) \frac{1}{\left|\Sigma_{y_i}^\star\left(y_i^\dagger\right)\right|} \exp\left(-\frac{1}{2}\left(x_i - \mu_{y_i}^\star\left(y_i^\dagger\right)\right)^T \left(\Sigma_{y_i}^\star\left(y_i^\dagger\right)\right)^{-1} \left(x_i - \mu_{y_i}^\star\left(y_i^\dagger\right)\right)\right).$$

To make the comparison, consider the one-dimensional case. The incentive compatibility condition is,

$$\pi_{y_i}^\star \frac{1}{\sigma_{y_i}^\star \sqrt{2\pi}} \exp\left(-\frac{\left(x_i - \mu_{y_i}^\star\right)^2}{2\left(\sigma_{y_i}^\star\right)^2}\right) \geqslant \pi_{y_i}^\star\left(y_i^\dagger\right) \frac{1}{\sigma_{y_i}^\star\left(y_i^\dagger\right)\sqrt{2\pi}} \exp\left(-\frac{\left(x_i - \mu_{y_i}^\star\left(y_i^\dagger\right)\right)^2}{2\left(\sigma_{y_i}^\star\right)^2\left(y_i^\dagger\right)}\right),$$

where,

$$\begin{aligned}
\pi_c^\star\left(y_i^\dagger\right) &= \frac{n_c - 1}{n_c} \pi_c^\star, \\
\mu_c^\star\left(y_i^\dagger\right) &= \frac{n_c \mu_c^\star}{n_c - 1} - \frac{x_i}{n_c - 1}, \\
\left(\sigma_c^\star\right)^2\left(y_i^\dagger\right) &= \frac{n_c\left(\sigma_c^\star\right)^2}{n_c - 1} - \frac{\left(x_i - \mu_c^\star\right)\left(x_i - \mu_c^\star\right)}{n_c - 1}
\end{aligned}$$

$$= \frac{n_c \left(\sigma_c^\star\right)^2}{n_c - 1} - \frac{n_c \left(x_i - \mu_c^\star\right)^2}{\left(n_c - 1\right)^2}.$$

The comparison is not clear as scenarios like the one in the diagram below could be possible.

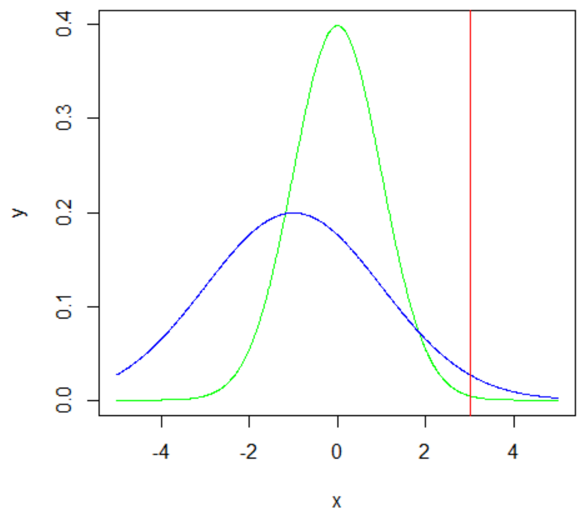

Figure 3: Normal Means

-

5 Support Vector Machines

5.1 One-vs-One

Since binary SVM is incentive compatible, no agent can gain from misreporting in any of the one-vs-one SVMs. Therefore, there will be no incentive to misreport in the multi-class SVM.

5.2 One-vs-Rest

If margin is used as the prediction probabilities, then it is possible to improve the margin by misreporting the third class label, for example on the 18-point data set.

5.3 Tree-Based

Since binary SVM is incentive compatible, no agent can gain from misreporting in any stage. Therefore, there will be no incentive to misreport in the multi-class SVM.

References

Cai, Yang, Constantinos Daskalakis, and Christos Papadimitriou (2015), "Optimum statistical estimation with strategic data sources." In *Conference on Learning Theory*, 280–296.

Chen, Yiling, Yang Liu, and Chara Podimata (2019), "Grinding the space: Learning to classify against strategic agents." *arXiv preprint arXiv:1911.04004*.

Chen, Yiling, Chara Podimata, Ariel D Procaccia, and Nisarg Shah (2018), "Strategyproof linear regression in high dimensions." In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 9–26.

Dekel, Ofer, Felix Fischer, and Ariel D Procaccia (2010), "Incentive compatible regression learning." *Journal of Computer and System Sciences*, 76, 759–777.

Dekel, Ofer and Ohad Shamir (2009a), "Good learners for evil teachers." In *Proceedings of the 26th annual international conference on machine learning*, 233–240.

Dekel, Ofer and Ohad Shamir (2009b), "Vox populi: Collecting high-quality labels from a crowd." In *COLT*.

Dütting, Paul, Zhe Feng, Harikrishna Narasimhan, David C Parkes, and Sai Srivatsa Ravindranath (2017), "Optimal auctions through deep learning." *arXiv preprint arXiv:1706.03459*.

Epasto, Alessandro, Mohammad Mahdian, Vahab Mirrokni, and Song Zuo (2018), "Incentive-aware learning for large markets." In *Proceedings of the 2018 World Wide Web Conference*, 1369–1378.

Gast, Nicolas, Stratis Ioannidis, Patrick Loiseau, and Benjamin Roussillon (2013), "Linear regression from strategic data sources." *arXiv preprint arXiv:1309.7824*.

Golowich, Noah, Harikrishna Narasimhan, and David C Parkes (2018), "Deep learning for multi-facility location mechanism design." In *IJCAI*, 261–267.

Meir, Reshef, Ariel D Procaccia, and Jeffrey S Rosenschein (2012), "Algorithms for strategyproof classification." *Artificial Intelligence*, 186, 123–156.

Mihailescu, Marian and Yong Meng Teo (2010), "Strategy-proof dynamic resource pricing of multiple resource types on federated clouds." In *International Conference on Algorithms and Architectures for Parallel Processing*, 337–350, Springer.

Perote, Javier and Juan Perote-Pena (2004), "Strategy-proof estimators for simple regression." *Mathematical Social Sciences*, 47, 153–176.

Procaccia, Ariel D and Moshe Tennenholtz (2009), "Approximate mechanism design without money." In *Proceedings of the 10th ACM conference on Electronic commerce*, 177–186.

Richardson, Adam, Ljubomir Rokvic, Aris Filos-Ratsikas, and Boi Faltings (2019), "Privately computing influence in regression models."

Shah, Nihar B and Dengyong Zhou (2016), "Double or nothing: Multiplicative incentive mechanisms for crowdsourcing." *The Journal of Machine Learning Research*, 17, 5725–5776.