Learning to audit data-driven sociotechnical systems

Data Justice Lab Workshop

Javier Sánchez-Monedero
@javisamo
sanchez-monederoj at cardiff.ac.uk

October 18, 2019

Cardiff University, UK

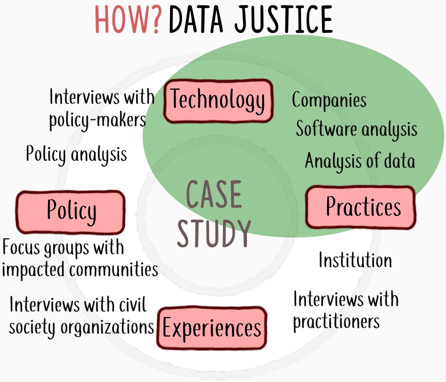
Data Justice Lab

datajusticelab.org

datajusticeproject.net

Data Justice

'Data Justice'



https://datajusticeproject.net/about

Summary of machine learning

Traditional programming

Explicit rules:

```
if email contains Viagra
  then mark is-spam;
if email contains ...;
if email contains ...;
```

Example from Jason's Machine Learning 101

Machine learning programs

Learn from examples:

```
try to classify some emails;
change self to reduce errors;
repeat;
```

...then use the model to label

Summary of machine learning

Traditional programming

Explicit rules:

```
if email contains Viagra
  then mark is-spam;
if email contains ...;
if email contains ...;
```

Example from Jason's Machine Learning 101

Machine learning programs

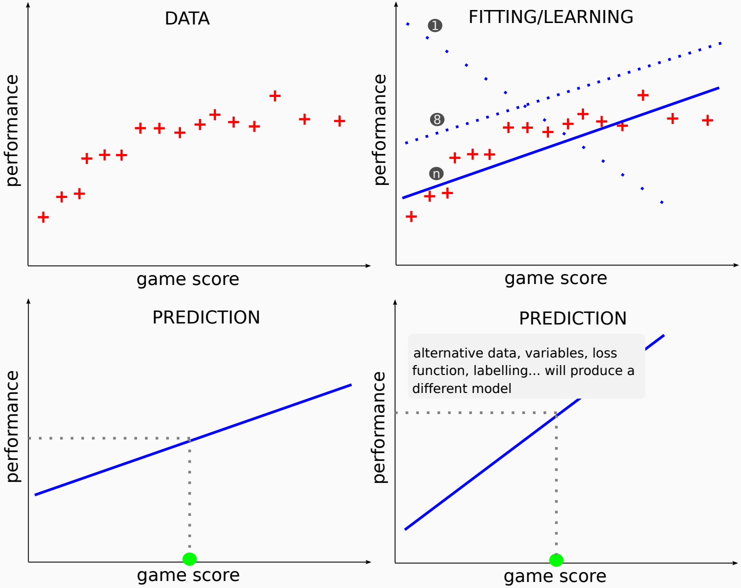
Learn from examples:

```
try to classify some emails;
change self to reduce errors;
repeat;
```

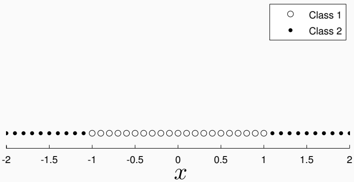
...then use the model to label

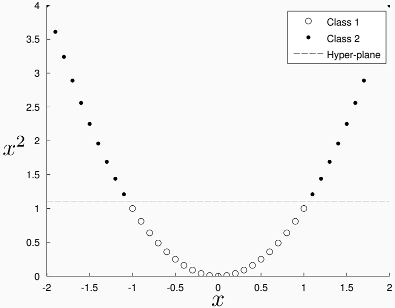
Since nobody is explicitly programming it, it is often assumed to be fair, non-discriminative, avoid human biases, etc.

Summary of machine learning



Data transformation





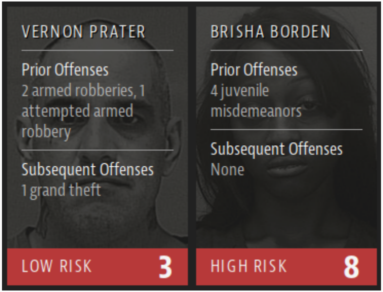
Many methods build/learn/create geometric transformations of the data to optimize the classification/prediction task.

The COMPAS case revisited

The COMPAS case revisited

- **COMPAS**: tool to assess the likelihood of a defendant becoming a recidivist.
- Builts a model with **historical records**
- **Input Variables**: number of priors, number of misdemeanor, gender, ethnic group, age, *environment*...
- **Target variable**: risk scale (1-10). High scores suggest inprisionement or bail.

Racial discrimination

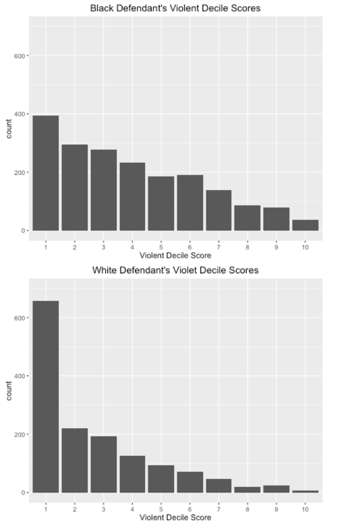


Source Angwin and Larson [2016]

ProPublica: the system is biased against blacks since it overestimates the risk for blacks (different false positive rates: 44.8% vs 23.4%)
Northpointe: the tool does not discriminates because it equally estimates high-risk scores (true positives are equal across groups: 63% vs 59%)

Compatible claims

Both definitions of fairness are mathematically compatible because the prevalence is different for 'blacks' and 'whites' Chouldechova [2017].



Fuente Larson and Angwin [2016]

What if we remove the race variable?

Table 1. Human versus COMPAS algorithmic predictions from 1000 defendants. Overall accuracy is specified as percent correct, AUC-ROC, and criterion sensitivity (d') and bias (β). See also Fig. 1.

	(A) Human (no race)	(B) Human (race)	(C) COMPAS
Accuracy (overall)	67.0%	66.5%	65.2%
AUC-ROC (overall)	0.71	0.71	0.70
d'/β (overall)	0.86/1.02	0.83/1.03	0.77/1.08
Accuracy (black)	68.2%	66.2%	64.9%
Accuracy (white)	67.6%	67.6%	65.7%
False positive (black)	37.1%	40.0%	40.4%
False positive (white)	27.2%	26.2%	25.4%
False negative (black)	29.2%	30.1%	30.9%
False negative (white)	40.3%	42.1%	47.9%

Source Dressel and Farid [2018]

...but there can be proxies to the 'race' variable.

Let's remove almost all the variables

Table 2. Algorithmic predictions from 7214 defendants. Logistic regression with 7 features (A) (LR_7), logistic regression with 2 features (B) (LR_2), a nonlinear SVM with 7 features (C) (NL-SVM), and the commercial COMPAS software with 137 features (D) (COMPAS). The results in columns (A), (B), and (C) correspond to the average testing accuracy over 1000 random 80%/20% training/testing splits. The values in the square brackets correspond to the 95% bootstrapped [columns (A), (B), and (C)] and binomial [column (D)] confidence intervals.

	(A) LR_7	(B) LR_2	(C) NL-SVM	(D) COMPAS
Accuracy (overall)	66.6% [64.4, 68.9]	66.8% [64.3, 69.2]	65.2% [63.0, 67.2]	65.4% [64.3, 66.5]
Accuracy (black)	66.7% [63.6, 69.6]	66.7% [63.5, 69.2]	64.3% [61.1, 67.7]	63.8% [62.2, 65.4]
Accuracy (white)	66.0% [62.6, 69.6]	66.4% [62.6, 70.1]	65.3% [61.4, 69.0]	67.0% [65.1, 68.9]
False positive (black)	42.9% [37.7, 48.0]	45.6% [39.9, 51.1]	31.6% [26.4, 36.7]	44.8% [42.7, 46.9]
False positive (white)	25.3% [20.1, 30.2]	25.3% [20.6, 30.5]	20.5% [16.1, 25.0]	23.5% [20.7, 26.5]
False negative (black)	24.2% [20.1, 28.2]	21.6% [17.5, 25.9]	39.6% [34.2, 45.0]	28.0% [25.7, 30.3]
False negative (white)	47.3% [40.8, 54.0]	46.1% [40.0, 52.7]	56.6% [50.3, 63.5]	47.7% [45.2, 50.2]

Source Dressel and Farid [2018]

Even when only using the number of priors and the age the model still overestimating the risk for the black community (column B)!

Risk as a proxy for race (and other groups)

Thoughts from Harcourt [2010]:

- Data-driven assessment has been reducing predictive variables and relying more on criminal history of the person
- Criminal history is linked to race, there it is a proxy for race.
- Risk assessment interventions in the US has always produced massive incarcelation of the black community.

ML problem and actual probleme

What is the model actually optimizing? $\hat{y} = f(\mathbf{x})$

ML problem and actual probleme

What is the model actually optimizing? $\hat{y} = f(\mathbf{x})$

Hype

- "Predictive policing"
- "Minority Report"
- ...

ML problem and actual probleme

What is the model actually optimizing? $\hat{y} = f(\mathbf{x})$

Hype

- “Predictive policing”
- “Minority Report”
- ...

Actual prediction

The system is not predicting future crimes, but arrest for future crimes.

Conclusions

Discussion

- What are the variables and what they represent?

Discussion

- What are the variables and what they represent?
- What is the actual task the system is solving/optimizing?

Discussion

- What are the variables and what they represent?
- What is the actual task the system is solving/optimizing?
- Limitations of statistical definitions of fairness

Discussion

- What are the variables and what they represent?
- What is the actual task the system is solving/optimizing?
- Limitations of statistical definitions of fairness
- Classify, predict, score, estimate...

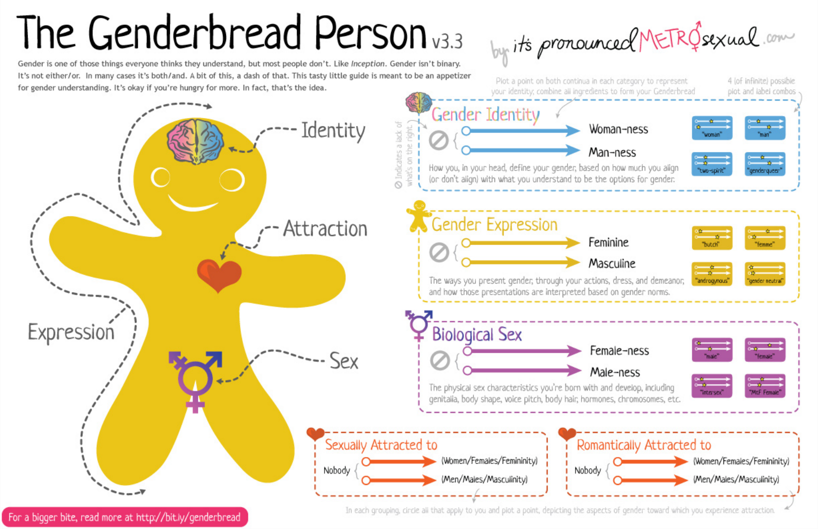
Discussion

- What are the variables and what they represent?
- What is the actual task the system is solving/optimizing?
- Limitations of statistical definitions of fairness
- Classify, predict, score, estimate...
- US understanding of discrimination and demographic groups

Discussion

- What are the variables and what they represent?
- What is the actual task the system is solving/optimizing?
- Limitations of statistical definitions of fairness
- Classify, predict, score, estimate...
- US understanding of discrimination and demographic groups
- Does the data-driven proposal works better than the current (human) process?

Information enconding



More at https://www.genderbread.org/

Working Paper

How to (partially) evaluate automated decision systems. Working paper by Javier Sánchez-Monedero and Lina Dencik. December 2018.
https://datajusticeproject.net/working-papers/

Thanks!

References i

J. Angwin and J. Larson. Machine Bias. *ProPublica*, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

A. Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, June 2017. ISSN 2167-6461, 2167-647X. doi: 10.1089/big.2016.0047. URL http://www.liebertpub.com/doi/10.1089/big.2016.0047.

J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1): eaao5580, Jan. 2018. ISSN 2375-2548. doi: 10.1126/sciadv.aao5580. URL http://advances.sciencemag.org/content/4/1/eaao5580.

B. E. Harcourt. Risk as a Proxy for Race. SSRN Scholarly Paper ID 1677654, Social Science Research Network, Rochester, NY, Sept. 2010. URL https://papers.ssrn.com/abstract=1677654.

J. Larson and J. Angwin. How We Analyzed the COMPAS Recidivism Algorithm, May 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.