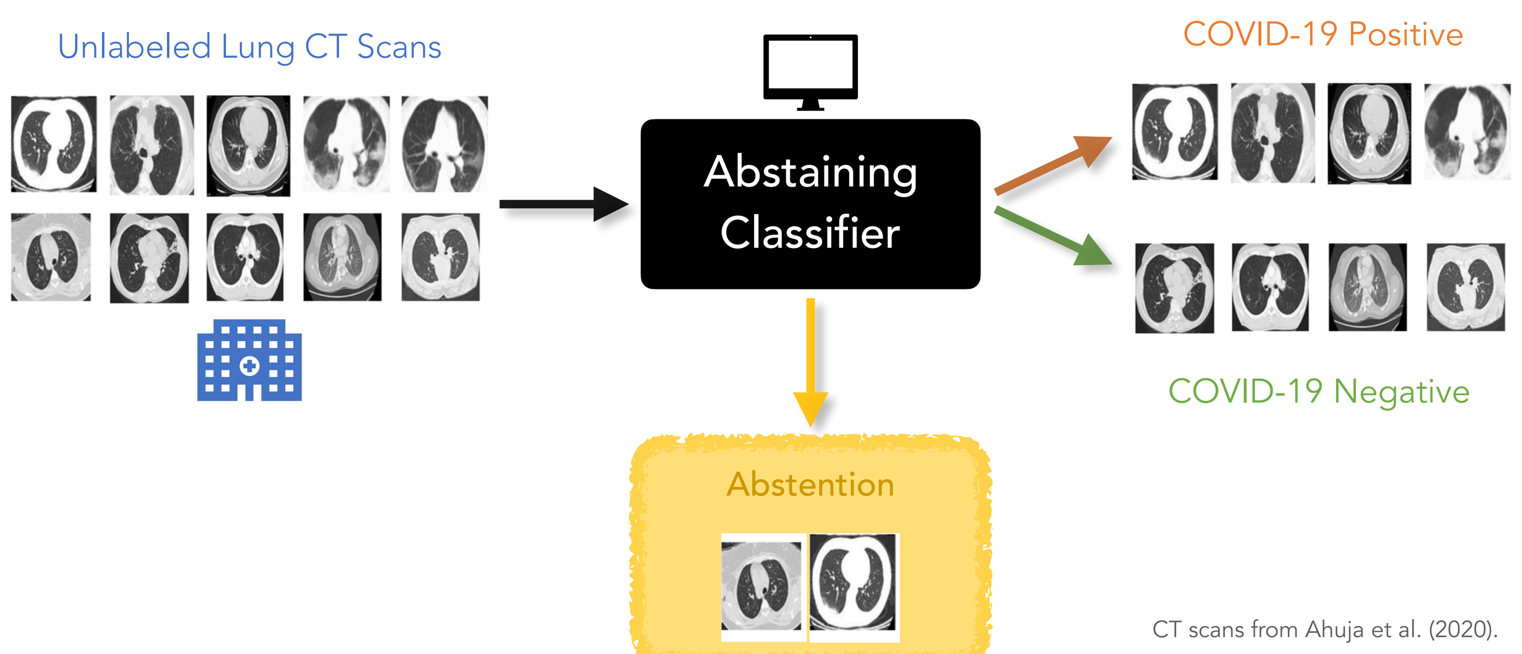
Counterfactually Comparing Abstaining Classifiers

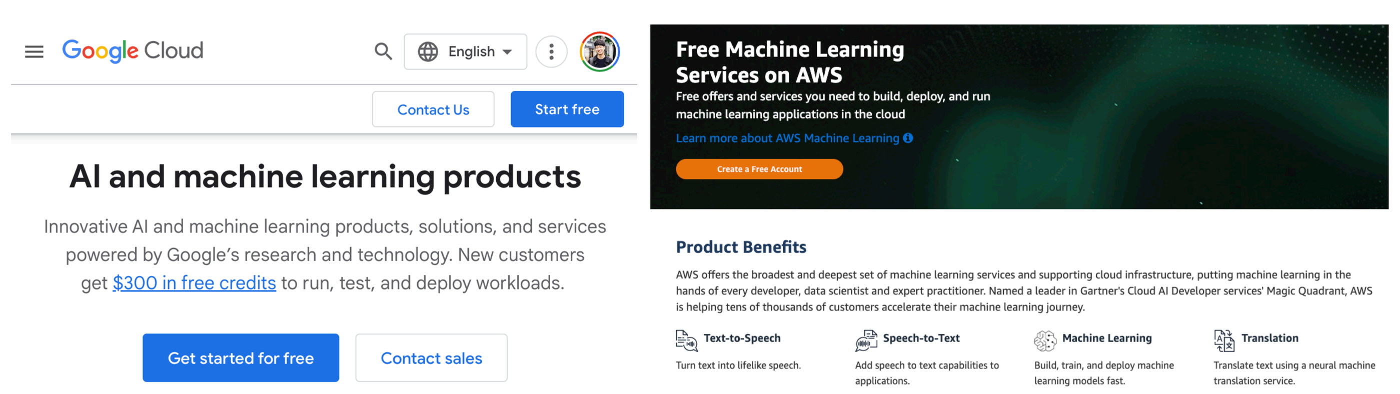
Yo Joong "YJ" Choe (UChicago), Aditya Gangrade (UMichigan), Aaditya Ramdas (Carnegie Mellon University)

Abstaining classifiers

Abstaining classifiers (Chow, 1957) have the option to withhold their predictions on inputs that they are uncertain about. They are used in safety-critical applications, such as medical imaging.



How can we evaluate and compare black-box abstaining classifiers?



- Suppose we want to evaluate and compare black-box ML prediction services for an image classification task.
- **During the free trial, each service deploys an abstaining classifier.** Each classifier utilizes its own (unknown) abstention mechanism.
- Once you pay for each service, it will use a non-abstaining classifier. *How can we compare the expected accuracies without accessing them?*

To the evaluator, abstentions are just missing predictions!

The counterfactual question

How would we compare black-box abstaining classifiers, had they not been allowed to abstain?

We propose a **black-box** evaluation framework for abstaining classifiers by leveraging tools from *missing data analysis* (Rubin, 1976) and *nonparametric causal inference* (e.g., Robins et al., 1994).

The counterfactual approach

An *abstaining classifier (AC)* is a pair of functions (f, π), where

- $f : \mathcal{X} \to \mathcal{Y}$ is the base classifier ($f(X)$: prediction);
- $\pi : \mathcal{X} \to [0, 1]$ is the abstention mechanism ($\pi(X)$: prob. of abstention).

<u>Protocol (Evaluating a black-box abstaining classifier).</u>

1. Classifier receives an input X.
2. Classifier decides whether or not it will abstain: $R \mid X \sim \text{Ber}(\pi(X))$.
   - If $R = 0$, then Evaluator sees its prediction & score: $S = s(f(X), Y)$.
   - If $R = 1$, then Evaluator does NOT see its score (**S is missing**).

Step 1: Defining the counterfactual score

The *counterfactual score* ψ of an AC (f, π) is its expected score:

$$\psi \overset{\text{def}}{=} \mathbb{E}[S] .$$

For comparison, estimate $\Delta = \psi^A - \psi^B = \mathbb{E}[S^A - S^B]$.

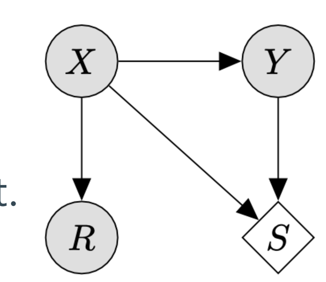
Step 2: Identification

Under identifying conditions,

$$\psi = \mathbb{E}[\mu_0(X)], \quad \text{where} \quad \mu_0(X) = \mathbb{E}[S \mid X, R = 0] .$$

<u>*What are the identifying conditions?*</u>

1. **Missing at random (MAR):** $S \perp\!\!\!\perp R \mid X$.
   - Satisfied as long as the evaluation set is *independent* of training set.
2. **Positivity:** There exists $\varepsilon > 0$ such that $\pi(X) \leq 1 - \varepsilon$.
   - Satisfied as long as the classifier does not *deterministically* abstain on an input region. (Otherwise it's impossible to estimate the score!)



Step 3: Doubly robust estimation

Now, define the *doubly robust (DR) estimator* $\hat{\psi}_{dr}$:

$$\hat{\psi}_{dr} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mu}_0(X_i) + \frac{1 - R_i}{1 - \hat{\pi}(X_i)} \left(S_i - \hat{\mu}_0(X_i) \right) \right],$$

where $\hat{\mu}_0$ and $\hat{\pi}$ are *nuisance function estimators* (e.g., ensemble methods).

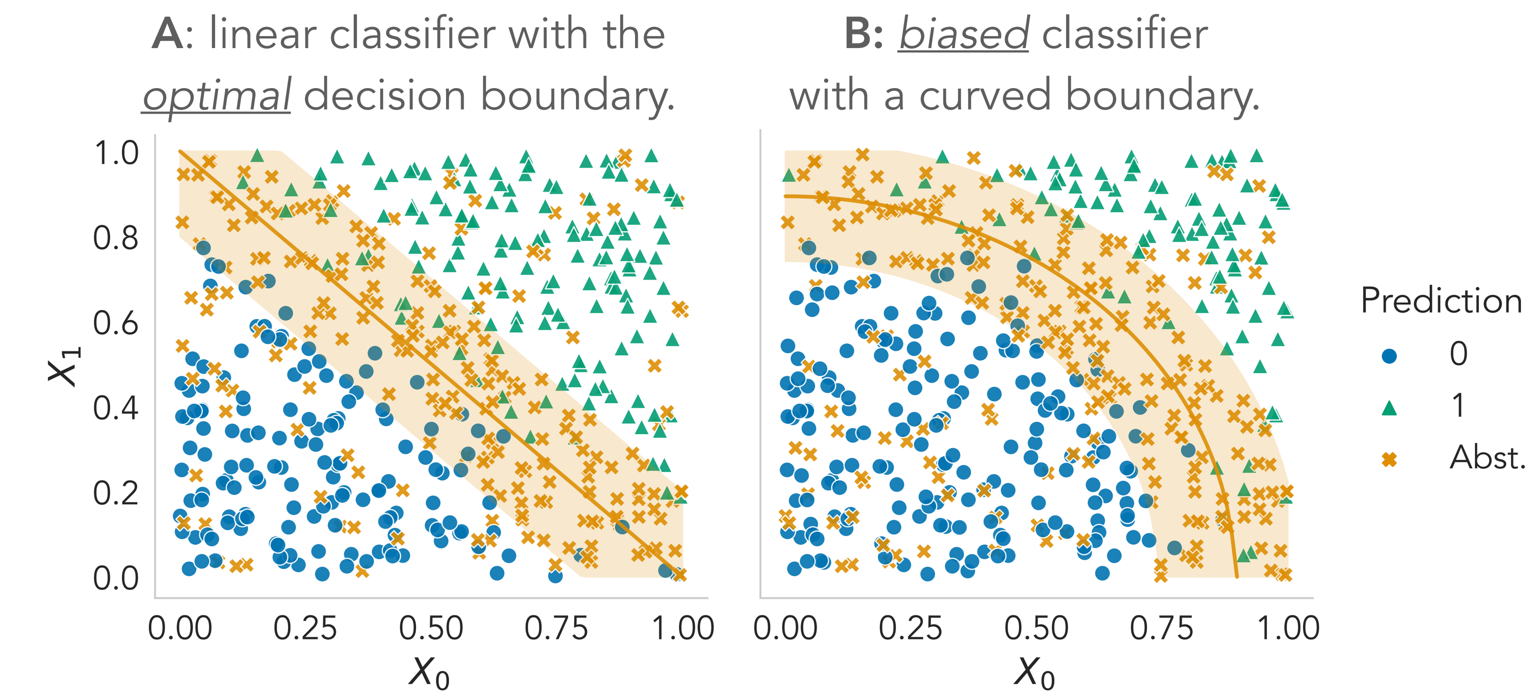
<u>Theorem (Informal).</u> *With sufficiently flexible nuisance function estimators $\hat{\mu}_0$ and $\hat{\pi}$, the DR estimator is* ***asymptotically normal and efficient for*** *ψ:*

$$\sqrt{n} \left(\hat{\psi}_{dr} - \psi \right) \rightsquigarrow \mathcal{N} \left(0, \text{Var}_{\mathbb{P}}(\text{IF}) \right) .$$

The nuisance functions are estimated via cross-fitting (Robins et al., 2008).

Experiments

Simulated data: Comparing abstaining binary classifiers (MAR)



Two abstaining classifiers, depicted using their decision boundary (orange), predictions (●/▲), and abstentions (x).

Nuisance fn.	95% CI's	Plug-in	IPW	DR
Random Forest	*Miscoverage*	0.64	0.14	0.05
	Width	*0.02*	*0.13*	*0.07*
Super Learner	*Miscoverage*	0.91	0.03	0.05
	Width	*0.01*	*0.12*	*0.06*

Miscoverage and width of the 95% CI for estimating Δ^{AB}, based on *accuracy*. Baselines: plug-in & IPW. N=2,000; averaged over 1,000 repeated simulations.

The doubly robust CI achieves the correct miscoverage rate while having a small width (i.e., it is efficient).

Real data: comparing abstaining CNNs for image classification

Base clf.	Abstention	Δ	Reject null?	95% CI
Same	Different	0.000	**No**	**(-0.014, 0.008)**
Different	Same	*-0.029*	***Yes***	***(-0.051, -0.028)***

Hypothesis tests and 95% CIs for comparing abstaining classifiers built using pre-trained VGG-16 networks on CIFAR-100 dataset (N=5,000). Null: $\Delta = 0$.

The theory is applicable to testing or estimating the counterfactual score difference between nonparametric predictors.

Find us online!

Paper

Code