CSE P 573: Guidelines for Deploying AI

Dan Weld/ University of Washington

[No slides taken from Dan Klein and Pieter Abbeel / CS188 Intro to AI at UC Berkeley – materials available at http://ai.berkeley.edu.]

Logistics

- Please fill out class survey!
  https://uw.iasystem.org/survey/205862
- Midterm
  - Mean 42.8
  - Max 54   (8 >= 50)
  - Min 23   (6 <= 35

2

Outline

- Biased Data
- Attacks on AI
- Maintenance Issues
- Intelligence in Interfaces

3

Your ML is Only as Good as the Training Data

Most training data is generated by humans

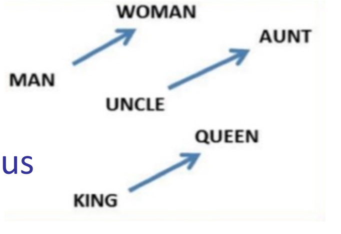
4

Science

"We show that standard machine learning can acquire stereotyped biases from textual data that reflect everyday human culture."

http://science.sciencemag.org/content/356/6334/183

Automating Sexism



- Word Embeddings
- Word2vec trained on 3M words from Google news corpus
- Allows analogical reasoning
- Used as features in machine translation, etc., etc.

man : king ↔ woman : queen
sister : woman ↔ brother : man

man : computer programmer ↔ woman : homemaker
man : doctor ↔ woman : nurse

https://arxiv.org/abs/1607.06520

Illustration credit: Abdullah Khan Zehady, Purdue

In fact…

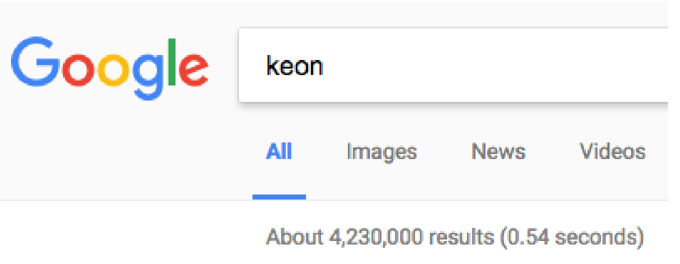
“Housecleaning Robot”

Google image search returns…

Not…

7

Racism in Search Engine Ad Placement



Searches of ‘black’ first names

Searches of ‘white’ first names

25% more likely to include ad for criminal-records background check

2013 study https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240

8

Predicting Criminal Conviction from Driver Lic. Photo

Convicted Criminals

Non-Criminals

?

- Convolutional neural network
- Trained on 1800 Chinese drivers license photos
- **90% accuracy**

https://arxiv.org/pdf/1611.04135.pdf

9

Should prison sentences be based on crimes that haven't been committed yet?

- US judges use proprietary ML to predict recidivism risk

NORTHPOINTE

- Much more likely to mistakenly flag black defendants
  - Even though race is not used as a feature

http://go.nature.com/29aznyw
https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing#.odaMKLgrw
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

10

What *is* Fair?

A	Protected attribute (*eg,* race)
X	Other attributes (*eg,* criminal record)
$Y' = f(X,A)$	Predicted to commit crime
Y	Will commit crime

- Fairness through unawareness

  $Y' = f(X)$ not $f(X, A)$ but Northpointe satisfied this!

- Demographic Parity

  $Y' \perp\!\!\!\perp A$ i.e. $P(Y'=1 \mid A=0)=P(Y'=1 \mid A=1)$

  Furthermore, if $Y \not\perp\!\!\!\perp A$, it rules out ideal predictor $Y'=Y$

C. Dwork et al. "Fairness through awareness" ACM ITCS, 214-226, 2012

What *is* Fair?

A	Protected attribute (*eg,* race)
X	Other attributes (*eg,* criminal record)
$Y' = f(X,A)$	Predicted to commit crime
Y	Will commit crime

- Calibration within groups

  $Y \perp\!\!\!\perp A \mid Y'$

  No incentive for judge to ask about A

- Equalized odds

  $Y' \perp\!\!\!\perp A \mid Y$ *i.e.* $\forall y$, $P(Y'=1 \mid A=0, Y=y) = P(Y'=1 \mid A=1, Y=y)$

  Same rate of false positives & negatives

- Can't achieve both!

  Unless $Y \perp\!\!\!\perp A$ or Y' perfectly $= Y$

J. Kleinberg et al "Inherent Trade-Offs in Fair Determination of Risk Score" arXiv:1609.05807v2

Guaranteeing Equal Odds

Given any predictor, Y'

Can create a new predictor satisfying equal odds

Linear program to find convex hull

Bayes-optimal ***computational affirmative action***

- Calibration within groups

  Y [redacted] A | Y'

  No incentive for judge to ask about A

- Equalized odds

  $Y' \perp\!\!\!\perp A \mid Y$ *i.e.* $\forall y,\ P(Y'=1 \mid A=0, Y=y) = P(Y'=1 \mid A=1, Y=y)$

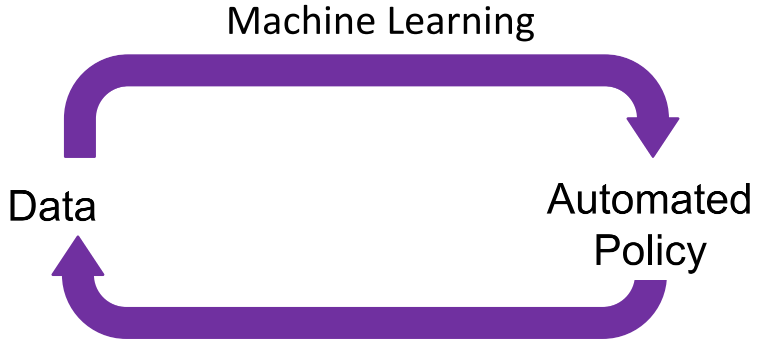
  Same rate of false positives & negatives

M. Hardt *et al* "Equality of Opportunity in Supervised Learning" arXiv:1610.02413v1

13

Important to get this Right!

Feedback Cycles

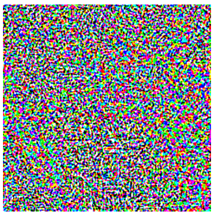


14

Attacks to Training Data

Adversarial Examples

+ 0.007 ×



=

57% Panda

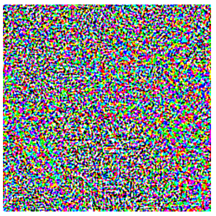
Access to NN parameters

"Explaining and harnessing adversarial examples," I. Goodfellow, J. Shlens & C. Szegedy, ICLR 2015

16

Adversarial Examples

+ 0.007 ×



=

57% Panda

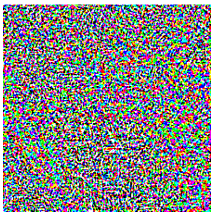
99.3% Gibbon

Access to NN parameters

“Explaining and harnessing adversarial examples,” I. Goodfellow, J. Shlens & C. Szegedy, ICLR 2015

Adversarial Examples

+ 0.007 ×



=

57% Panda

99.3% Gibbon

Only need x Queries to NN

Attack is robust to fractional changes in training data, NN structure

“Explaining and harnessing adversarial examples,” I. Goodfellow, J. Shlens & C. Szegedy, ICLR 2015

What’s This Sign Say?

Vision Algorithm Sees

https://arxiv.org/pdf/1707.08945.pdf

Maintenance

Machine Learning: The High Interest Credit Card of Technical Debt

https://ai.google/research/pubs/pub43146