Personal data - Fairness

Exercise 1 - types of harm

Below are some cases involving unfavourable decisions that may occur in the context of decisions taken towards people. Give a more general/abstract description of the type of harm in each case.

1. Having a good job opportunity not displayed on one's professional social network
2. Being denied health insurance coverage
3. Need to go through more security checks than other groups of people to access a building or an app
4. Receive less accurate medical diagnosis because skin cancer AI was trained on light skin
5. A search engine that shows only men when searching for "CEO".
6. A speech-to-text videoconferencing systems that cannot transcribe when speaking with a regional accent
7. "forget" to include a given colleague to casual coffee break and lunch
8. A person of colour being followed by a security guard in a store

Answer:

1. allocative harm : being denied an opportunity
2. allocative harm : being denied a resource
3. quality of service : need to put more effort
4. quality of service : performance disparity
5. stereotyping : assigning a property (CEO) to some category (men) hence excluding women.
6. erasure : acting as if some people or cultures do not exist.
7. social isolation
8. dignity, humiliation : doesn't prevent you from doing something but sends a public signal that you are inherenly suspicious

Do these descriptions cover *intent* or *impact* ?

Which is best suited for legal regulation regarding automated decision-taking systems ?

Exercise 2 - protected and unprotected features

Draw a small example of tabular data that may be used by some bank to decide whether to accept a request for a loan. Lines are loan applicants and columns are applicant attributes. In your example, propose at least two attributes that you think could fairly be used to decide whether to accept a loan (unprotected attributes) and at least two attributes that should not be used (protected attributes). Values of individuals on protected attributes determine what are often called privileged and underprivileged groups. Include an intersectional situation. Give more protected attributes.

Answer:

columns :

- income (unprotected)
- on-going loans (unprotected)
- gender (protected)
- ethnic background (protected)
- include a column with decision

lines :

- privileged
- underprivileged
- can be privileged on some feature and underprivileged on some other (this multivariate analysis is intersectionality)

protected : age, disability, gender reassignment, marriage/partnership, pregnancy, color, nationality, origins, religion, sex, sexual orientation, discrimination by perception (being mistakely perceive as), discrimination by association to someone who has a protected characteristic (being denied a promotion because you have a disabled child).

public social situations : at work, school, services (shops, banks, hospitals, gov offices, buying or renting, public transport).

Exercise 3 - statistical parity

Statistical parity is a possible criterion for fairness. Statistical parity is obtained if the decision-taking process is such that p(d|a)= p(d), where d is the decision and is a protected attribute.

- Show that this is equivalent to independence between the decision and the attribute
- Show that is also equivalent to demanding that the share of the protected group benefiting from a favourable decision is that same as its share in the general population.

How can a lack of statistical parity be solved ? (application : bank loan application)

Answer:

- lower the success decision threshold for underpriviledged so that the percentage of people passing the test is the same for both groups, provided their core financial metrics are the same.

Exercice 4 - proxy variables

Latanya Sweeney is a famous researcher. One day she googled her name and noticed she was proposed some google sponsored links towards lists of inmates in the local prison (true). Can you explain this story ? (she did) Did she supply a protected attribute ? Imagine a scenario based of geolocation and delivery where discrimination takes place indirectly, with impact but with no intent.

Exercice 5 - two-stage discrimination

Imagine you are a recruiter in a two-stage recruitment process (first stage : filtering based on the resume, then interview for those whose resume was considered good enough). Let us consider there might be a legal control of statistical parity from the first stage. Let us also consider decisions in the interview phase are fair. How can the overall process nonetheless be made unfair ?

Exercice 6 - individual fairness

Instead of statistical parity (also known as *group fairness*), let us consider the following intuition : *similar people should get similar treatment* : could you refine this idea, so that it leads to a mechanism for detecting unfair situations ?

Answer: similar people as evaluated on unprotected attributes only !

Exercice 7 - loan repayment

Let us consider a population asks for a bank loan. A known subset of that population is given a loan by the bank. Among these people, a known subset repays back the loan, while another subset cannot repay back the loan. How can you detect possible discrimination ?

Answer:

1. The Confusion Matrix Approach

Since you know who actually paid back their loan, you can build a Confusion Matrix for each subpopulation. This allows you to see where the “errors” are happening. True Positives (TP): People who were predicted to pay back and actually did. False Negatives (FN): People who would have paid back but were denied the loan (This is where the “harm” usually lies). False Positives (FP): People who were given a loan but defaulted.

This is the most critical metric when you know repayment outcomes. It asks: Of the people who are actually “good” for the money (Y=1), do they have an equal chance of getting the loan regardless of their group?

Mathematical check: $Y\hat{}=$ loan approval, $Y=$pay back

$P(Y\hat{}=1|Y=1,G=A)=P(Y\hat{} =1|Y=1,G=B)$

If Group A has a much higher False Negative Rate than Group B, you have identified systemic discrimination against “qualified” applicants in Group A.

Extra-topic : performative prediction

1. If you try to predict the weather for tomorrow, may your prediction change the weather ?
2. If you predict traffic jams for a given location and time tomorrow, may your prediction affect the traffic jam and hence the prediction ? At what condition ?
3. Give 2 or 3 other examples, including one which related to classification

Answer: stock market : positive feedback ; public health : predicting flu : people wash their hands ; strategic classification : a spam filter predicts spams : spammer see what gets blocked and change their mail tricks - data points can be seen as rational agents who have a preference for certain labels