Incentive Compatibility of Multi-Class Logistic Regression

Young Wu

April 27, 2021

Model Setup

- There are n agents. Agent i has public x_i and private y_i.
- The principal uses a publicly known classifier.
- Agent i chooses to report some $y_i^\dagger$ to maximize the probability that the classifier classifies x_i correctly.
- Question: for which classifiers, the agents do not have incentive to misreport their y_i

Incentive Compatibility

- A dataset is incentive compatible (IC) given a classifier if there is an i and some $y_i^\dagger \neq y_i$ such that,

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

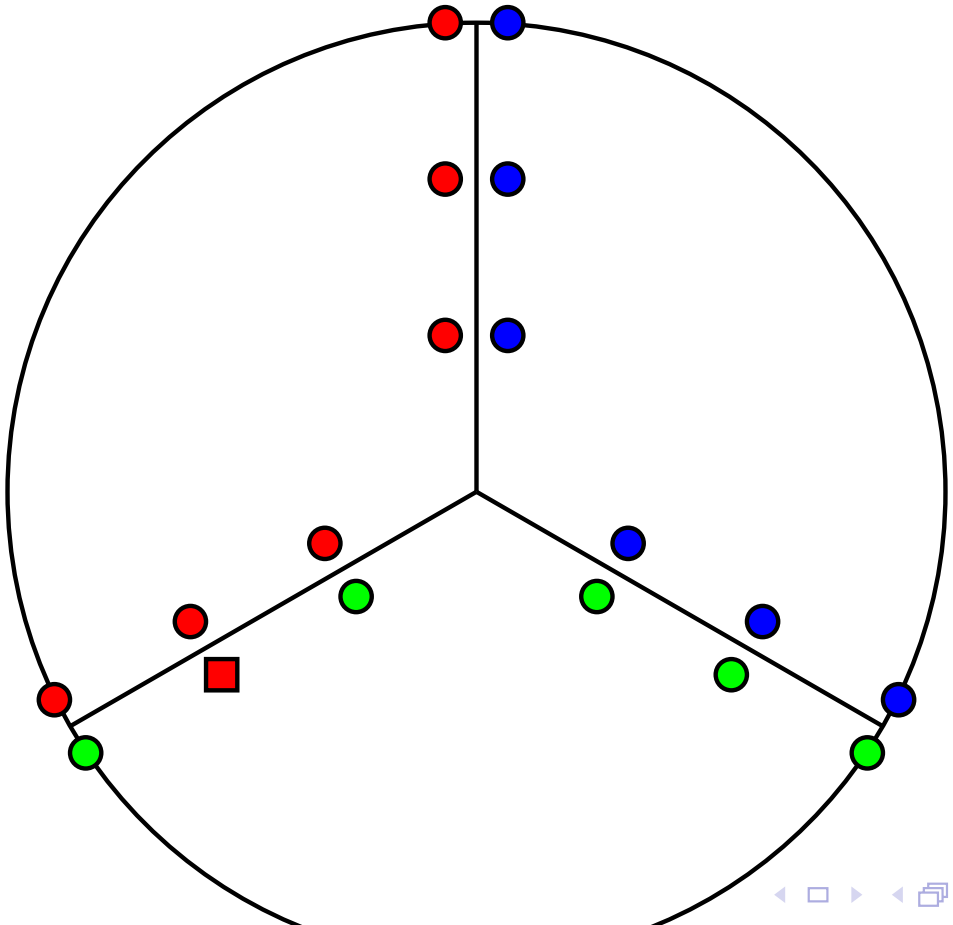
- A classifier is incentive compatible if all datasets are incentive compatible given this classifier.

Logistic Regression Example

- One numerical example dataset is not incentive compatible,

$$\mathbb{P}\left\{Y = \text{ red } \middle| w^\star, x_{\text{square}}\right\} = 0.3290, \text{ but}$$

$$\mathbb{P}\left\{Y = \text{ red } \middle| w^\star\left(y^\dagger_{\text{square}} = \text{ blue }\right), x_{\text{square}}\right\} = 0.4966.$$



Logistic Regression Experiments

- R nnet package is used to estimate the logistic regression coefficients on 10000 datasets of 100 iid standard normally distributed points. If the labels are assigned at random, none of the datasets are incentive incompatible. If the random weights are used to create the labels (but with small error), then among the 10000 dataset:

1. 99952 are IC.
2. 48 are not IC.
3. 7 are not IC for two or more agents.
4. for 4 datasets from the 48, after one agent misreports, at least one other agent has the incentive to also misreport as a "defense".

IC ERM Classifiers

- Binary classifiers (Proposition 2, requires a technical condition on a loss function. The condition is satisfied for the standard logistic regression)
- ERM with zero-one loss (Proposition 3)
- ERM with L_1 loss (Dekel, Fischer, Procaccia 2010)

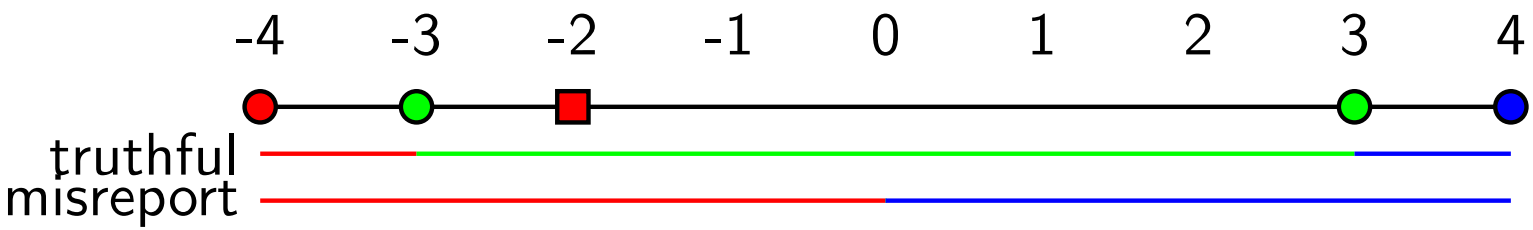
IC Probabilistic Classifiers

- Binary classifiers (Corollary 1, does not require additional distributional assumptions)
- Bayes classifiers (Proposition 4)
- Gaussian mixture model (A special case of Bayes classifier)
- Kernel density estimators (Corollary 2)

Artificial Example

- Three-class threshold classifier with thresholds symmetric around 0 and squared error margin,

$$\mathbb{P}\left\{Y = \text{ red } \middle| w^\star, x_{\text{square}}\right\} = 0, \text{ but}$$
$$\mathbb{P}\left\{Y = \text{ red } \middle| w^\star\left(y^\dagger_{\text{square}} = \text{ blue }\right), x_{\text{square}}\right\} = 1.$$



Intuition

- An agent in class a is mis-classified as b may want to misreport as c to move the decision boundary between a and b.
- Two conditions to prevent this incentive incompatibility problem:

1. Adding one class a agent at x moves the decision boundary between a and b away from x (so that x is relatively more likely to be classified as a).
2. Adding one class c agent at x does not change the decision boundary between a and b.

Formal Conditions

- Monotonic condition (MC) and Independence of Irrelevant Alternatives condition (IIA), for any x and a, b,

1. $$\frac{\mathbb{P}\{Y = a | x, w^\star(S)\}}{\mathbb{P}\{Y = b | x, w^\star(S)\}} \leqslant \frac{\mathbb{P}\{Y = a | x, w^\star(S \cup \{(x, y = a)\})\}}{\mathbb{P}\{Y = b | x, w^\star(S \cup \{(x, y = a)\})\}}.$$
2. $$\frac{\mathbb{P}\{Y = a | x, w^\star(S)\}}{\mathbb{P}\{Y = b | x, w^\star(S)\}} = \frac{\mathbb{P}\{Y = a | x, w^\star(S \cup \{(x', y' \notin \{a, b\})\})\}}{\mathbb{P}\{Y = b | x, w^\star(S \cup \{(x', y' \notin \{a, b\})\})\}}.$$

Outline of Proof

- Let $w^\star$ be the estimated parameters when everyone reports truthfully and $w^\dagger$ be the estimated parameters when some agent i misreports $y_i^\dagger \neq y_i$.
- Suppose for a contradiction that $\mathbb{P}\left\{Y = y_i | x_i, w^\star\right\} < \mathbb{P}\left\{Y = y_i | x_i, w^\dagger\right\}$.
- MC and IIA implies $\mathbb{P}\left\{Y_i = y_i^\dagger | x_i, w^\star\right\} > \mathbb{P}\left\{Y = y_i | x_i, w^\dagger\right\}$.
- Then, the optimality of $w^\star$ contradicts with the optimality of $w^\dagger$.

ERM Classifiers

- MC and IIA may not be sufficient.
- A normalization condition is required: the sum of the losses over all classes for any point must be constant.
- ERM with zero-one loss satisfies MC and the normalization condition, but does not always satisfy the IIA condition.

Separable Classifiers

- If separate parameters are estimated for separate classes (independent of the points from other classes), and the classification probabilities can be written as a fraction of the values from different classes, then MC and IIA are always satisfied, therefore always IC.
- Examples include Bayes classifiers and kernel density estimators (Taylor 1997 form).

Separable Logistic Regression

- Logistic Regression is not separable.
- One-vs-one and One-vs-all Logistic Regressions are also not separable, but they are IC (because the binary classifiers are IC).