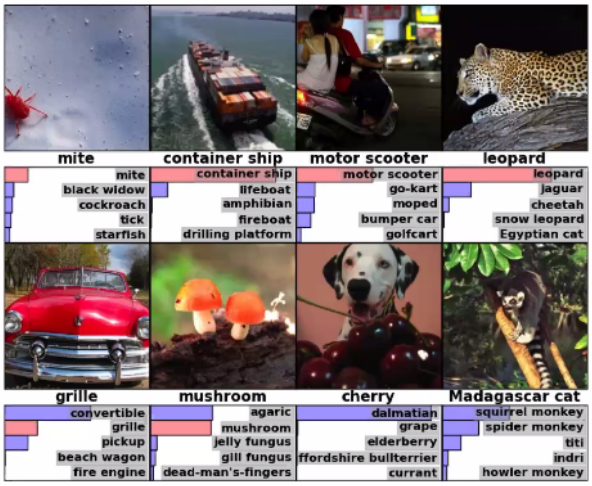
Day 3

Speaker: Prof. Chetan Arora, IITD

Title: Fairness in Visual Recognition Tasks.

Success of DNN Models in Computer Vision



Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton
ImageNet classification with deep convolutional neural networks. NIPS 2012

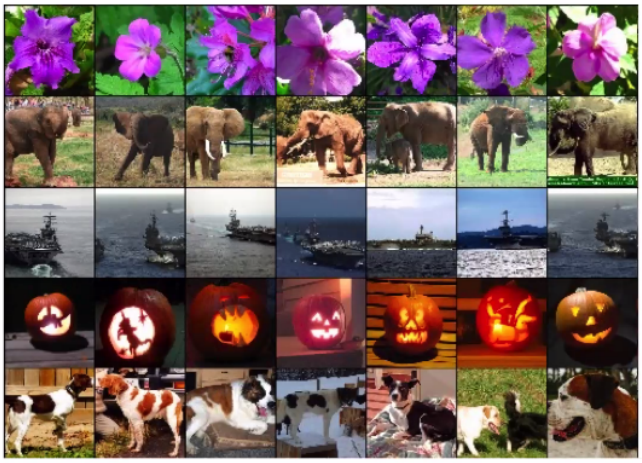
Breakthrough in DL → Accuracy ↑

* Mite → Mite → Black widow → Cockroach } Top 4 results similar

Eg. mite, leopard

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

DNNs: Generic Features!



Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton
ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Problem Formulation

Example: A company wants to hire a software engineer (SWE) and is going to advertise for the same. An ML system needs to predict which persons to show the advertisement based upon if he/she is currently a SWE.

- X: features of an individual (browsing history etc.)
- A: sensitive or protected attributes (gender etc.)
- $C = c(X, A)$: predicted score/class (show ad or not). Also denoted as $\hat{Y}$
- Y: target variable (whether the person is a SWE)

Notation: $\mathbb{P}_a\{E\} = \mathbb{P}\{E|A = a\}$

Euclidian Vectors]

ing Vectors

S

the

entire data set (in this case CIFAR 10, and ImageNet).

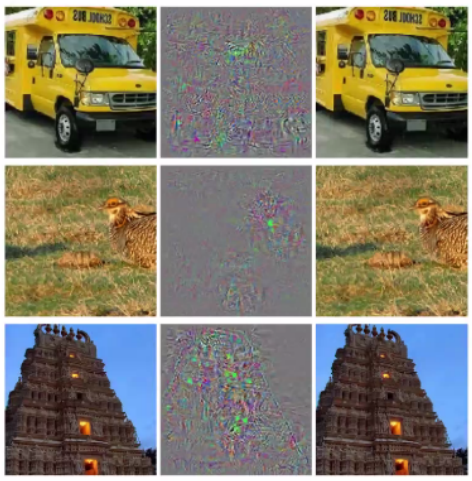
- State-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data.
- This phenomenon is qualitatively unaffected by explicit regularization and occurs even if we replace the true images by completely unstructured random noise.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals.
Understanding deep learning requires rethinking generalization.

* Labelling had not been significant. * Image memorization.
* Not affected by regularization

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

But How Little We Really "Understand" DNNs



Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus
Intriguing properties of neural networks. ICLR 2014

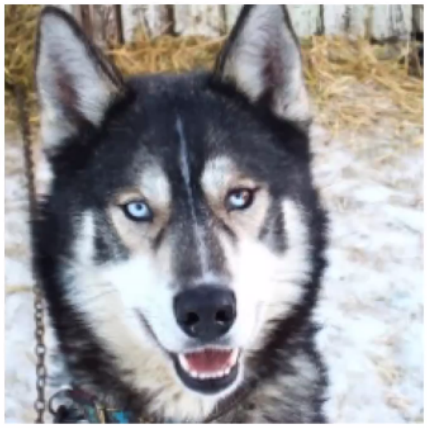
* Read Paper

Image → White Noise → ? (All predicted Ostrich)

[Same Distribution Dataset → Bad Prediction]

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Can We Trust Deep Network Predictions?



Husky classified as wolf

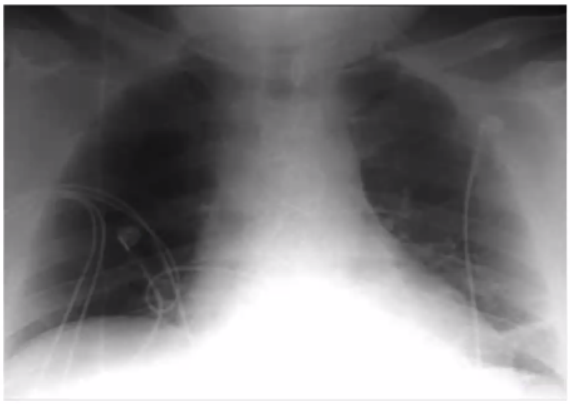
M. T. Ribeiro, S. Singh, and C. Guestrin.
Why should I trust you?: Explaining the predictions of any classifier.

→ Which part the n/w thinks are important for classification?
Here snow is considered a imp. part for determing a wolf.

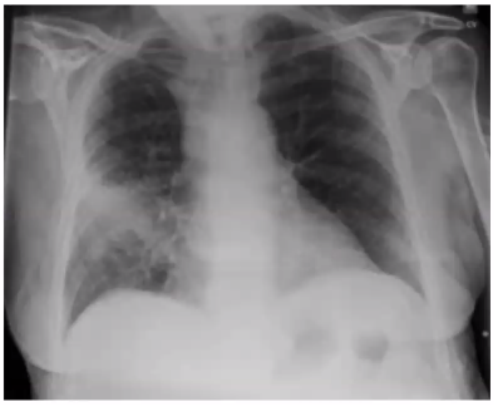
∴ Embedding may not say we learnt the semantic meaning.
but actually the similar distribution dataset
[unintended dataset]

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Can We Trust Deep Network Predictions?



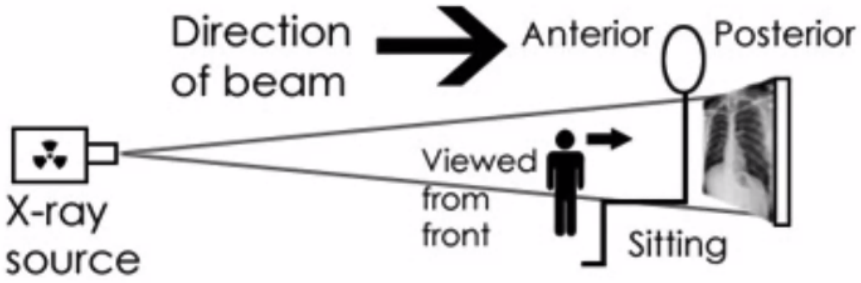
Classified Positive for Pneumonia



Classified Negative for Pneumonia

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

AP vs PA View



* Usually. Patient ⟶ PA View
  If P can't stand ⟶ AP View

Here, N/w may get bias with AP & PA View rather than Pneumonia or non-Pneumonia.

* Dataset always have the possibility of not learning its semantic meaning

→

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

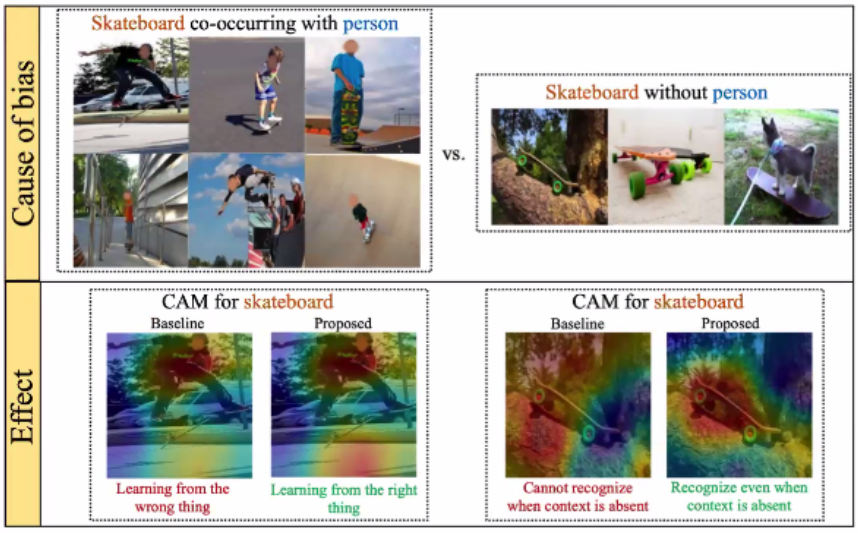
Systematic Bias: Appearance

https://objectnet.dev/

ImageNet → Organised Data. { ObjectNet → Wild Images of data.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Systematic Bias: Context



Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, Deepti Ghadiyaram
Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias. CVPR 2020

* ↑Accuracy ⟶ doesn't mean it is right for the right reason.

* Adversarial Attacks [To study the weakness of the NN]

To read ⟶ One Pixel Attack. Paper. *

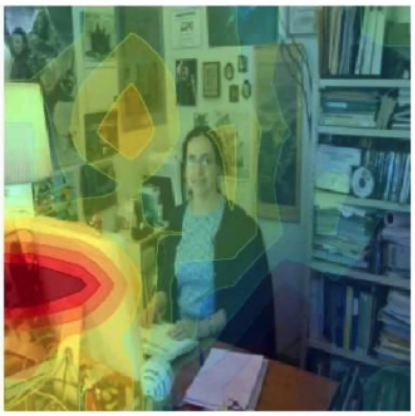
Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Systematic Bias → Fairness

https://www.gapminder.org/dollar-street

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Systematic Bias → Fairness



A man sitting at a desk with a laptop computer

Image Source: Women Also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018

For ImageNet Experiment:
- 256X256 Resolution → Both man/woman.
- 64X64 " → Only man.

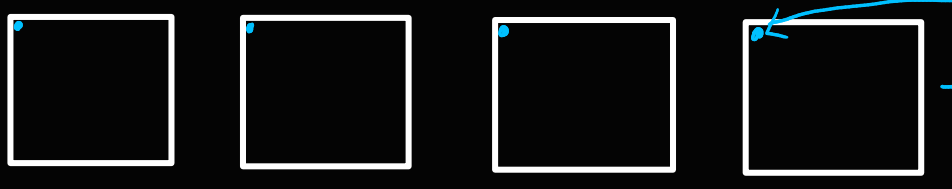
* If the n/w is biased against a particular group or class → *
* Systematic Bias or Fairness Issue in CV.

{ Size of the dataset is not a problem but the quality of the dataset. }
[To know if the dataset is having higher order correlation]

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness

- CV systems based on DNNs have been widely adopted in critical areas.
- Trained on large datasets that contain biases or spurious variations → Unintended Bias
- DNNs have been shown to inherit these biases and make decisions based on the sensitive, and protected attributes.



→ Blue Pixel adds Systematic Bias.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness in ML Systems

- Names that are associated with black people are found to be significantly more associated with unpleasant than with pleasant terms, compared to names associated with whites.
- The models learned on such text data for opinion or sentiment mining have a possibility of inheriting the prejudices reflected in the data.

A. Caliskan-Islam, J. J. Bryson, and A. Narayanan.
Semantics derived automatically from language corpora necessarily contain human biases.

Case Study:

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness in ML Systems

- Strong ethnic bias in COMPAS score: a predictive model for the "risk of crime recidivism"
- According to the score:
  - A black who did not re-offend were classified as high risk twice as much as whites who did not re-offend, and
  - White repeat offenders were classified as low risk twice as much as black repeat offenders

Offensive results of

http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

→ Not all Biases are Bad!

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Inductive Bias

Set of assumptions that the learner uses to predict outputs given inputs that it has not encountered

- **Minimum cross-validation error:** When trying to choose among hypotheses, select the hypothesis with the lowest cross-validation error.
- **Maximum margin:** when drawing a boundary between two classes, attempt to maximize the width of the boundary.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Undesirable Biases in ML Systems

- **Selection Bias**
  - Bias introduced by the selection of data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population
- **Reporting Bias**
  - Selective revealing or suppression of information. Authors under-reporting unexpected or undesirable experimental results

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Differentiation Vs Discrimination

- Differentiation is the basis of prediction in any ML system.
- Unjustified basis for differentiation is Discrimination
  - **Practical irrelevance**
    - Gender, Religion, Caste, Financial Status etc.
  - **Moral irrelevance**
    - Caste, Disability etc.

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Doctrines of Discrimination

1. Disparate Treatment → Treating 2 groups differently.

- **Formal**
  - Using protected attributes explicitly
- **Intentional**
  - Using proxies for the protected attributes

Eg. Indirect action.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Doctrines of Discrimination

2. Disparate Impact → Impact is different for 2 groups.

- **Unjustified**
  - Predicted outcome is significantly different for two groups
- **Avoidable**
  - There is way to achieve the same score with lesser disparate impact

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Does ML Prevent Disparate Treatment?

In favour of ML i.e no bias in the model.

- Automated decision leaves no scope of human discretion.
- Model learns what the data supports
- Protected attributes can be withheld

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

How Machines Learn to Discriminate

Biased sample selection caused discrimination.

- **Tainted examples**
  - Unreliable labels
  - Group A are tax evaders.
- **Proxies**
  - Considering features that are correlated with protected attribute

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

How Machines Learn to Discriminate

- **Limited features**
  - Features may be less informative or less reliably collected for certain parts of the population
  - A feature set that supports accurate predictions for the majority group may not for a minority group
- **Sample size disparity**
  - Objective functions are biased against the minority classes

MacBook → Rich
Correlation
may cause Bias.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Problem Formulation

Example: A company wants to hire a software engineer (SWE) and is going to advertise for the same. An ML system needs to predict which persons to show the advertisement based upon if he/she is currently a SWE.

- X: features of an individual (browsing history etc.)
- A: sensitive or protected attributes (gender etc.)
- $C = c(X, A)$: predicted score/class (show ad or not). Also denoted as $\hat{Y}$
- Y: target variable (whether the person is a SWE)

subset

Notation: $\mathbb{P}_a\{E\} = \mathbb{P}\{E|A = a\}$

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fundamental Criterion for Fairness

- **Independence:** $\hat{Y}$ independent of A
- **Separation:** $\hat{Y}$ independent of A conditional on Y
- **Sufficiency:** Y independent of A conditional on $\hat{Y}$

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

First criterion: Independence

- Require $\hat{Y}$ and A to be independent. Denoted as $\hat{Y} \perp A$
- That is, for all groups g_1, g_2 and all predictions c:

$$\mathbb{P}_{g_1}\{\hat{Y} = c\} = \mathbb{P}_{g_2}\{\hat{Y} = c\}$$

- When $\hat{Y}$ is a binary 0/1-variable:

$$\mathbb{P}_{g_1}\{\hat{Y} = 1\} = \mathbb{P}_{g_2}\{\hat{Y} = 1\},$$

- Also called **demographic parity**, or **statistical parity**

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

First criterion: Independence

Approximate versions

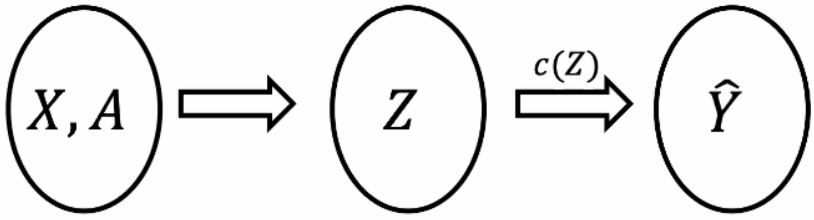
- $\frac{\mathbb{P}_{g_1}\{\hat{Y}=1\}}{\mathbb{P}_{g_2}\{\hat{Y}=1\}} \geq 1 - \epsilon$
- Also known in the legal context as "80% Rule". Declare unfair if the impact differs by more than 20%
- Additive version: $\left|\mathbb{P}_{g_1}\{\hat{Y} = 1\} - \mathbb{P}_{g_2}\{\hat{Y} = 1\}\right| \leq \epsilon$

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Independence

Representation Learning Approach

- Maximize *Mutual Information* between X and latent representation Z while minimizing the same between A and Z



$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} \mathbb{P}(x,y) \log \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)}$$

R. Zemel, Y. Wu, K. Swersky T. Pitassi, and C. Dwork
Learning Fair Representations. ICML 2013

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Learning Not to Learn

→ To read

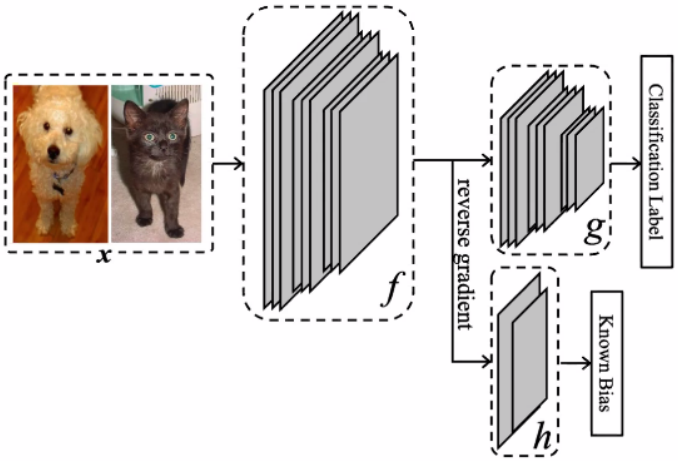


Image Source: Learning Not to Learn: Training Deep Neural Networks with Biased Data. CVPR 2019.

* To remove the class bias ↑ [To Be Updated]

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Problems with Independence based Fairness

- Ignores possible correlation between Y and A:
  - Recall our example problem: predict which persons to show the advertisement based upon if he/she is currently a SWE.
  - There could be less woman software engineers.
- Rules out perfect predictor $\hat{Y} = Y$.
- Permits laziness: Accept the qualified in one group, random people in other.
  - Allows to trade false negatives for false positives: Can create random false positives to make $\mathbb{P}_{g_1}\{\hat{Y} = 1\} = \mathbb{P}_{g_2}\{\hat{Y} = 1\}$.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fundamental Criterion for Fairness

- **Independence:** $\hat{Y}$independent of A
- **Separation:** $\hat{Y}$independent of A conditional on Y
- **Sufficiency:** Y independent of A conditional on $\hat{Y}$

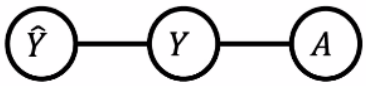
Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Second criterion: Separation

- Require $\hat{Y}$ and A to be independent, conditional on target variable Y
- Denoted as $\hat{Y} \perp A \mid Y$
- For all the groups g_1, g_2 and all values c and y:

$$\mathbb{P}_{g_1}\{\hat{Y} = c \mid Y = y\} = \mathbb{P}_{g_2}\{\hat{Y} = c \mid Y = y\}$$

- PGMs: Random variable $\hat{Y}$ **separated** from A if $\hat{Y} \perp A \mid Y$



Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Second criterion: Separation

$$\mathbb{P}_{g_1}\{\hat{Y} = c \mid Y = y\} = \mathbb{P}_{g_2}\{\hat{Y} = c \mid Y = y\}$$

- Also known as **Equalized Odds** OR **Equal Opportunity**
- For the outcome $Y = 1$, the constraint requires that $\hat{Y}$ has equal true positive rates across the two demographics A = a and A = b.
- For $Y = 0$, the constraint equalizes false positive rates.

+ Equalizing the error rates for all classes

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Desirable Properties of Separation

$$\mathbb{P}_a\{\hat{Y} = c \mid Y = y\} = \mathbb{P}_b\{\hat{Y} = c \mid Y = y\}$$

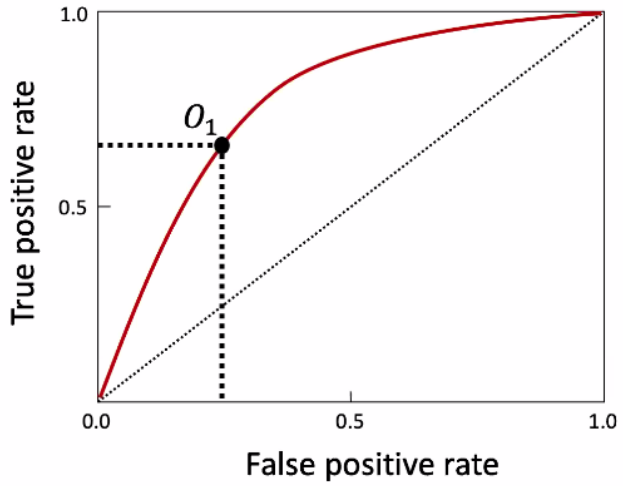
- Optimality compatibility: $\hat{Y} = Y$ is allowed
- Penalizes laziness: Incentive to reduce errors uniformly in all groups
  - Recall, none of the above is achieved by Independence.
- Assume Y is unbiased: Problematic with tainted and skewed samples

To read →

[To Be Updated]

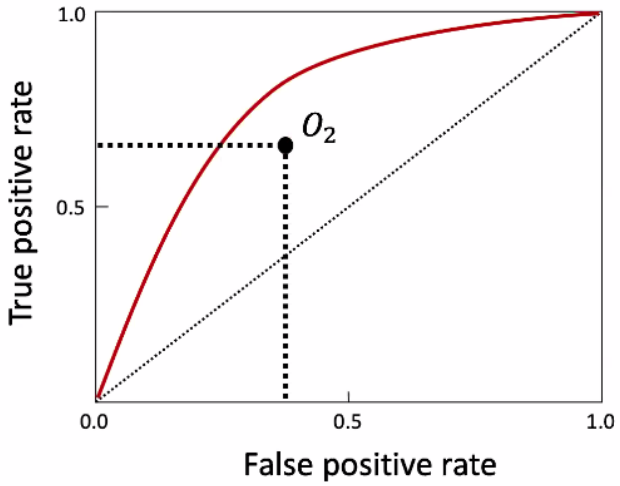
Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Separation



Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Separation

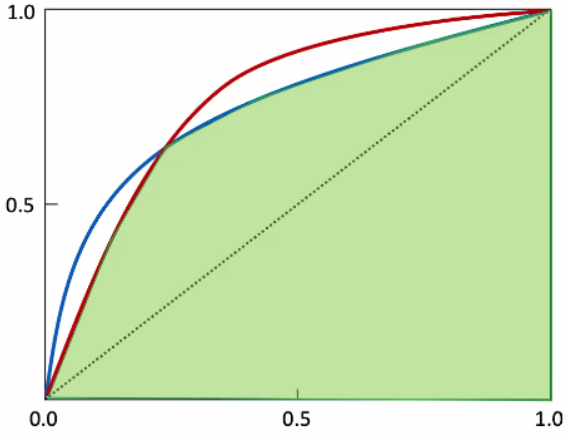


Create Random False Positives

Operating point O_1 can always be converted to O_2 by creating random false positives

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Separation



If Red and Blue are the curves corresponding to the two groups, then any point in the green region is feasible and satisfies equal opportunity

Moritz Hardt, Eric Price, Nathan Srebro
Equality of Opportunity in Supervised Learning. NIPS 2016

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fundamental Criterion for Fairness

- **Independence:** $\hat{Y}$ independent of A
- **Separation:** $\hat{Y}$ independent of A conditional on Y
- **Sufficiency:** Y independent of A conditional on $\hat{Y}$

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Third criterion: Sufficiency

- Random variable $\hat{Y}$ is sufficient for A, if $Y \perp A \mid \hat{Y}$

- For the purpose of predicting Y, we don't need to see A when we have $\hat{Y}$

$$\mathbb{P}_{g_1}\{Y = y \mid \hat{Y} = c\} = \mathbb{P}_{g_2}\{Y = y \mid \hat{Y} = c\}$$

- Also called **Predictive Rate Parity**

* Compas Company had PRP but it failed.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Third criterion: Sufficiency

$$Y \perp A \mid \hat{Y}$$

- For the purpose of predicting Y, we don't need to see A when we have $\hat{Y}$

$$\mathbb{P}_a\{Y = y \mid \hat{Y} = c\} = \mathbb{P}_b\{Y = y \mid \hat{Y} = c\}$$

Examples:

- For the purpose of showing the SWE advertisement, I don't need to know gender, if I already know the programming score.
- For the purpose of approving loan, I don't need to the know the race, when I know the credit rating

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Sufficiency

$$Y \perp A \mid \hat{Y}$$

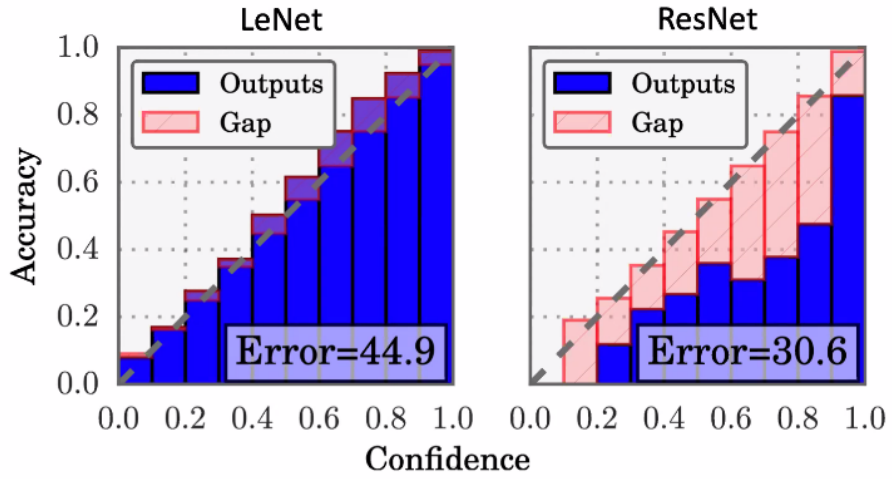
- Sufficiency satisfied by Bayes optimal score:
$$\hat{Y} = r(X, A) = \mathbb{E}[Y \mid X = x, A = a]$$

- Can be achieved using **calibration by group**: $\mathbb{P}\{Y = 1 \mid \hat{Y} = c, A = a\} = r$

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Neural Network Calibration



Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger
On Calibration of Modern Neural Networks. ICML 2017

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Sufficiency

$$\mathbb{P}_{g_1}\{Y = 1 \mid \hat{Y} = r\} = \mathbb{P}_{g_2}\{Y = 1 \mid \hat{Y} = r\}$$

- Formally, we say that a score R is **calibrated** if for all score values r in the support of R we have
$$\mathbb{P}\{Y = 1 \mid \hat{Y} = r\} = r$$

- Fairness by sufficiency can be achieved by group-wise calibrating the scores

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Sufficiency

$$\mathbb{P}_g\{Y = 1 \mid \hat{Y} = r\} = r$$

Plat Scaling:

- Given uncalibrated score $\hat{r}$, fit a sigmoid function: $S = \frac{1}{1+\exp(\alpha\hat{r}+\beta)}$
- Minimize log loss: $-\mathbb{E}[r \log S + (1-r)\log(1-S)]$

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Sufficiency

$$\mathbb{P}_g\{Y = 1 \mid \hat{Y} = r\} = r$$

Plat Scaling:

- Minimize log loss: $-\mathbb{E}[r \log S + (1-r)\log(1-S)]$, where $S = \frac{1}{1+\exp(\alpha\hat{r}+\beta)}$
- Objective function nudges S to take similar value as r(differentiate w.r.t. S)
- $\frac{r}{S} - \frac{1-r}{1-S} = 0. \Rightarrow \frac{r}{S} = \frac{1-r}{1-S} \Rightarrow r(1-S) = S(1-r). \Rightarrow r = S$

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness Tradeoffs

Any two of the three criteria we saw are mutually exclusive except in degenerate cases.

For example:

- If $A \not\perp Y$ and $\hat{Y} \not\perp Y$, then either independence or separation holds, not both.

* All the three fairness criteria cannot be treated the same time.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness Tradeoffs

- If $A \not\perp Y$ and $\hat{Y} \not\perp Y$, then either independence or separation holds, not both.
- Recall that Separation $\Rightarrow \hat{Y} \perp A \mid Y$, and Independence $\Rightarrow \hat{Y} \perp A$
- If separation holds then $\frac{\mathbb{P}\{\hat{Y},A\}}{\mathbb{P}\{Y\}} = \frac{\mathbb{P}\{\hat{Y}\}}{\mathbb{P}\{Y\}} \times \frac{\mathbb{P}\{A\}}{\mathbb{P}\{Y\}}$
- If independence holds then $\frac{\mathbb{P}\{\hat{Y},A\}}{\mathbb{P}\{Y\}} = \frac{\mathbb{P}\{\hat{Y}\}\mathbb{P}\{A\}}{\mathbb{P}\{Y\}}$.
- The two can not be true at the same time, unless $A \perp Y$ or $\hat{Y} \perp Y$

Case Study:

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

The COMPAS Debate

Bernard Parker, left, was rated high risk; Dylan Fugett, right, was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

- There's software used across the country to predict future criminals. And it's biased against blacks.

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica May 23, 2016

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Essence of COMPAS debate

ProPublica's main charge: Black defendants face higher false positive rate.

Northpointe's main defence: Scores are calibrated by group.

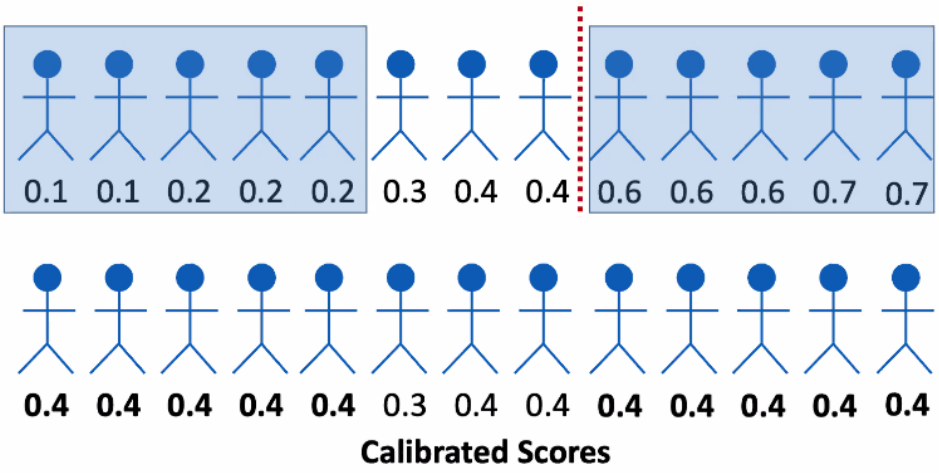
- Recall that sufficiency $\Rightarrow Y \perp A \mid R$, and can be achieved using **calibration by group**: $\mathbb{P}\{Y = 1 \mid R = r, A = a\} = r$ implemented by Platt Scaling

Though we discuss only calibration here, but neither calibration nor equality of false positive rates rule out blatantly unfair practices

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, Aziz Huq
Algorithmic decision making and the cost of fairness. KDD 2017

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Calibration is Insufficient



Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, Aziz Huq
Algorithmic decision making and the cost of fairness. KDD 2017

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Calibration is Insufficient: Another Example

	White	Black
0 previous convictions	5%	5%
1-2 previous convictions	20%	20%
3+ previous convictions	40%	40%
Average recidivism	20%	20%

Sam Corbett-Davies, Stanford University
Optimization and Fairness Symposium

{To be Updated}

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Calibration is Insufficient: Another Example

	White	Black
0 previous convictions		5%
1-2 previous convictions		20%
3+ previous convictions		40%
Average recidivism	20%	

Sam Corbett-Davies, Stanford University
Optimization and Fairness Symposium