Data Wrangling and Data Analysis

Imbalanced Data and Algorithmic Fairness

Hakim Qahtan

Department of Information and Computing Sciences
Utrecht University

Topics for Today

- Imbalanced Data
- Algorithmic Fairness
  - Fairness Measures
  - Bias Mitigation Algorithms

Reading Material & Exercises

- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35

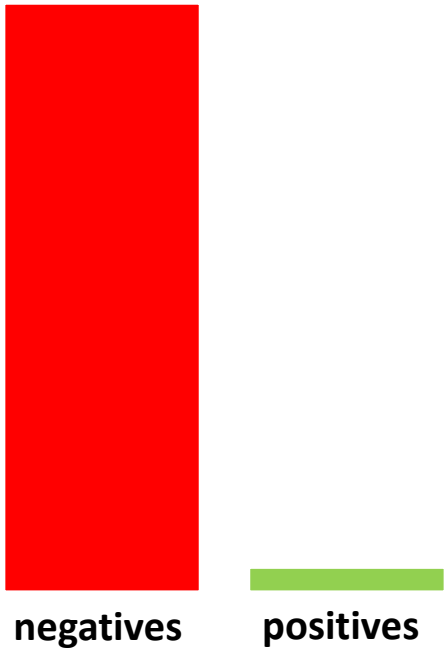
Imbalanced Data

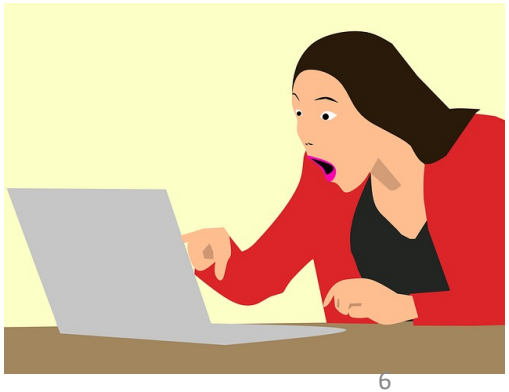
Imbalanced Data

- You trained a model to predict cancer from image data using CNN with dynamic kernel activations
- Your accuracy is **99.9%**

Imbalanced Data

- After plotting your class distribution
  - thousands of negative examples but just a couple of positives.





Imbalanced Data Problem

- Classifiers try to reduce the overall error (increase the accuracy) so they can be biased towards the majority class.
- Limitation of accuracy measure:
  - The problem of imbalanced classes
  - Consider a 2-class problem
    - Number of class 1 examples = 9990
    - Number of class 0 examples = 10
  - If the model predicts everything as class 1
    - $Accuracy = \frac{TP+TN}{P+N} = \frac{9990}{10000} = 99.9\%$

Your dataset is imbalanced.

Now what??

What Can We Do?

- Collect more data (difficult in many domains)
- Delete data from the majority class
- Create synthetic data
- Adapt your learning algorithm (cost sensitive classification)

Random Over/Under Sampling

- Random oversampling: randomly duplicate data points from the minority class.
- Random under-sampling: randomly delete data points from the majority class.
- Problems:
  - Loss of information (in the case of under sampling)
  - Overfitting and fixed boundaries (over sampling)

Create Synthetic Data

SMOTE

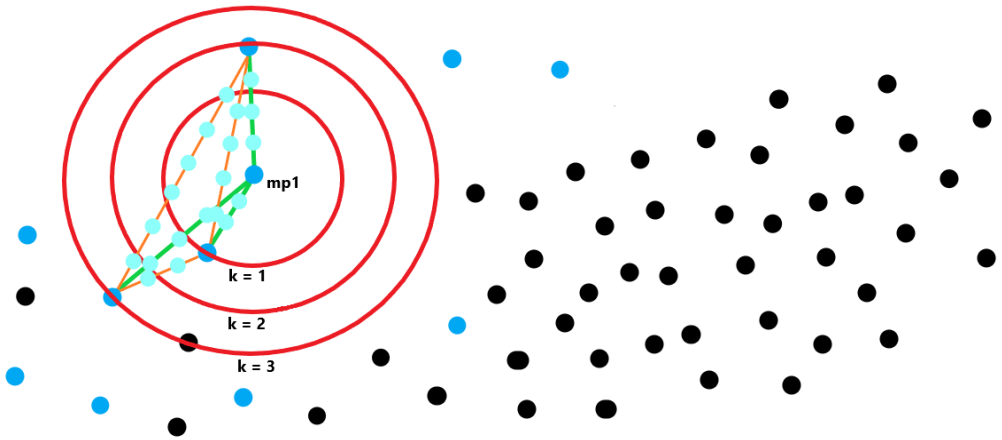
SMOTE

- Synthetic Minority Over-sampling Technique (Chawla).
- Creates new data points from the minority class.
- Operates in the feature space.

SMOTE

1. Take the difference between a sample point and one of its nearest neighbors.
2. Multiply the difference by a random number between 0 and 1 and add it to the feature vector.

This causes the selection of a random point along the line segment between two specific features.



SMOTE – Things to Consider

- Do not create synthetic points on the entire dataset before splitting into train/test sets.
- Generate the synthetic examples to enrich the training data only
- **Problem with Smote:** might introduce the artificial minority class examples too deeply in the majority class space.

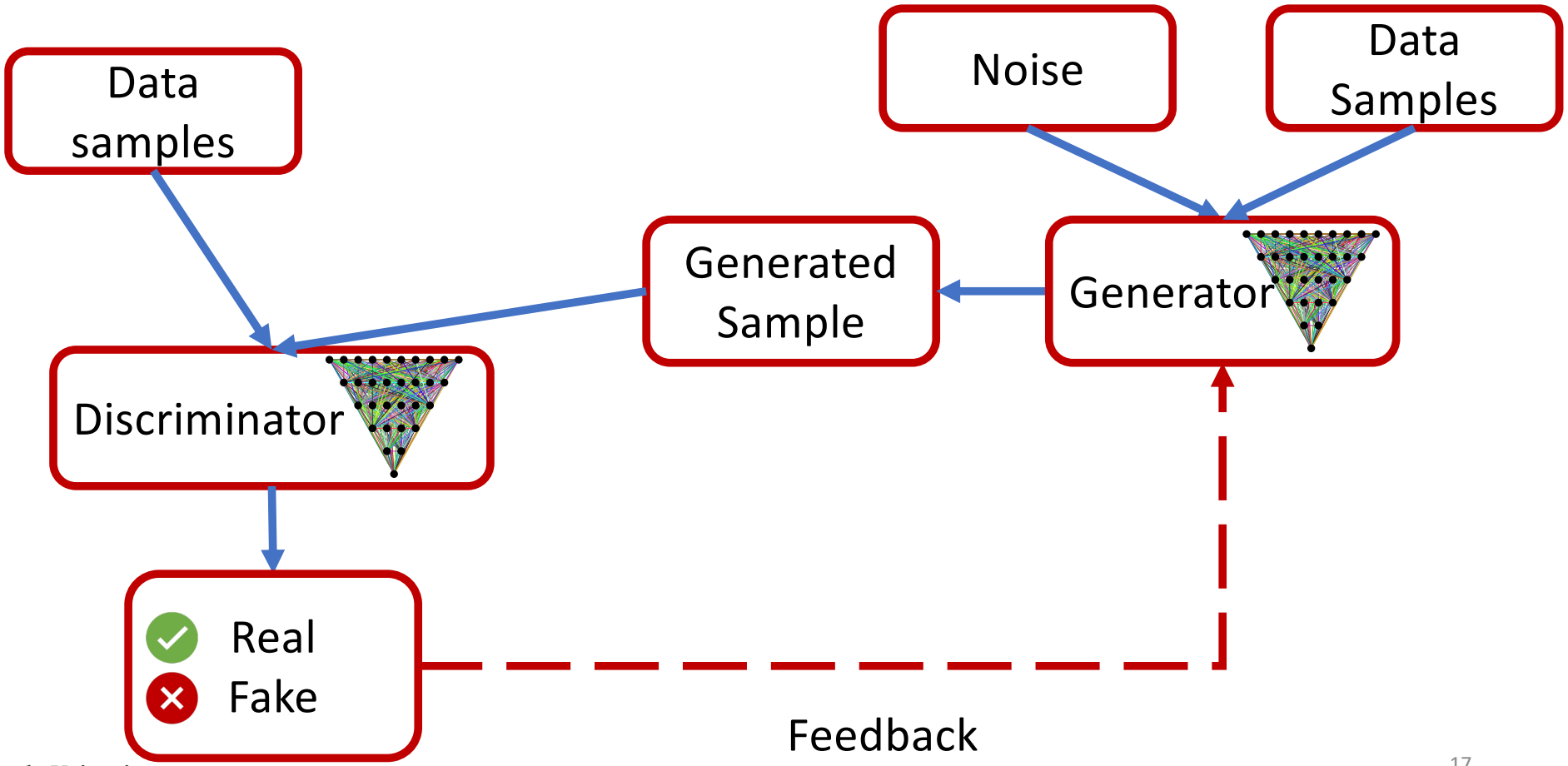
Create Synthetic Data

GANs

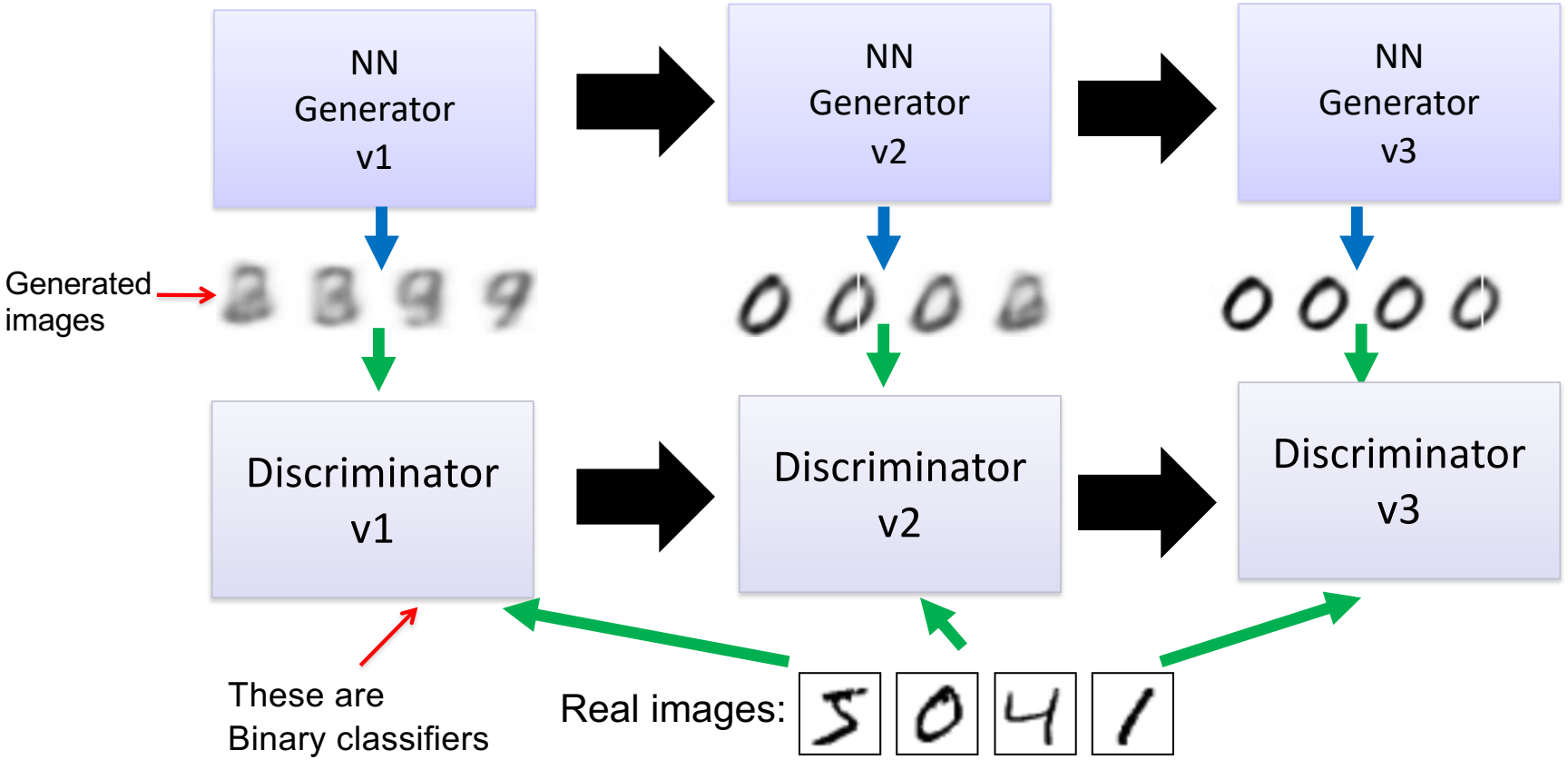
Generative Adversarial Networks (GANs)

- System of two neural networks competing against each other in a zero-sum game framework
- They were first introduced by Ian Goodfellow et al. in 2014
- Can learn to draw samples that are similar to the original examples

Generative Adversarial Networks (GANs)



Generative Adversarial Networks (GANs)



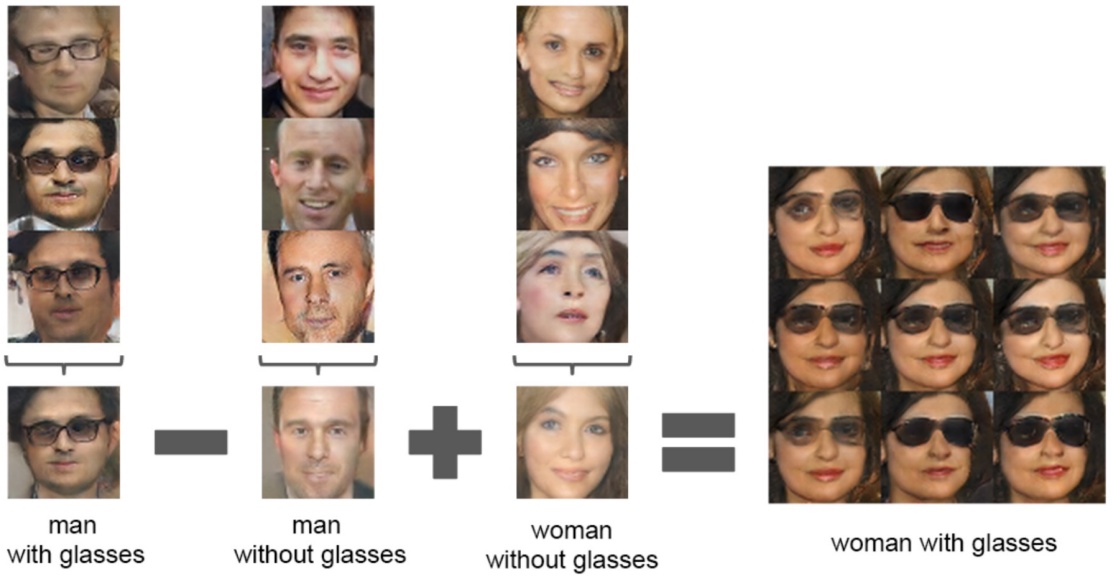
Generative Adversarial Networks (GANs)

- The **generator** tries to mimic examples from a training dataset, which is sampled from the true data distribution
  - It does so by transforming a random source of noise received as input into a synthetic sample

- The **discriminator** receives a sample, but it is not told where the sample comes from
  - Its job is to predict whether it is a data sample or a synthetic sample

How GANs Have Been Used?

- Have been used in generating images, videos, poems, some simple conversation
- Note, image processing is easy (all animals can do it), NLP is hard (only human can do it)
- This co-evolution approach might have far-reaching implications
  - This may hold the key to making computers a lot more intelligent


man
with glasses
man
without glasses
woman
without glasses
woman with glasses

How to Train GANs?

- Objective of generative network - increase the error rate of the discriminative network
  - Loss function
- Objective of discriminative network – decrease binary classification loss

Variations of the GANs

- Several new concepts built on top of GANs have been introduced –
- **InfoGAN** – Approximate the data distribution and learn interpretable, useful vector representations of data.
- **Conditional GANs (CTGAN)** - Able to generate samples taking into account external information (class label, text, another image). Force G to generate a particular type of output.

Failure Cases

- The discriminator becomes too strong too quickly and the generator ends up not learning anything
- The generator only learns very specific weaknesses of the discriminator
- The generator learns only a very small subset of the true data distribution

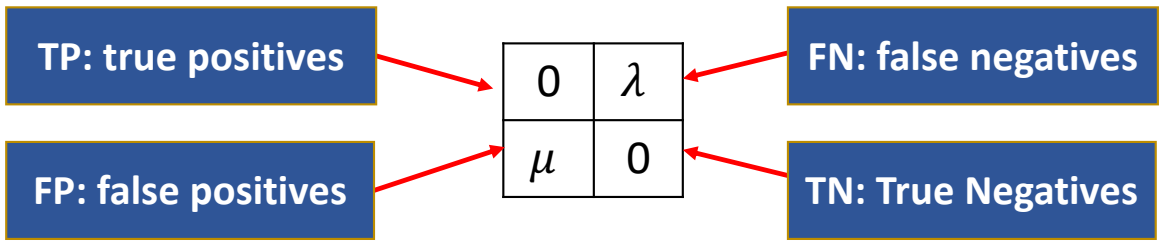
Major Problems

- Networks are difficult to converge
- Ideal goal – Generator and discriminator to reach some desired equilibrium but this is rare
- GANs are yet to converge on large problems (E.g. Imagenet).

Cost Sensitive Classification

Cost-sensitive classification

- Based on the classifier predicted probabilities.
- Binary traditional case: predict positive if probability is > 0.5
- Probability threshold can be changed using a cost matrix:



Classify as positive if: probability of positive > $\frac{\mu}{\mu+\lambda}$

Type I error
(false positive)

Type II error
(false negative)

Assign Class Weights

- In sklearn models, you can assign class weights.

fit() function in sklearn models has a class_weight parameter (see e.g., https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html).

Weights associated with classes in the form {class_label: weight}. If not given, all classes are supposed to have weight one.

e.g., class_weight = {0: 1., 1: 9.}

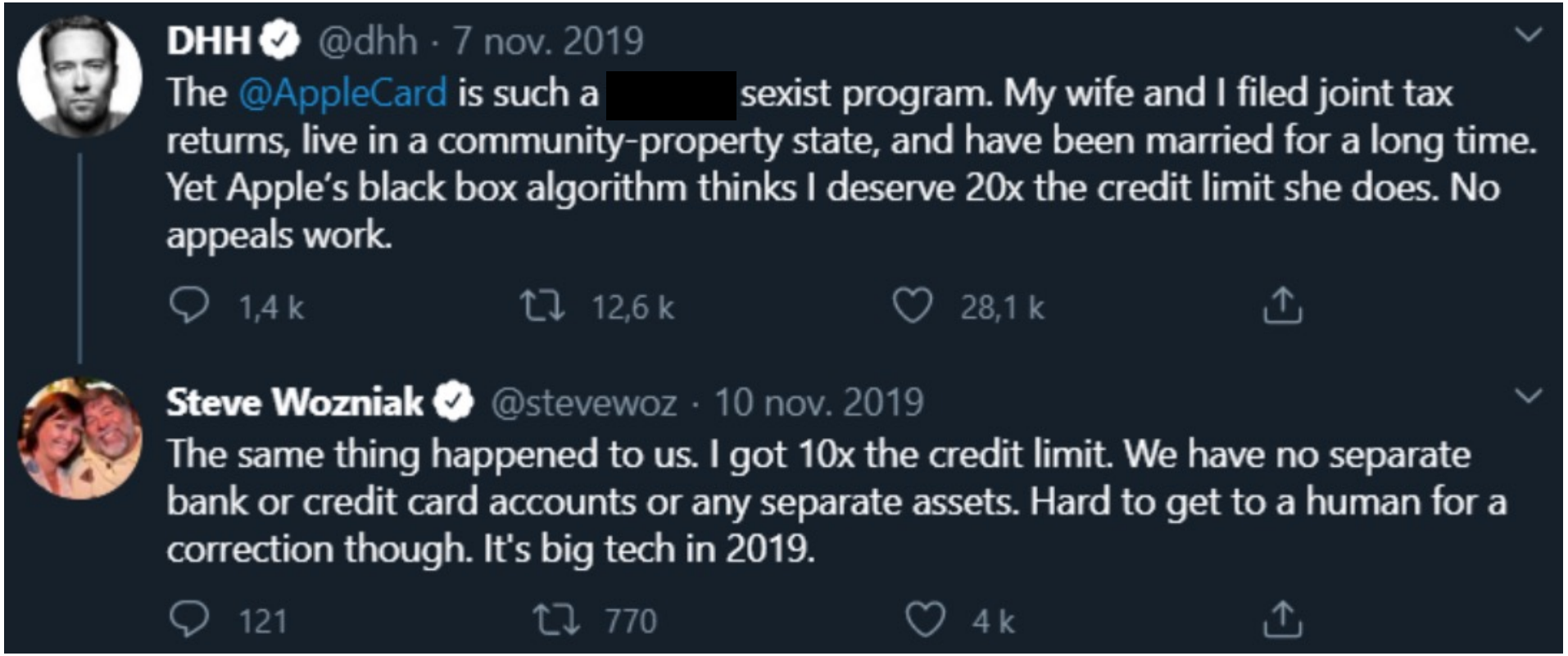
More Classification Evaluation Measures

- $precision(P) = TP/(TP + FP)$
- $recall(R) = TP/(TP + FN)$
- F-measure: $\mathrm{F}_1 = (2 \times R \times P)/(R + P)$
  - In general: $\mathrm{F}_\beta = (1 + \beta^2)\,(R \times P)/(\beta^2 \times P + R)$
- $sensitivity = (TP\ /\ (TP + FN))$
- $specificity = (TN\ /\ (FP + TN))$
- $Balanced\ Accuracy = (sensitivity + specificity)\ /\ (2)$

Bias and Fairness

Unfair Applications

Unfair Algorithms – Apple Card



Unfair Algorithms – Word Embeddings in NLP

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Bias and Fairness

Data Bias

Why ML Algorithms are Biased?

- The bias in the algorithms outcomes is not related to the way they are built

- The used dataset in training and constructing the ML models for prediction is the main reason
  - Problems comes from the bias in the training datasets

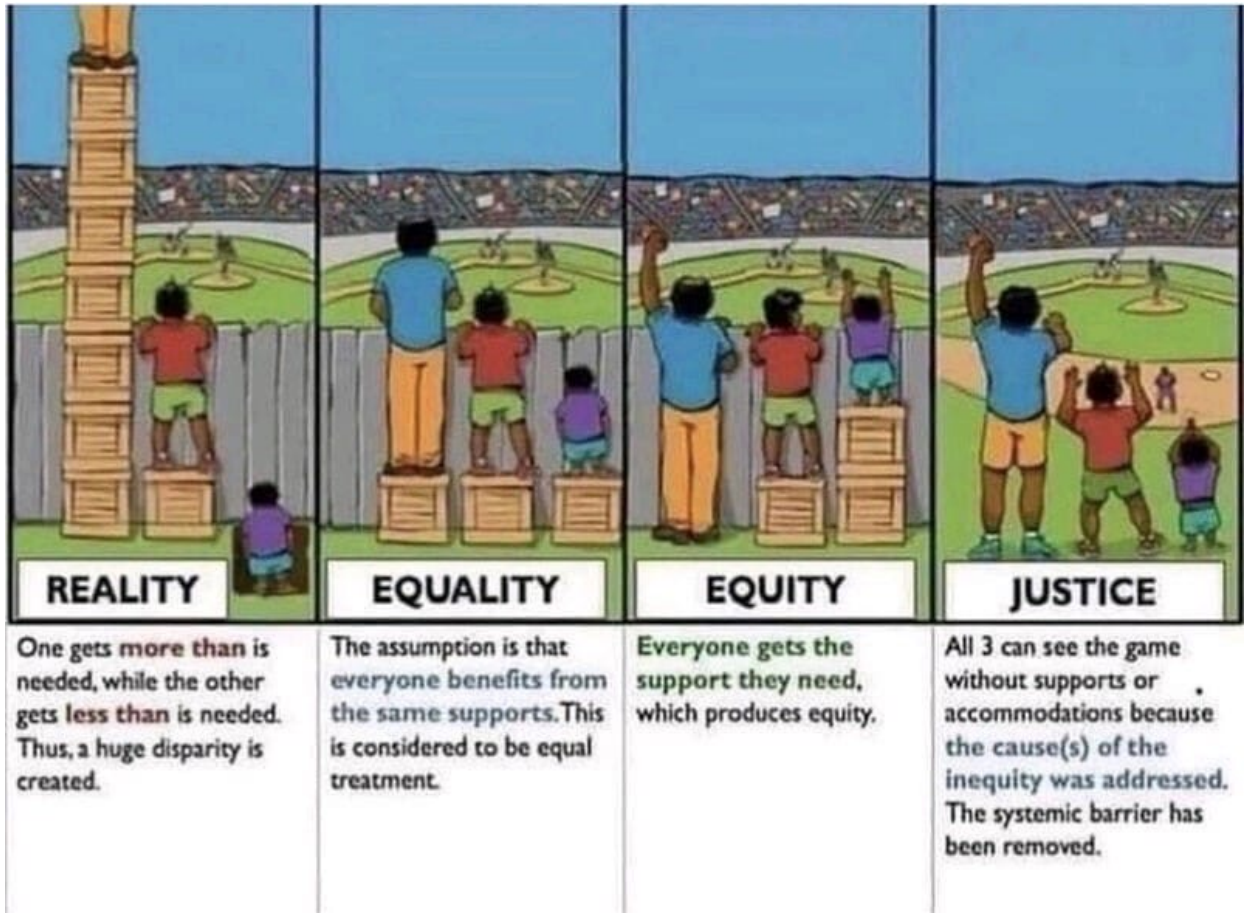
Bias in the Data

- Historical bias in the decision variable
- Limited / less informative features
- Biased data collection
- Imbalanced representation of different demographic groups

Bias and Fairness

Fairness Measures

What is Fairness?



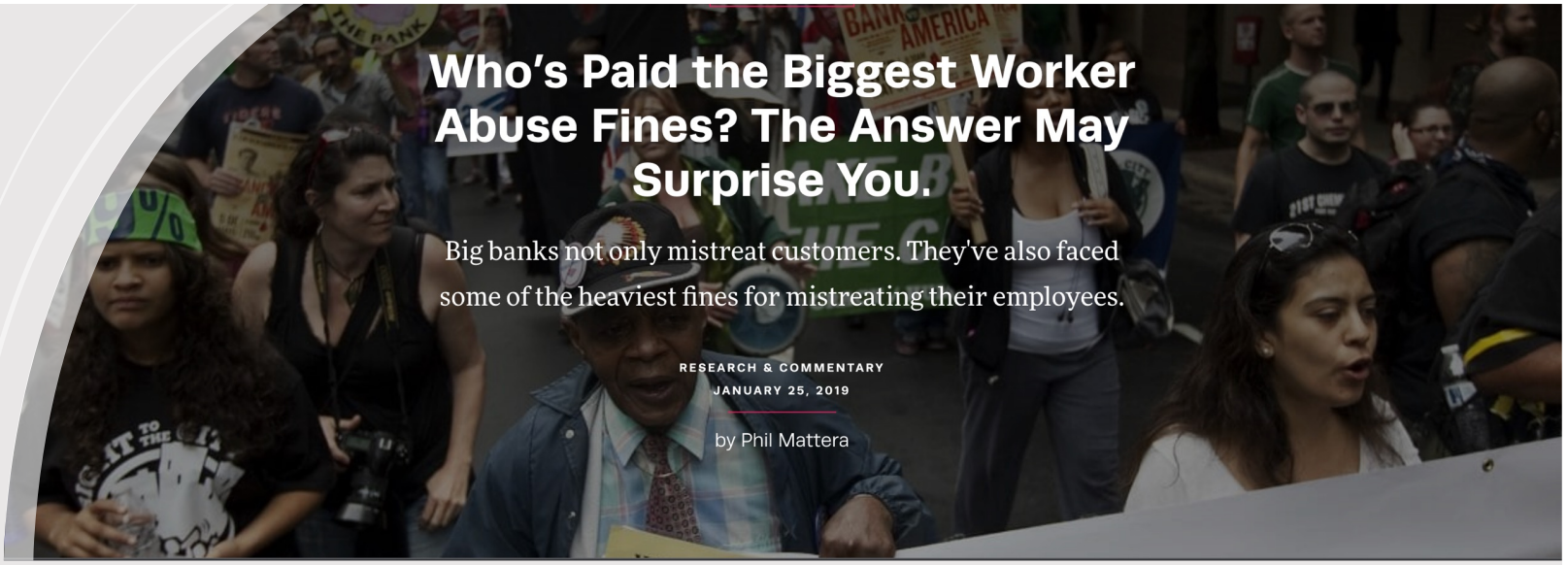
What is Fairness?

"Fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the decision-making context"

Bank of America fined $335m for minority discrimination

21 December 2011

https://www.bbc.com/news/business-16296146



Parent Companies with the Highest Disclosed Discrimination Penalties

Rank	Parent	Penalty Total	Cases
1	Bank of America	$210,296,593	8
2	Coca-Cola	$200,616,000	9
3	Novartis	$183,000,000	2
4	Morgan Stanley	$150,385,000	6
5	Abercrombie & Fitch	$90,115,600	4
6	FedEx	$80,035,138	15
7	Boeing	$79,935,059	7
8	Verizon Communications	$71,504,891	6
9	Wells Fargo	$68,099,000	5
10	SoftBank (parent of Sprint)	$62,852,756	3
11	Walmart	$52,888,021	27

https://inequality.org/research/penalties-workplace-abuse/

Protected Attributes

Protected characteristics

It is against the law to discriminate against someone because of a protected characteristic.

What are protected characteristics?

It is against the law to discriminate against someone because of:

- age
- disability
- gender reassignment
- marriage and civil partnership
- pregnancy and maternity
- race
- religion or belief
- sex
- sexual orientation

These are called protected characteristics.

Related content

Equality Act 2010

Advice and guidance

ADVICE AND GUIDANCE

Age discrimination

ADVICE AND GUIDANCE

Disability discrimination

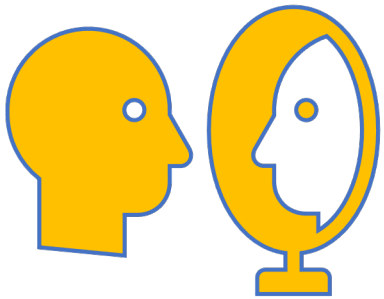
https://www.equalityhumanrights.com/equality/equality-act-2010/protected-characteristics

Fairness Measures

- Demographic groups are determined based on sensitive attribute
  - Also called protected attributes
- Privileged and unprivileged groups are determined based on the sensitive attributes and the decision label
  - Groups that receive undesirable decision more frequently are unprivileged
- Checking parity between the demographic groups
- Cannot always identify hidden unfairness

Individual Fairness Measures

- Individuals with similar features except the sensitive (protected) attributes must have the same/similar outcomes
- A similarity/distance measure is needed
- Requires strong assumptions regarding the relationship between features (variables) and the decision label

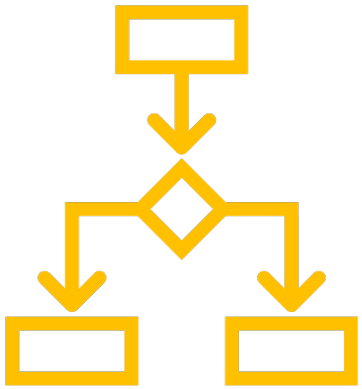


Group Fairness Measures

- Define multiple subgroups in a dataset
- Check parity between these subgroups
- A statistical constraint is needed:
  - E.g.: false positive rates

Causal Fairness Measures

- Causal relationships between the attributes and the outcome labels
- Sensitive attributes should not affect the outcome labels
- Identify “proxy” attributes
- Constructing the correct causal graph is a must
- Mostly a domain expert is needed



Fairness Measures in Details

- $D = \{X, S, Y\}$ is a dataset,
- X: the set of attributes that does not contain sensitive information regarding individuals
- S: is the set of sensitive attributes containing sensitive information
- $Y/\hat{Y} \in \{0,1\}$: the original/predicted class label of individuals, which indicates the decision outcome
- G/G': the values of unprivileged/privileged group

Fairness Measures – Demographic Parity (DP) Difference

- The instances in both unprivileged and privileged groups should have equal probability to receive positive outcomes

$$DP_{diff} = P\left[Y(\mathbf{x}) = 1 \mid S(\mathbf{x}) = G'\right] - P\left[Y(\mathbf{x}) = 1 \mid S(\mathbf{x}) = G\right] \approx 0.$$

- This measure takes values between 0 and 1 where 0 is the optimal

Fairness Measures – Disparate Impact (DI) Ratio

- The ratio between the probability of privileged and unprivileged groups getting positive or desired outcomes

$$DI(D) = \frac{P\left[Y(\mathbf{x}) = 1 | S(\mathbf{x}) = G\right]}{P\left[Y(\mathbf{x}) = 1 | S(\mathbf{x}) = G'\right]}.$$

- A dataset or a classifier is considered fair (by law) if its DI-ratio is between 0.8 and 1.25 (1 is the optimal)

Fairness Measures – Consistency

- An individual fairness measure determines how similar the labels are for the similar instances in a dataset based on the k-neighbors of the instance

$$Consistency = 1 - \frac{1}{|D|}\sum_{i=1}^{|D|}\left|\widehat{y}(\mathbf{x}_i) - \frac{1}{|kNN(\mathbf{x}_i)|}\sum_{\mathbf{x}_j \in kNN(\mathbf{x}_i)}\widehat{y}(\mathbf{x}_j)\right|$$

- This measure: takes values between 0 and 1 with 1 is the optimal
- **NOTE:** DP-diff., DI-ratio and consistency can be computed from the original dataset and the outcomes of a ML model

Fairness Measures – Equalized Odds (EO) Difference

- EO states that instances from privileged and unprivileged groups should have equal True Positive Rate (TPR) and False Positive Rate (FPR)

$$P_1 = P\left[\widehat{Y}(\mathbf{x}) = 1 \mid S(\mathbf{x}) = G', Y(\mathbf{x}) = 1\right],$$
$$P_2 = P\left[\widehat{Y}(\mathbf{x}) = 1 \mid S(\mathbf{x}) = G, Y(\mathbf{x}) = 1\right],$$
$$P_3 = P\left[\widehat{Y}(\mathbf{x}) = 1 \mid S(\mathbf{x}) = G', Y(\mathbf{x}) = 0\right],$$
$$P_4 = P\left[\widehat{Y}(\mathbf{x}) = 1 \mid S(\mathbf{x}) = G, Y(\mathbf{x}) = 0\right].$$

- For a classifier to be fair:

$$P_1 = P_2 \text{ and } P_3 = P_4$$

- i.e.

$$AEO_{diff} = \frac{(P_1 - P_2) + (P_3 - P_4)}{2}$$

Fairness Measures – Predictive Parity (PP)

- A classifier is fair in terms of predictive parity if the prob. that an example is positive in the original dataset given that it is predicted positive from both privileged and unprivileged groups is the same

$$P\left[Y(\mathbf{x}) = 1 \mid \widehat{Y}(\mathbf{x}) = 1, S(\mathbf{x}) = G\right] = P\left[Y(\mathbf{x}) = 1 \mid \widehat{Y}(\mathbf{x}) = 1, S(\mathbf{x}) = G'\right]$$

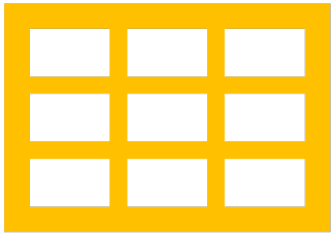
- This measure and the AEO measure can be applied on the outcome of an ML only (cannot be computed from the original dataset)

Bias and Fairness

Mitigation Algorithms

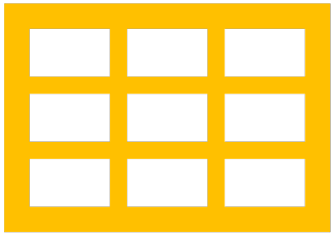
Mitigation Algorithms – Pre-Processing Techniques

- Pre-process the dataset only
- Different types of strategies:
  - Fairness through “unawareness”: deletes the sensitive attributes in a dataset
  - Preferential sampling (re-sampling): data objects are sampled with replacement
  - Massaging (Relabeling): changes the actual class label of some of the instances in the training set
  - Reweighing: assigns weights to each instance in the training set



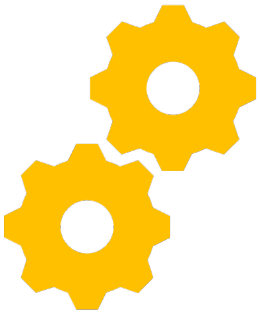
Mitigation Algorithms – Pre-Processing Techniques

- Example
- Learning Fair Representations (LFR) [ref]:
  - Intermediate representation of training set
  - Accurate representation but conceals information about sensitive attributes
  - In the final mapping, the class labels are changed to satisfy group and individual fairness



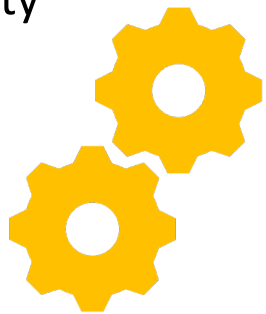
Mitigation Algorithms – In-Processing Techniques

- Adjust/tune the classification algorithm
- Applied during the model training
- Regularization of the model
- Dependent on the implemented classifier



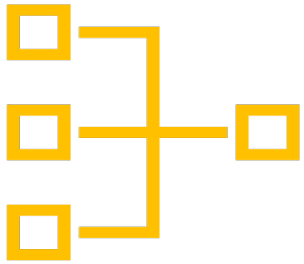
Mitigation Algorithms – In-Processing Techniques

- Example
- Adversarial Debiasing (Adv. Deb.): In-processing [ref]
  - An adversarial learning technique, one predictor and one adversary
  - Predictor predicts the class label, adversary predicts the sensitive attribute
  - Must satisfy demographic parity, equalized odds and equal opportunity



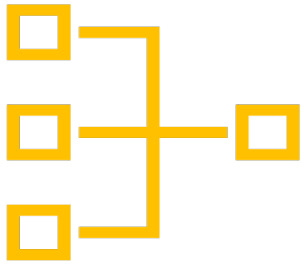
Mitigation Algorithms – Post-Processing Techniques

- Eliminate the discrimination from the final predictions
- Change the predicted outcomes of classifiers
- Changes are based on certain rules or constraints such as equalized odds.
- Thresholding the critical regions
  - E.g., change the samples that are predicted positive from the privileged group with prob. < 0.6 to negative.



Mitigation Algorithms – Post-Processing Techniques

- Example
- Calibrated Equalized Odds (Calibrated EO): Post Processing [ref]
  - Targets to satisfy both calibration and equalized odds
  - To achieve the goal, it changes some of the predicted class labels



Bias Detection Tools

- Only detect/quantify the amount of bias
- Audit focus
- Some of the developed tools:
  - FairML
  - Themis
  - What-If Tool (Google)

Bias Mitigation Tools

- Detect / quantify the amount of bias AND eliminate or mitigate it
- Internal / business focus
- Some of the developed tools:
  - Themis-ML
  - FairLearn (Microsoft)
  - AI Fairness 360 (IBM)

Wrap-Up

- We discussed
  - Bias and Imbalanced Data
  - Fairness measures
  - Bias Mitigation Algorithms