Lecture 3
Security and Ethical Aspects of Data

Amaya Nogales Gómez
amaya.nogales-gomez@univ-cotedazur.fr

MSc Data Science & Artificial Intelligence
Université Côte d'Azur

October 25, 2021

Tentative Content

1. Introduction
   - Preliminaries and Machine Learning basics
   - Motivation for ethics in Artificial Intelligence
2. Sources of unfairness
   - Bias in data
   - Algorithmic unfairness: Examples
3. Fairness criteria
   - Types of discrimination
   - Definitions of fairness
4. Algorithms and methods for fair ML
   - Pre-processing methods
   - In-processing methods
   - Post-processing methods
   - Legal and policy perspectives

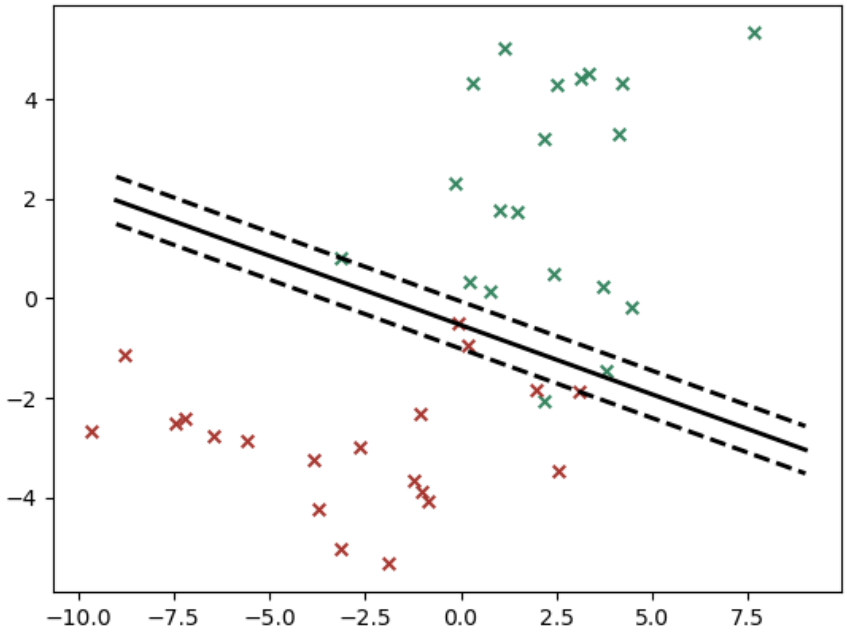
The SVM formulation

$$\min_{\omega, b, \xi} \frac{1}{2} \sum_{j=1}^{d} \omega_j^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, n$$
$$\omega \in \mathbb{R}^d$$
$$b \in \mathbb{R}$$
$$\xi_i \geq 0 \qquad \forall i = 1, \ldots, n.$$

Non-linearly separable data



- An object i will be correctly classified if $0 \leq \xi_i < 1$
- Misclassified if $\xi_i > 1$.
- In the case $\xi_i = 1$, we get a tie (objects coincide with the hyperplane).
- $\sum_{i=1}^{n} \xi_i$ is an upper bound of the number of misclassified objects.

Advertising

Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

- The authors explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields.
- This ad was explicitly intended to be gender-neutral in its delivery.
- **Fewer women saw the ad than men**.

Advertising

Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

- The authors explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields.
- This ad was explicitly intended to be gender-neutral in its delivery.
- **Fewer women saw the ad than men**.
- Why? Because younger women are a prized demographic, i.e., they are more expensive to show ads to.

`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260`

Machine Learning lifecycle

1. Data collection
2. Data preparation
3. Model development
4. Model evaluation
5. Model postprocessing
6. Model deployment

Bias in data

Historical bias

- It arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model.
- It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

Representation bias

- It arises while defining and sampling a development population.
- It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population.

Measurement Bias

- It arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities.
- The chosen set of features and labels may leave out important factors or introduce group or input-dependent noise that leads to differential performance.

Aggregation bias

- It arises during model construction, when distinct populations are inappropriately combined.
- In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups.

Evaluation bias

- It occurs during model iteration and evaluation.
- It can arise when the testing populations do not equally represent the various parts of the use population.
- Evaluation bias can also arise from the use of performance metrics that are not appropriate for the way in which the model will be used.

Deployment Bias

- It occurs after model deployment, when a system is used or interpreted in inappropriate ways.

Sensitive features

- In many classification tasks, the features X contain or implicitly encode sensitive characteristics of an individual.
- We let A to designate a binary variable that captures one sensitive characteristic.
- Different settings of A correspond to different groups of the population.
- This choice of notation is not meant to suggest that we can cleanly partition the set of features into two independent categories such as "neutral" and "sensitive".
- **The choice of sensitive attributes will generally have profound consequences as it decides which groups of the population we highlight, and what conclusions we draw from our study.**

Notation

- Dataset (X, A, Y)
- X: non-protected features
- A: protected attribute
- $Y \in \{-1, +1\}$ label, class membership
- $\hat{Y} = f(X, A)$: prediction

Disclaimer: abuse of notation

- feature $\equiv$ attribute $\equiv$ variable $\equiv$ characteristic
- protected $\equiv$ sensitive

Discrimitation Law: two doctrines

Disparate treatment

A decision making process suffers from disparate treatment if its decisions are partly based on the subject's sensitive attribute information.

Formal or intentional

Disparate impact

A decision making process suffers from disparate impact if its outcomes disproportionately hurt people with certain sensitive attribute values.

Unjustified or avoidable

- While it is desirable to design decision making systems free of disparate treatment as well as disparate impact, controlling for both forms of unfairness simultaneously is challenging.
- Avoid disparate treatment $\rightarrow$ disparate impact
- Avoid disparate impact $\rightarrow$ disparate treatment

Disparate Treatment

Formal

Explicitly considering the sensitive feature, even if it is relevant

Intentional

Purposefully attempting to discriminate without direct reference to sensitive feature

The $p\%$ rule

A decision boundary satisfies the "80%-rule" (or more generally the "$p\%$-rule"), if the ratio between the percentage of users with a particular sensitive attribute value having $\hat{Y} = +1$ and the percentage of users without that value having $\hat{Y} = +1$ is no less than 80:100 (p:100).
For a given binary sensitive attribute $A \in \{0, 1\}$, one can write the $p\%$-rule as:

$$\min\left(\frac{\mathcal{P}(\hat{Y} = +1|A = 1)}{\mathcal{P}(\hat{Y} = +1|A = 0)}, \frac{\mathcal{P}(\hat{Y} = +1|A = 0)}{\mathcal{P}(\hat{Y} = +1|A = 1)}\right) \geq \frac{p}{100}$$

Disparate impact

1. Accuser must first establish that decision procedure has a disparate impact, i.e., does it satisfy the 80%-rule?
2. Defendant must provide a justification for making decisions in this way. Is there a 'business necessity'? Is it 'job-related'?
3. Finally, accuser has the opportunity to show that defendant could achieve same goal using a different procedure that would result in a smaller disparity. Is there an 'alternative practice'?

Formal non-discrimination criteria

Independence

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

Separation

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

Sufficiency

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

Independence

The random variables $(A, \hat{Y})$ satisfy independence if $A \perp \hat{Y}$.

- Independence has been explored through many equivalent terms or variants, referred to as **demographic parity**, **statistical parity**, **group fairness**, **disparate impact** and others.
- In the case of binary classification, independence simplifies to the condition
$$P(\hat{Y} = +1|A = 0) = P(\hat{Y} = +1|A = 1).$$
- Thinking of the event $\hat{Y} = +1$ as "acceptance", the condition requires the acceptance rate to be the same in all groups.

Separation

Random variables $(\hat{Y}, A, Y)$ satisfy separation if $\hat{Y} \perp A \mid Y$.

- It acknowledges the sensitive characteristic may be correlated with the target variable.
- The separation criterion allows correlation between the score and the sensitive attribute to the extent that it is justified by the target variable.

Sufficiency

Random variables $(\hat{Y}, A, Y)$ satisfy sufficiency if $Y \perp A \mid \hat{Y}$.

- It requires a parity of positive/negative predictive values across all groups, i.e.,

$$\mathcal{P}\left(Y = +1 | \widehat{Y} = \hat{y}, A = 0\right) = \mathcal{P}\left(Y = +1 | \widehat{Y} = \hat{y}, A = 1\right), \forall \hat{y} \in \{+1, -1\}$$

Relationships between criteria: Independence vs Sufficiency

Proposition 1

Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.

Proof.(Proof by contradiction) By the contraction property for conditional independence,

$$A \perp \hat{Y} \text{ and } A \perp Y \mid \hat{Y} \rightarrow A \perp (Y, \hat{Y}) \rightarrow A \perp Y.$$

That is, $A \perp (Y, \hat{Y})$ means that A is independent of the pair of random variables $(Y, \hat{Y})$. And dropping $\hat{Y}$ cannot introduce a dependence between A and Y. □

Independence vs Separation

Proposition 2

Assume Y is binary, A is not independent of Y, and $\hat{Y}$ is not independent of Y. Then, independence and separation cannot both hold.

Proof. In order to prove it by contradiction, we need to prove that

$$A \perp \hat{Y} \text{ and } A \perp \hat{Y} \mid Y \rightarrow A \perp Y \text{ or } \hat{Y} \perp Y.$$

By the law of total probability,

$$\mathcal{P}(\hat{Y} = \hat{y}|A = a) = \sum_{y} \mathcal{P}(\hat{Y} = \hat{y}|A = a, Y = y)\mathcal{P}(Y = y|A = a)$$

Applying the assumption $A \perp \hat{Y}$ and $A \perp \hat{Y}|Y$, this equation simplifies to

Proof. (continuation)

$$\mathcal{P}(\hat{Y} = \hat{y}) = \sum_y \mathcal{P}(\hat{Y} = \hat{y}|Y = y)\mathcal{P}(Y = y|A = a) \tag{1}$$

Applied differently, the law of total probability states

$$\mathcal{P}(\hat{Y} = \hat{y}) = \sum_y \mathcal{P}(\hat{Y} = \hat{y}|Y = y)\mathcal{P}(Y = y) \tag{2}$$

Combining (1) and (2), we have

$$\sum_y \mathcal{P}(\hat{Y} = \hat{y}|Y = y)\mathcal{P}(Y = y) = \sum_y \mathcal{P}(\hat{Y} = \hat{y}|Y = y)\mathcal{P}(Y = y|A = a)$$

Replacing $p = \mathcal{P}(Y = 0), p_a = \mathcal{P}(Y = 0|A = a), \hat{y}_y = \mathcal{P}(\hat{Y} = \hat{y}|Y = y)$, above:

$$p\hat{y}_0 + (1-p)\hat{y}_1 = p_a\hat{y}_0 + (1-p_a)\hat{y}_1.$$

Equivalently, $p(\hat{y}_0 - \hat{y}_1) = p_a(\hat{y}_0 - \hat{y}_1)$. This equation only holds if $\hat{y}_0 = \hat{y}_1$, which implies $\hat{Y} \perp Y$ or if $p = p_a$ for all a, in which case $Y \perp A$. □

Fairness Through Unawareness

It implies that $f(X, A)$ does not use the value of A and is a legal requirement in many domains where processing sensitive information about individuals is forbidden in order to guarantee no disparate treatment. A predictor $\widehat{Y}$ satisfies unawareness if it does not use the protected attribute A, i.e.,

$$\mathcal{P}\left(\widehat{Y}|X, A\right) = \mathcal{P}\left(\widehat{Y}|X\right).$$

Demographic Parity or Statistical Parity

A stronger definition of fairness compared to unawareness is **demographic parity**. A predictor $\widehat{Y}$ satisfies demographic parity with respect to protected attribute A, if $\widehat{Y}$ is independent of A, i.e.,

$$\mathcal{P}\left(\widehat{Y}|A\right) = \mathcal{P}\left(\widehat{Y}\right).$$

or equivalently

$$\mathcal{P}\left(\widehat{Y}|A=0\right) = \mathcal{P}\left(\widehat{Y}|A=1\right).$$

The protected and unprotected groups should receive the same distribution of output values.

Demographic parity

$$\mathcal{P}\left(\hat{Y}|A=0\right)=\mathcal{P}\left(\hat{Y}|A=1\right).$$

Pros

- Legal support: the "four-fifth rule" or "80%-rule" states that the selection rate for any protected group should be no less than four-fifths of that for the non-protected group.
- If this rule is violated, justification must be provided.
- *"Business necessity means that using the procedure is essential to the safe and efficient operation of the business and there are no alternative procedures that are substantionally equally valid and would have less adverse impact."*

Demographic parity

$$\mathcal{P}\left(\hat{Y}|A=0\right)=\mathcal{P}\left(\hat{Y}|A=1\right).$$

Cons

- This definition ignores any possible correlation between Y and A and in particular excludes the perfect predictor $Y=\hat{Y}$ when base rates are different,
  i.e., $\mathcal{P}(Y=+1|A=0)\neq\mathcal{P}(Y=+1|A=1)$.
- The notion permits that we accept the qualified applicants in one demographic, but random individuals in another, so long as the percentages of acceptance match.

Equalized odds

A predictor $\widehat{Y}$ satisfies **equalized odds** with respect to protected attribute A and outcome Y, if $\widehat{Y}$ and A are independent conditional on Y.
In this case, the equalized odds are equivalent to:

$$\mathcal{P}\left(\widehat{Y} = +1 | A = 0, Y = y\right) = \mathcal{P}\left(\widehat{Y} = +1 | A = 1, Y = y\right), \quad \forall y \in \{-1, +1\}.$$

The protected and unprotected groups should have equal true positive and false positive rates.

Equalize odds

$$\mathcal{P}\left(\widehat{Y} = +1 | A = 0, Y = y\right) = \mathcal{P}\left(\widehat{Y} = +1 | A = 1, Y = y\right), \forall y \in \{-1, +1\}.$$

Pros

- Optimality compatibility: $\hat{Y} = Y$ is allowed.
- Penalizes laziness: it provides incentive to reduce errors uniformly in all groups.

Equal Opportunity

We say that a binary predictor $\widehat{Y}$ satisfies **equal opportunity** with respect to A and Y if

$$\mathcal{P}\left(\widehat{Y} = +1 | A = 0, Y = +1\right) = \mathcal{P}\left(\widehat{Y} = +1 | A = 1, Y = +1\right)$$

The protected and unprotected groups should have equal true positive rate.

Equal Opportunity

$$\mathcal{P}\left(\widehat{Y} = +1 | A = 0, Y = +1\right) = \mathcal{P}\left(\widehat{Y} = +1 | A = 1, Y = +1\right)$$

Pros

- It allows for stronger utility (accuracy).

Cons

- Weaker notion of non-discrimination: it may not help with different notion of bias, i.e. historical bias.

Predictive equality

We say that a binary predictor $\widehat{Y}$ satisfies **predictive equality** with respect to A and Y if

$$\mathcal{P}\left(\widehat{Y} = +1 | A = 0, Y = -1\right) = \mathcal{P}\left(\widehat{Y} = +1 | A = 1, Y = -1\right)$$

The protected and unprotected groups should have equal false positive rate.

Predictive Equality

$$\mathcal{P}\left(\widehat{Y} = +1 | A = 0, Y = -1\right) = \mathcal{P}\left(\widehat{Y} = +1 | A = 1, Y = -1\right)$$

Pros

- It allows for stronger utility (accuracy).

Cons

- Weaker notion of non-discrimination: it may not help with different notion of bias, i.e. historical bias.

Predictive Parity

We say that a binary predictor $\widehat{Y}$ satisfies **predictive parity** with respect to A and Y if

$$\mathcal{P}\left(Y = +1 | \widehat{Y} = +1, A = 0\right) = \mathcal{P}\left(Y = +1 | \widehat{Y} = +1, A = 1\right)$$

Predictive parity requires the same positive predictive value (i.e., precision) in both groups.

Predictive Parity

$$\mathcal{P}\left(Y = +1 | \widehat{Y} = +1, A = 0\right) = \mathcal{P}\left(Y = +1 | \widehat{Y} = +1, A = 1\right)$$

Pros

- Optimality compatibility: $\hat{Y} = Y$ satisfies predictive parity.
- Equal chances of success given acceptance.

Cons

- Same as for equal opportunity and predictive equality, it may not help "closing the gap" between protected and non-protected groups.

And even more...

Group fairness, individual fairness, fairness through awareness, treatment equality, test fairness, counterfactual fairness, conditional statistical parity, conditional use accuracy equality...

List of demographic fairness criteria

Criteria	Category	Relationship
Group fairness	Independence	Equivalent
Demographic parity	Independence	Equivalent
Equal opportunity	Separation	Relaxation
Equalized odds	Separation	Equivalent
Unawareness	Separation	Equivalent
Predictive equality	Separation	Relaxation
Predictive parity	Sufficiency	Relaxation

The impossibility theorem of fairness

Demographic Parity vs Predictive Parity

If A and Y are not independent, then either demographic parity or predictive parity holds.

Demographic Parity vs Equalized odds

If A is not independent of Y and $\hat{Y}$ is not independent of Y, then either demographic parity or equalized odds holds.

Fairness metrics

Equalized Odds

$$UNF_{EOdds} = |\mathcal{P}\left(\widehat{Y} = +1|Y = +1, A = 0\right) - \mathcal{P}\left(\widehat{Y} = +1|Y = +1, A = 1\right)|$$

$$+|\mathcal{P}\left(\widehat{Y} = +1|Y = -1, A = 0\right) - \mathcal{P}\left(\widehat{Y} = +1|Y = -1, A = 1\right)|.$$

Predictive Equality

$$UNF_{PE} = |\mathcal{P}\left(\widehat{Y} = +1|Y = -1, A = 0\right) - \mathcal{P}\left(\widehat{Y} = +1|Y = -1, A = 1\right)|.$$

Equal opportunity

$$UNF_{EOpp} = |\mathcal{P}\left(\widehat{Y} = +1|Y = +1, A = 0\right) - \mathcal{P}\left(\widehat{Y} = +1|Y = +1, A = 1\right)|.$$

Fairness metrics (continuation)

Demographic Parity

$$UNF_{DP} = |\mathcal{P}\left(\widehat{Y} = +1|A = 0\right) - \mathcal{P}\left(\widehat{Y} = +1|A = 1\right)|.$$

Predictive parity

$$UNF_{PP} = |\mathcal{P}\left(Y = +1|\widehat{Y} = +1, A = 0\right) - \mathcal{P}\left(Y = +1|\widehat{Y} = +1, A = 1\right)|.$$

Disparate Impact or The $p\%$-rule

$$UNF_{DI} = \min\left(\frac{\mathcal{P}(\hat{Y} = +1|A = 1)}{\mathcal{P}(\hat{Y} = +1|A = 0)}, \frac{\mathcal{P}(\hat{Y} = +1|A = 0)}{\mathcal{P}(\hat{Y} = +1|A = 1)}\right) \times 100.$$

Recommended bibliography

- Barocas, S. and Hardt, M. *NIPS 2017 Tutorial on Fairness in Machine Learning*. `https://mrtz.org/nips17/#/`
- Hardt, M. and Price, E. and Srebro, N., *Equality of Opportunity in Supervised Learning*. `https://arxiv.org/abs/1610.02413`