Bounding the fairness and accuracy of classifiers from population statistics

ICML 2020

Sivan Sabato and Elad Yom-Tov

The 1-slide summary

- We show how to study a classifier without even a black box access to the classifier and without validation data.

The 1-slide summary

- We show how to study a classifier without even a black box access to the classifier and without validation data.
- Our methodology makes provable inferences about classifier **quality**.

The 1-slide summary

- We show how to study a classifier without even a black box access to the classifier and without validation data.
- Our methodology makes provable inferences about classifier **quality**.
- The quality combines the accuracy and the **fairness** of the classifier.

The 1-slide summary

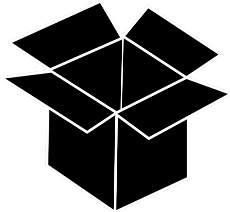
- We show how to study a classifier without even a black box access to the classifier and without validation data.
- Our methodology makes provable inferences about classifier **quality**.
- The quality combines the accuracy and the **fairness** of the classifier.
- We make inferences using a small number of aggregate statistics.

The 1-slide summary

- We show how to study a classifier without even a black box access to the classifier and without validation data.
- Our methodology makes provable inferences about classifier **quality**.
- The quality combines the accuracy and the **fairness** of the classifier.
- We make inferences using a small number of aggregate statistics.
- We demonstrate in experiments a wide range of possible applications.

The 1-slide summary

- We show how to study a classifier without even a black box access to the classifier and without validation data.
- Our methodology makes provable inferences about classifier **quality**.
- The quality combines the accuracy and the **fairness** of the classifier.
- We make inferences using a small number of aggregate statistics.
- We demonstrate in experiments a wide range of possible applications.



Introduction

- Classifiers affect many aspects of our lives.
- But some of these classifiers cannot be directly validated:
  - Unavailability of representative individual-level validation data
  - Company of government secret: not even black-box access
- What can we infer about a classifier using only aggregate statistics?

What can we tell about an unpublished classifier?

A motivating example:

- A health insurance company classifies whether a client is as "at risk" for some medical condition.
- We do not know how this classification is done;
- We have no individual classification data.

What can we tell about an unpublished classifier?

A motivating example:

- A health insurance company classifies whether a client is as "at risk" for some medical condition.
- We do not know how this classification is done;
- We have no individual classification data.
- But we would still like to study the properties of the classifier:
  - Accuracy
  - Fairness

What can we tell about an unpublished classifier?

A motivating example:

- A health insurance company classifies whether a client is as "at risk" for some medical condition.
- We do not know how this classification is done;
- We have no individual classification data.
- But we would still like to study the properties of the classifier:
  - Accuracy
  - Fairness
- Can this be done with minimal information about the classifier?

Fairness

- Fairness is defined with respect to some attribute of the individual.
  - E.g., race, age, gender, state of residence
- We will be interested in attributes with several different values.
- A **sub-population** includes the individual who share the attribute value (e.g., same race/age bracket/state, etc.).

Fairness

- Fairness is defined with respect to some attribute of the individual.
  - E.g., race, age, gender, state of residence
- We will be interested in attributes with several different values.
- A **sub-population** includes the individual who share the attribute value (e.g., same race/age bracket/state, etc.).
- A fair classifier **treats all sub-populations the same**.
- **Equalized Odds** [Hardt et. al, 2016]:
  The false positive rate (FPR) and the false negative rate (FNR) are fixed across all sub-populations.

Using population statistics

Back to the example: Use available information

Using population statistics

Back to the example: Use available information

- Size of each sub-population
- Prevalence rate of the condition in each sub-population
- Fraction of positive predictions in each sub-population.

State	Population Fraction	Have condition	Classified as positive
California	12.2%	0.3%	0.4%
Texas	8.6%	1.2%	5%
...	...	...	...

Using population statistics

Back to the example: Use available information

- Size of each sub-population
- Prevalence rate of the condition in each sub-population
- Fraction of positive predictions in each sub-population.

State	Population Fraction	Have condition	Classified as positive
California	12.2%	0.3%	0.4%
Texas	8.6%	1.2%	5%
...	...	...	...

- What is the accuracy of this classifier? What is the fairness?

Using population statistics

Back to the example: Use available information

- Size of each sub-population
- Prevalence rate of the condition in each sub-population
- Fraction of positive predictions in each sub-population.

State	Population Fraction	Have condition	Classified as positive
California	12.2%	0.3%	0.4%
Texas	8.6%	1.2%	5%
...	...	...	...

- What is the accuracy of this classifier? What is the fairness?
- Without individual data, there are many possibilities:

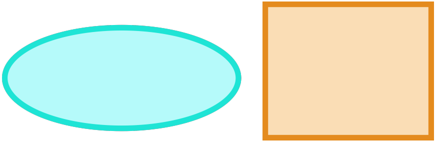
Using population statistics

Back to the example: Use available information

- Size of each sub-population
- Prevalence rate of the condition in each sub-population
- Fraction of positive predictions in each sub-population.

State	Population Fraction	Have condition	Classified as positive
California	12.2%	0.3%	0.4%
Texas	8.6%	1.2%	5%
...	...	...	...

- What is the accuracy of this classifier? What is the fairness?
- Without individual data, there are many possibilities:



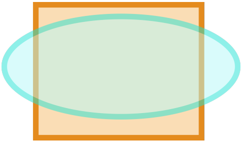
Using population statistics

Back to the example: Use available information

- Size of each sub-population
- Prevalence rate of the condition in each sub-population
- Fraction of positive predictions in each sub-population.

State	Population Fraction	Have condition	Classified as positive
California	12.2%	0.3%	0.4%
Texas	8.6%	1.2%	5%
...	...	...	...

- What is the accuracy of this classifier? What is the fairness?
- Without individual data, there are many possibilities:



The relationship between accuracy and fairness

- If fairness or error are constrained, this also constrains the other.

The relationship between accuracy and fairness

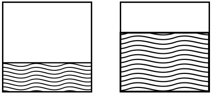
- If fairness or error are constrained, this also constrains the other.
- Example:

	Population Fraction	Have condition	Classified as positive
State A	1/2	1/3	1/2
State B	1/2	2/3	2/3

The relationship between accuracy and fairness

- If fairness or error are constrained, this also constrains the other.
- Example:

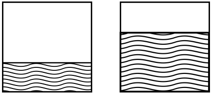
	Population Fraction	Have condition	Classified as positive
State A	1/2	1/3	1/2
State B	1/2	2/3	2/3

  - True positives: .

The relationship between accuracy and fairness

- If fairness or error are constrained, this also constrains the other.
- Example:

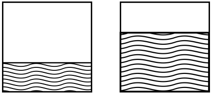
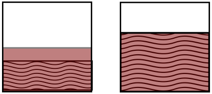
	Population Fraction	Have condition	Classified as positive
State A	1/2	1/3	1/2
State B	1/2	2/3	2/3

  - True positives: .
  - Which are the predicted positives?

The relationship between accuracy and fairness

- If fairness or error are constrained, this also constrains the other.
- Example:

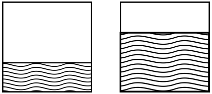
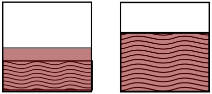
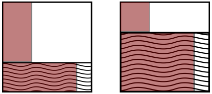
	Population Fraction	Have condition	Classified as positive
State A	1/2	1/3	1/2
State B	1/2	2/3	2/3

  - True positives: .
  - Which are the predicted positives?
  - Smallest error: . Error of 12.5%, unfair.

The relationship between accuracy and fairness

- If fairness or error are constrained, this also constrains the other.
- Example:

	Population Fraction	Have condition	Classified as positive
State A	1/2	1/3	1/2
State B	1/2	2/3	2/3

  - True positives: .
  - Which are the predicted positives?
  - Smallest error: . Error of 12.5%, unfair.
  - Fair solution: . 25% error.

Balancing accuracy and fairness

- The two measures:

Balancing accuracy and fairness

- The two measures:
  - `error`: Fraction of the population classified with the wrong label.

Balancing accuracy and fairness

- The two measures:
  - `error`: Fraction of the population classified with the wrong label.
  - `unfairness`: Fraction of the population treated differently than a common baseline. We expand on this next.

Balancing accuracy and fairness

- The two measures:
  - `error`: Fraction of the population classified with the wrong label.
  - `unfairness`: Fraction of the population treated differently than a common baseline. We expand on this next.
- Combine both desiderata:
  For $\beta \in [0, 1]$,

$$\texttt{discrepancy}_\beta := \beta \cdot \texttt{unfairness} + (1 - \beta) \cdot \texttt{error},$$

Balancing accuracy and fairness

- The two measures:
  - `error`: Fraction of the population classified with the wrong label.
  - `unfairness`: Fraction of the population treated differently than a common baseline. We expand on this next.
- Combine both desiderata:
  For $\beta \in [0, 1]$,

$$\texttt{discrepancy}_\beta := \beta \cdot \texttt{unfairness} + (1 - \beta) \cdot \texttt{error},$$

- β defines a trade-off between (un)fairness and error.

Balancing accuracy and fairness

- The two measures:
  - $\texttt{error}$: Fraction of the population classified with the wrong label.
  - $\texttt{unfairness}$: Fraction of the population treated differently than a common baseline. We expand on this next.
- Combine both desiderata:
  For $\beta \in [0, 1]$,

$$\texttt{discrepancy}_\beta := \beta \cdot \texttt{unfairness} + (1 - \beta) \cdot \texttt{error},$$

- β defines a trade-off between (un)fairness and error.
- By **lower-bounding** $\texttt{discrepancy}_\beta$, we can answer:

Balancing accuracy and fairness

- The two measures:
  - `error`: Fraction of the population classified with the wrong label.
  - `unfairness`: Fraction of the population treated differently than a common baseline. We expand on this next.
- Combine both desiderata:
  For $\beta \in [0, 1]$,

$$\texttt{discrepancy}_\beta := \beta \cdot \texttt{unfairness} + (1 - \beta) \cdot \texttt{error},$$

- β defines a trade-off between (un)fairness and error.
- By **lower-bounding** $\texttt{discrepancy}_\beta$, we can answer:
  - What is the minimal unfairness that the classifier must have, given an upper bound on its error?

Balancing accuracy and fairness

- The two measures:
  - `error`: Fraction of the population classified with the wrong label.
  - `unfairness`: Fraction of the population treated differently than a common baseline. We expand on this next.
- Combine both desiderata:
  For $\beta \in [0, 1]$,

$$\texttt{discrepancy}_\beta := \beta \cdot \texttt{unfairness} + (1 - \beta) \cdot \texttt{error},$$

- β defines a trade-off between (un)fairness and error.
- By **lower-bounding** $\texttt{discrepancy}_\beta$, we can answer:
  - What is the minimal unfairness that the classifier must have, given an upper bound on its error?
  - What is the minimal error that the classifier must have, given an upper bound on its unfairness?

Balancing accuracy and fairness

- The two measures:
  - $\texttt{error}$: Fraction of the population classified with the wrong label.
  - $\texttt{unfairness}$: Fraction of the population treated differently than a common baseline. We expand on this next.
- Combine both desiderata:
  For $\beta \in [0, 1]$,

$$\texttt{discrepancy}_\beta := \beta \cdot \texttt{unfairness} + (1 - \beta) \cdot \texttt{error},$$

- β defines a trade-off between (un)fairness and error.
- By **lower-bounding** $\texttt{discrepancy}_\beta$, we can answer:
  - What is the minimal unfairness that the classifier must have, given an upper bound on its error?
  - What is the minimal error that the classifier must have, given an upper bound on its unfairness?
  - What is the minimal combined cost of this classifier?

Quanitifying unfairness

- Decompose the conditional distribution of predictions given labels:
  - A baseline distribution which is common to all sub-populations; FPR $= \alpha^1$ and FNR $= \alpha^0$,
  - A nuisance distribution for each sub-population s; FPR $= \alpha^1_s$ and FNR $= \alpha^0_s$,
  - The distribution for sub-population s is a mixture:

$$\eta_s \cdot \mathrm{Nuisance}_s + (1 - \eta_s) \cdot \mathrm{Baseline}.$$

Quanitifying unfairness

- Decompose the conditional distribution of predictions given labels:
  - A baseline distribution which is common to all sub-populations; FPR $= \alpha^1$ and FNR $= \alpha^0$,
  - A nuisance distribution for each sub-population s; FPR $= \alpha^1_s$ and FNR $= \alpha^0_s$,
  - The distribution for sub-population s is a mixture:

$$\eta_s \cdot \text{Nuisance}_s + (1 - \eta_s) \cdot \text{Baseline}.$$

  - Define **unfairness** as the fraction of the population that is **treated differently from the baseline treatment** $= \sum_s \eta_s$.

Quanitifying unfairness

- Decompose the conditional distribution of predictions given labels:
  - A baseline distribution which is common to all sub-populations; FPR $= \alpha^1$ and FNR $= \alpha^0$,
  - A nuisance distribution for each sub-population s; FPR $= \alpha^1_s$ and FNR $= \alpha^0_s$,
  - The distribution for sub-population s is a mixture:

$$\eta_s \cdot \mathrm{Nuisance}_s + (1 - \eta_s) \cdot \mathrm{Baseline}.$$

  - Define **unfairness** as the fraction of the population that is **treated differently from the baseline treatment** $= \sum_s \eta_s$.
  - The decomposition to baseline and nuisance is unobserved.

Quanitifying unfairness

- Decompose the conditional distribution of predictions given labels:
  - A baseline distribution which is common to all sub-populations; FPR $= \alpha^1$ and FNR $= \alpha^0$,
  - A nuisance distribution for each sub-population s; FPR $= \alpha^1_s$ and FNR $= \alpha^0_s$,
  - The distribution for sub-population s is a mixture:

$$\eta_s \cdot \text{Nuisance}_s + (1 - \eta_s) \cdot \text{Baseline}.$$

  - Define **unfairness** as the fraction of the population that is **treated differently from the baseline treatment** $= \sum_s \eta_s$.
  - The decomposition to baseline and nuisance is unobserved.
  - Set η_s to the minimum consistent with the input statistics.

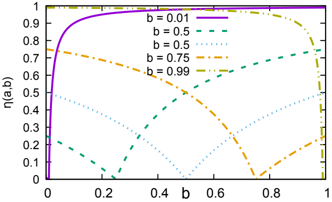
Quanitifying unfairness

- Decompose the conditional distribution of predictions given labels:
  - A baseline distribution which is common to all sub-populations; FPR $= \alpha^1$ and FNR $= \alpha^0$,
  - A nuisance distribution for each sub-population s; FPR $= \alpha^1_s$ and FNR $= \alpha^0_s$,
  - The distribution for sub-population s is a mixture:

$$\eta_s \cdot \text{Nuisance}_s + (1 - \eta_s) \cdot \text{Baseline}.$$

  - Define **unfairness** as the fraction of the population that is **treated differently from the baseline treatment** $= \sum_s \eta_s$.
  - The decomposition to baseline and nuisance is unobserved.
  - Set η_s to the minimum consistent with the input statistics.

$$\eta(\alpha^y, \alpha^y_s) = \begin{cases} 1 - \alpha^y_s/\alpha^y & \alpha^y_s < \alpha^y \\ 1 - (1 - \alpha^y_s)/(1 - \alpha^y) & \alpha^y_s > \alpha^y \\ 0 & \alpha^y_s = \alpha^y. \end{cases}$$



Lower-bounding $\texttt{discrepancy}_\beta$

- Given known FPRs and FNRs $\{\alpha_s^y\}$ in each sub-population,

$$\texttt{discrepancy}_\beta(\{\alpha_s^y\}) =$$

$$\beta \cdot \min_{(\alpha^0,\alpha^1)\in[0,1]^2} \sum_{g\in\mathcal{G}} w_s \sum_{y\in\mathcal{Y}} \pi_s^y \eta(\alpha^y, \alpha_s^y) + (1-\beta)\cdot \sum_{g\in\mathcal{G}} w_s \sum_{y\in\mathcal{Y}} \pi_s^y \alpha_s^y.$$

$$w_s := P(\text{attribute value is } s)$$

$$\pi_s := P(\text{positive label} \mid s)$$

$$\hat{p}_s := P(\text{positive prediction} \mid s)$$

Lower-bounding $\texttt{discrepancy}_\beta$

- Given known FPRs and FNRs $\{\alpha_s^y\}$ in each sub-population,

$$\texttt{discrepancy}_\beta(\{\alpha_s^y\}) =$$
$$\beta \cdot \min_{(\alpha^0,\alpha^1)\in[0,1]^2} \sum_{g\in\mathcal{G}} w_s \sum_{y\in\mathcal{Y}} \pi_s^y \eta(\alpha^y, \alpha_s^y) + (1-\beta)\cdot \sum_{g\in\mathcal{G}} w_s \sum_{y\in\mathcal{Y}} \pi_s^y \alpha_s^y.$$

$$w_s := P(\text{attribute value is } s)$$
$$\pi_s := P(\text{positive label} \mid s)$$
$$\hat{p}_s := P(\text{positive prediction} \mid s)$$

- We derive a lower bound on $\min_{\{\alpha_s^y\}} \texttt{discrepancy}_\beta(\{\alpha_s^y\})$ subject to the constraints imposed by $\{w_s, \pi_s, \hat{p}_s\}$.

Lower-bounding $\texttt{discrepancy}_\beta$

- Given known FPRs and FNRs $\{\alpha_s^y\}$ in each sub-population,

$$\texttt{discrepancy}_\beta(\{\alpha_s^y\}) =$$
$$\beta \cdot \min_{(\alpha^0,\alpha^1)\in[0,1]^2} \sum_{g\in\mathcal{G}} w_s \sum_{y\in\mathcal{Y}} \pi_s^y \eta(\alpha^y, \alpha_s^y) + (1-\beta)\cdot \sum_{g\in\mathcal{G}} w_s \sum_{y\in\mathcal{Y}} \pi_s^y \alpha_s^y.$$

$$w_s := P(\text{attribute value is } s)$$
$$\pi_s := P(\text{positive label} \mid s)$$
$$\hat{p}_s := P(\text{positive prediction} \mid s)$$

- We derive a lower bound on $\min_{\{\alpha_s^y\}} \texttt{discrepancy}_\beta(\{\alpha_s^y\})$ subject to the constraints imposed by $\{w_s, \pi_s, \hat{p}_s\}$.

Theorem

The minimum of $\texttt{discrepancy}_\beta(\{\alpha_s^y\})$ subject to the constraints imposed by $\{w_s, \pi_s, \hat{p}_s\}$ is obtained by an assignment in a small number of one-dimensional solution sets.

Experiments: Tightness of lower bound

- (In all experiments, sub-populations are defined by state of residence.)

Experiments: Tightness of lower bound

- (In all experiments, sub-populations are defined by state of residence.)
- We obtain a lower bound; how tight is it in practice?

Experiments: Tightness of lower bound

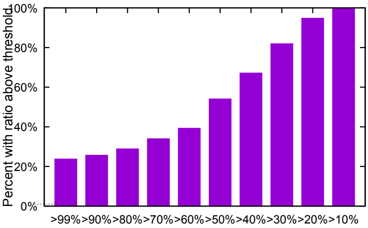
- (In all experiments, sub-populations are defined by state of residence.)

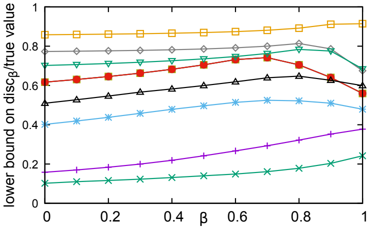
- We obtain a lower bound; how tight is it in practice?
- Generated hundreds of classifiers from the US Census data set.
- The classifiers are known and we can calculate their true properties.

Experiments: Tightness of lower bound

- (In all experiments, sub-populations are defined by state of residence.)

- We obtain a lower bound; how tight is it in practice?
- Generated hundreds of classifiers from the US Census data set.
- The classifiers are known and we can calculate their true properties.
- Left plot: Compared the lower bound on $\texttt{discrepancy}_1 \equiv \texttt{unfairness}$ with the true unfairness.
- Right plot: For randomly selected classifiers, the ratio between the true value and the lower bound for $\beta \in [0, 1]$.





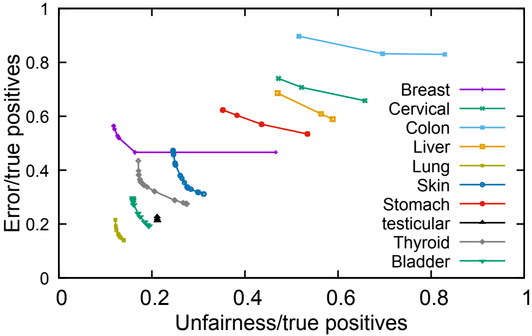
Experiments: Making inferences in the wild (1)

- In the following experiments, $\texttt{discrepancy}_\beta$ is unknown.
- We calculate (unfairness,error) Pareto-curves as a function of β.

Experiments: Making inferences in the wild (1)

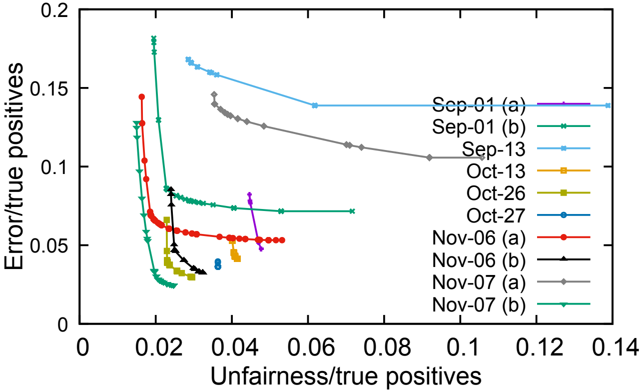
- In the following experiments, $\texttt{discrepancy}_\beta$ is unknown.
- We calculate (unfairness,error) Pareto-curves as a function of β.

- Experiment 1: Identify if anonymous individuals have a certain cancer from their search queries in Bing.
- Classify as positive if user searched for said cancer.
- True positive rates per state from CDC data.
- Results lower-bound the quality of these classifiers.



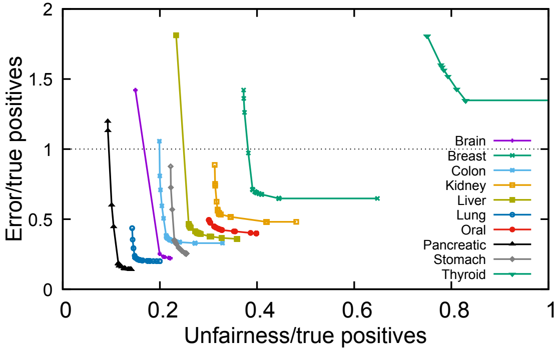
Experiments: Making inferences in the wild (2)

- Experiment (2): Studied 10 pre-election polls from the 2016 US presidential elections.
- Treat each poll as a classifier from individual to vote.
- How biased are these polls in their treatment of different states?



Experiments: Making inferences in the wild (3)

- Experiment (3): Compare cancer mortality rates in different states
- "True positive" rates: cancer mortality rates in each state
- "Predicted" rates: expected mortality in the state based on cancer prevalence and **overall** US mortality.
- "Classifier" maps an individual to an outcome (living/deceased)
- Error and unfairness can speculatively point to patterns in health care access or in cancer strains.



Summary

- We showed how a small set of aggregate statistics can be used to make strong inferences about the quality of the classifier.
- The methodology can be applied to a range of applications:
  - Estimating the quality of a classifier during development stages
  - Studying classifiers of public importance
  - Analysis of statistical phenomena by defining an appropriate classifier
- Extending this toolbox is an important research direction with many open problems.

Summary

- We showed how a small set of aggregate statistics can be used to make strong inferences about the quality of the classifier.
- The methodology can be applied to a range of applications:
  - Estimating the quality of a classifier during development stages
  - Studying classifiers of public importance
  - Analysis of statistical phenomena by defining an appropriate classifier
- Extending this toolbox is an important research direction with many open problems.