Societal consequences of machine learning

Societal consequences

Often data about individuals is used in data-driven applications.

Some examples:

1. Credit card transactions for fraud detection.
2. Medical/genetic test results for disease association studies.
3. Predictive policing.

Such applications face major social/policy issues, including **privacy** and **fairness**.

Privacy

Privacy

Suppose each point in data set D corresponds to potentially sensitive information about an individual.

- **Question**: Can the output of an algorithm $\mathcal{A}$ run on D reveal information about an individual?

  Answer: Yes!

  But often the answer we get out is also socially useful (e.g., helps us detect credit card fraud).

What kind of privacy can we expect from such data-driven applications?

Side-information

Suppose $D = \{x_i\}_{i=1}^n$, where $x_i \in [0, \infty)$ is the salary of individual i.

And suppose $\mathcal{A}(D)$ simply returns the average of the salaries in D.

- I don't know Alice's salary, but I know her position in her department, and I know that the salary for that position is $3\times$ the average salary.
- If I learn the average salary (i.e., output of $\mathcal{A}(D)$), then I also learn Alice's salary.

Potential privacy "attackers" can have side-information.

Real example (Narayanan and Shmatikov, 2008):
"Anonymized" Netflix data set could be de-anonymized by using public IMDb movie ratings.

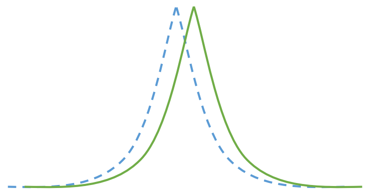
Question: for people who used both IMDb and Netflix, what information was actually "private" in the Netflix data?

What can we do

Differential privacy: An attacker should not learn anything about an individual i from $\mathcal{A}(D)$ that he could not have already learned from $\mathcal{A}(D \setminus \{x_i\})$.

Formal definition (Dwork, McSherry, Nissim, & Smith, 2006): A randomized algorithm $\mathcal{A}$ provides ε-**differential privacy** if, for all data sets D and D' that differ in just one individual's data point,

$$\mathbb{P}(\mathcal{A}(D) = z) \;\in\; (1 \pm \varepsilon) \cdot \mathbb{P}(\mathcal{A}(D') = z) \quad \forall z\,.$$



In other words, whatever an attacker can learn about Alice from $\mathcal{A}(D)$ is just about the same as what he could have learned if Alice's data point was replaced with arbitrary (e.g., garbage) data point.

Simple example

Average salary

1. Releasing average does not provide differential privacy because attacker who knows everyone's but Alice's salary will be able to learn Alice's salary.
2. Releasing average + $\mathrm{N}(0, \sigma^2)$ noise provides ε-differential privacy when

$$\sigma \gg \frac{\text{maximum salary}}{\varepsilon n}\,.$$

   (Technically, should use a slightly different noise distribution.)

For statistical analysis: often data set itself is just a random sample, and we care more about broader population.

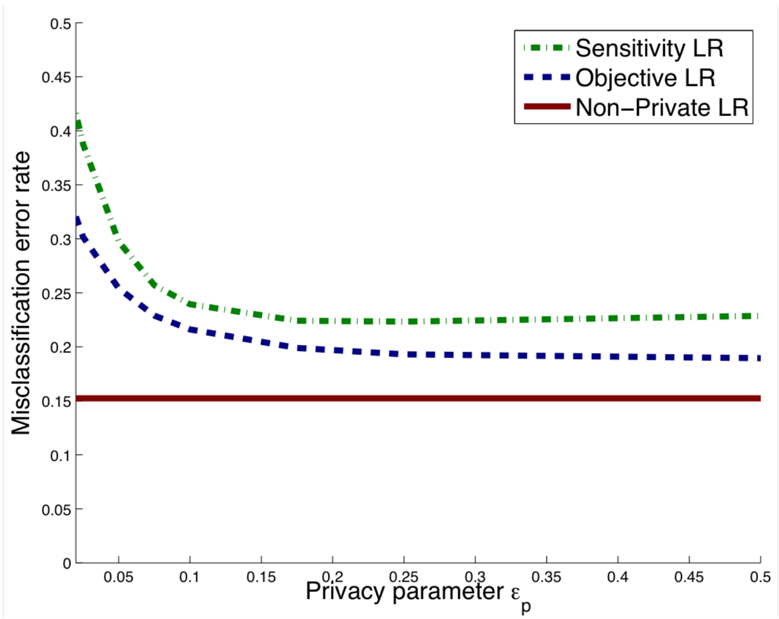
Average on sample will differ from true population mean by $\Theta(n^{-1/2})$ anyway!

Notes:

- With side-information ("Alice's salary $= 3\times$ average"), can still learn Alice's salary up to some small error.
  But whether or not Alice's data is used in this computation does not change this.
- The "maximum salary" in the numerator is troubling—can nullify utility.

Privacy-preserving logistic regression

(Chaudhuri, Monteleoni, and Sarwate, 2011)



Privacy: summary

- Many ways data-driven applications and analyses can leak private information.
- Simple anonymization / obfuscation are easily broken!
- Often, there is a trade-off between privacy and utility.

Fairness

Unfair / harmful behavior

Data-driven applications can have unpredictable and harmful behavior.

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI
December 24, 2012

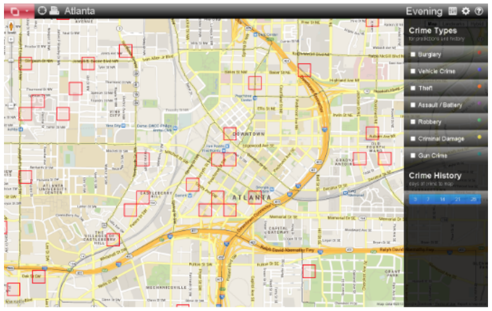
It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

In what appears to be an unintended side effect of Staples' pricing methods—likely a function of retail competition with its rivals—the Journal's testing also showed that areas that tended to see the discounted prices had a higher average income than areas that tended to see higher prices.

Unfair / harmful behavior

Data-driven applications can have unpredictable and harmful behavior.



Fairness?

What is **fairness**?
Difficult to precisely define.

- **No disparate impact**:

$$P(\widehat{Y} = + \mid Z = 1) \;\approx\; P(\widehat{Y} = + \mid Z = 0)\,,$$

  where Z is a binary protected attribute (e.g., race, religion, sex).

- **Equal treatment**:

$$\widehat{Y} \text{ and } Z \text{ are conditionally independent given } Y\,.$$

- **Equality of opportunity** (Rawls):

$$P(\widehat{Y} = + \mid Z = 1, Y = +) \;\approx\; P(\widehat{Y} = + \mid Z = 0, Y = +)\,.$$

- ...

Something to do with statistical discrepancies
(e.g., in monetary costs, personal harms, service quality).

Example: predictive policing

Financial Times article from August 22, 2014 by Gillian Tett:

> *After all, as the former CPD computer experts point out, the algorithms in themselves are neutral. "This program had absolutely nothing to do with race ... but multi-variable equations," argues Goldstein. Meanwhile, the potential benefits of predictive policing are profound.*

Apparently it is easy to forget that variables often have semantics.

Recommend reading:

- `http://mathbabe.org/2014/08/25/gilian-tett-gets-it-very-wrong-on-racial-profiling/`
- `https://www.teamupturn.com/reports/2016/stuck-in-a-pattern`

Nature of the training data

Often, available data is inappropriate.

Suppose goal is to predict whether a suspect should be arrested.

- NYC "Stop, Question, and Frisk" data **only reflects actions of past NYPD officers**.

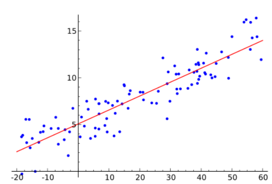
  Is this the desired behavior to learn?

- Even if we had "corrected" labels, the **distribution of suspects reflects who past officers chose to stop**: a textbook case of selection bias.

Simple fixes fail

Difficult even with "proper" data.

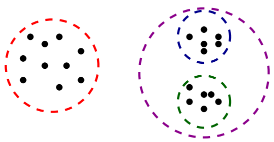
- Can't just remove protected features (e.g., gender), because other features could be correlated with it.



- Can't just tweak output to have "statistical parity"

$$\text{e.g., } P(f(X) = 1 \mid \text{gender} = 0) \;=\; P(f(X) = 1 \mid \text{gender} = 1)$$

  because disparity could manifest in subpopulations.



Example: Staples

WSJ observed that online price for an item depends on **how far you live from a brick-and-motar Staples store**.

- Doesn't explicitly look at your income.
- But where you live is probably correlated with your income.

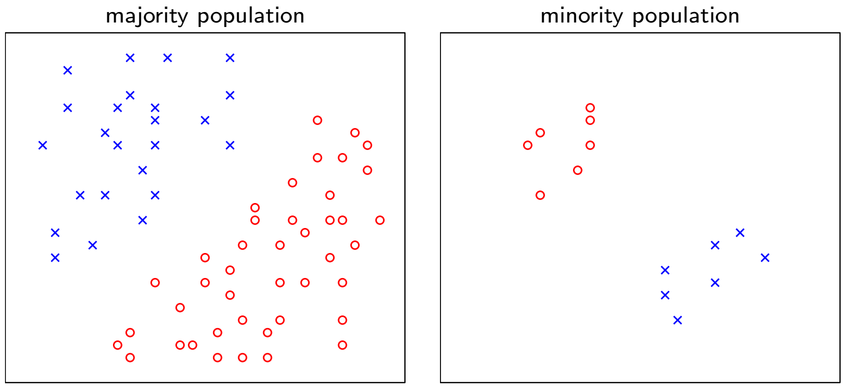
Moreover, **effect could manifest in subpopulations**, even if it doesn't manifest in overall population.

- For example, might see dependence between price and income in New York, but opposite dependence in Kansas.
  (Caveat: this isn't necessarily what actually happened.)
- At national level, dependence could appear to vanish!
  (Related to Simpson's paradox.)

Sample size difficulties

Often have more training data about "majority" populations, less about "minority" populations.

- **Service quality** of application may be higher (e.g., more accurate) for majority populations, lower (e.g., less accurate) for minority populations.
- **Extreme case**:



Fairness: summary

- Fairness, however you may define it, is difficult to assess and ensure in applications.
- Available data can be inappropriate.
- Some common "fixes" do not actually help.
- Intrinsic qualities (e.g., size) of even "proper" data can lead to undesirable outcomes.