Some exercises on personal data

Fairness in decisions involving personal data

Exercise 1

Below are some examples of unfavourable decisions that may occur in the context of automated decision systems. Describe in a more general/abstract way the type of harm in each case. Find at least one example in each type of harm.

Category 1 not being admitted into a school to which one applied (ex1), right to student financial allowance rejected (ex2)
Category 2 : having a good job opportunity announcement not being displayed on one's professional social network
Category 3 : having to wait a long time for a taxi in some areas of town
Category 4 : having one's ID card randomly controlled in the street more often than the average person

Exercise 2

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 example values for four fictitious people : two for which a positive loan decision would be taken, two for which a negative loan decision would be taken. In both positive and negative cases, each time incluse a case which you consider as fair and a case you consider as unfair)

Exercise 3

Consider the two following cases :

- the owner of red cars pay high insurance premium that the owner of cars of other colours
- a small company prefers to hire a man than a woman despite equivalent qualifications

Show that in both cases, the decision is based on stereotyping, i.e. assigning to all people in a category qualities (or lack of qualities) assigned globally to the category.

Which of these cases is more likely to result from machine learning or from personal prejudice ?
Is the case of the car insurance not statistically sound (in terms of cause or correlation) ?
If we introduce a black box in the car that measures the way it is driven, what personal data can be solved and what personal data problem is introduced ?
What is the abstraction of introducing a black box ?

Exercise 4

Bureaucracy consists in applying systematic procedures and rules for taking decisions.

1) Why is this good ?
2) What is the relation to automatic decision taking ?
3) Imagine some bureaucracy rejects your application without explanation and refuse to give you an explanation. It is a matter of discrimination ? Is there a difference between a market-based bureaucracy and a monopolistic, state-based bureaucracy ?

Exercise 5

In Boston, it was noted some delivery company delivered much less « one day delivery » in a given central district of the city, where black people where overrepresented, compared to the whole city. Do you think they necessarily discriminated against black people ? What are the unprotected and protected attributes in this scenario ? If we suppress the protected attribute from the table, does it prevent discrimination ? In your loan application, scenario can you think of some unprotected attributed that could be correlated to the protected attributed and some that may be uncorrelated to protected attributes ?

Exercise 6

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.

Given a machine learning system that takes automatic decisions on loan request, how can we make it fair, according to the statistical parity criterion ? On one or two practical examples you may find, is this is really satisfactory ?

Imagine you are a recruiter in a two-stage recruitment process (first resume, then oral interview for selected resumes). There is a legal control of statistical control of parity based on the acceptance of resumes. Let us consider fair decisions in the interview process. How can the overall process be made unfair ? (on purpose or not)

Coming back to the loan application problem, let us now assume that we have a first attribute for past requests : whether the person has been able to replay her loan or not. Can we now measure true positive decisions and false positive decisions ? What about true negative and false negative cases ? Can this improve the situation with regard to the previous recruitment scenario, especially if we want to make a machine-learning based decision technique ?

The previous questions addressed *group fairness*. Let us now consider an alternative criterion : individual fairness.
Does it make sense to consider that a decision is fair if, for two similar individual individuals, the same decision is taken ?

Who wants to know more can find a nice bibliography (https://fairmlclass.github.io/) and a good free book (https://fairmlbook.org/) by Moritz Hardt (Max Planck Institut/Berkeley) and the AI Fairness 360 IBM software

Factorization of the user-item matrix

Let us consider a matrix M indicating numerical evaluations provided by 100 people on 1000 items. As usual, only a small proportion of matrix entries are filled with an evaluation.

Instead of using similarities between lines or between columns, this exercise examines the other main technique for recommendation : matrix factorization. In the fast few years, many methods and software tools (in python, C++, R, Matlab, Java,…) have been developed for factorizing a matrix M as the product of two matrices U and V. This method is known as Non-Negative Matrix Factorization and belongs to larger family of matrix factorization techniques, widely used in data mining (in particular text mining), signal analysis and compression.

Remarks :

- we don't usually achieve M=U.V but M is only approximately equal to U.V, the closer the better.
- the factorization technique is able to cope with empty entries in M (i.e. M is a sparse matrix). In this case, comparison of M and U.V is only done where an entry for M is available. Yet all entries of U and V are filled.
- This factorization is not something you could easily calculate manually, like a matrix product.

Write down the factorization. A new dimension appears. This new dimension is typically chosen to be much smaller than the dimensions of M (say, 10).

How can we use matrix factorization to make prediction of missing ratings in M ?
How many entries are there in U and V, how does it compare to the number of entries in M ?
How could you interpret the line-column scalar product involved in computing each entry of U.V ?
How could you interpret the new dimension introduced in the factorization process ?
How could you interpret the matrix U.transposed(U) and the matrix V.transposed(V) ?

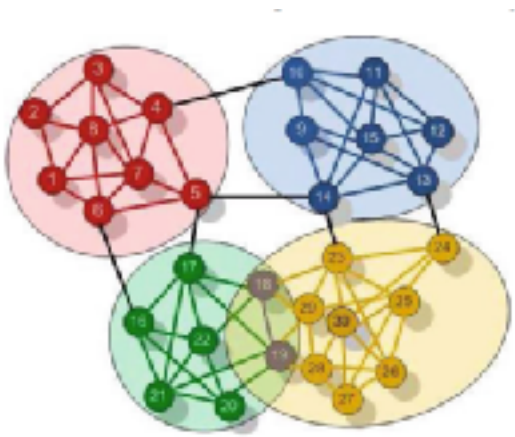
How does the concept of co-occurrence (« people who buy X generally also buy Y ») appear in matrix factorization ?

When describing the similarity between using users decomposed on topics, do we face the same spareness issue as when measuring the similarity between users in k-nearest neighbours technique ?

What is the relation between matrix factorization and multi-layer perceptron neural networks ? As a consequence, what would you expect (hope) from using neural networks ?

Factorization and social graph

This exercise addresses the graph of relations between people on a social network (undirected edges, eg people linked are symmetrical friends).



Consider the adjacency matrix A of the graph.
Interpret the values in A.transpose(A). Because A is symmetric, it can be factorized, with approximation as previously, as B.transpose(B). The new dimension introduced is generally significantly smaller than the number of lines in A. Write down this factorization. What interpretation could you give to the lines and columns of B ? What interpretation to the approximate reconstruction of A as the product B.transpose(B) ? Can a person belong to several communities at the same time ? How Could you use use this to make link prediction in the social network (i.e. predict that two people are likely to know each other (or to have common characteristics or interest, as far as the social graph can tell, although they are currently not linked) ?

Example with 30 people and 4 communities.
[fig. by Wang et al. , Data Mining Knowledge Discovery 2010]

Performative prediction

What if traffic jams predicted by google maps were not made public by google ?
What happens if they were suddenly they are made public ?

How does this relate to the fact that higher YouTube ranking tends to favor video viewing duration, which in turns favors high YouTube ranking ? What is the difference with the previous case ?
It is loop process desirable for Youtube ?