Some exercises on personal data

Fairness in decisions involving personal data

Identify various kinds of harms (indicate generic terms and examples) that can affect individuals, against which unfavourable decisions may be taken by automatic consequential decision-taking systems.

Draw a small example of tabular data that may be used by some bank to decide whether to accept a loan request. In your example, include attributes that you think could fairly be used to decide whether to accept a loan (unprotected attributes), and attributes that should not be used (protected attributes). The latter type defines what are often called privileged and underprivileged groups.

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 ?

Statistical parity is a possible criterion for fairness. Statistical parity is obtained if the decision-taking process is such that $p(d \mid a) = p(d)$, 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 sens 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).

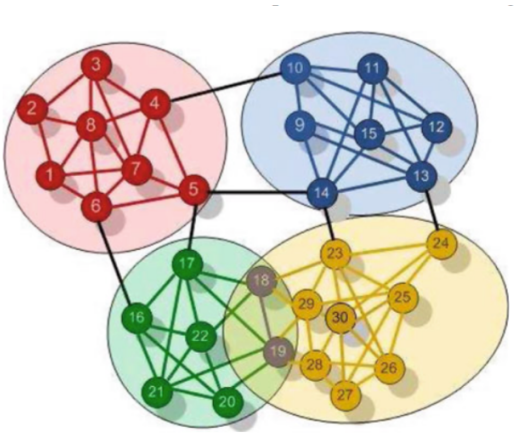
How can we use matrix factorization to make prediction of missing ratings in M ?
How many entries are there in U and V, how does it compare to the number of entries in M ?
How could you interpret the line-column scalar product involved in computing each entry of U.V ?
How could you interpret the new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

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).

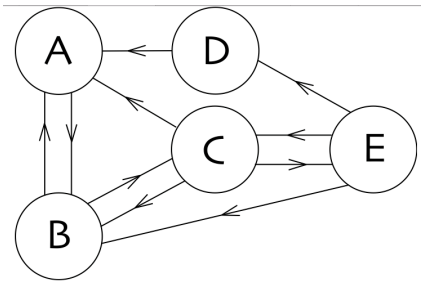


Example with 30 people and 4 communities.
[fig. by Wang et al., Data Mining Knowledge Discovery 2010]

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) ?

Estimating authority

This exercise addresses the case of a directed graph (edges are directed).



Can this model hyperlinks on the web ?
Can this model 'follow' relations on twitter ?
Is the number of incoming edges on a node a good way of measuring authority ?
Should all incoming edges be weighed the same way for measuring authority ?

Let's view the problem of estimating authority as follows : starting from a (uniformly sampled) random node on the graph, one moves on the graph to a neighbouring node (random move, uniformly sampled among the outgoing edges). Such moves are repeated a large number of times.

- On the above example, after one move, what is the probability distribution of being in each node ? After n moves ? Let us consider that this distribution, after a large number of moves (« random walk »), provides an indicator of authority of each node. Make a quick implementation in your favourite language and check whether it seems to converge, and if « sinks » in the graph seem to have higher authority. Try various graph configurations to check if the estimated authority corresponds to your intuition.

- Let's assume (Frobenius theorem (1910) says it is unique and it does not depend on the starting point) this probability distribution stabilises (converges) over time : what is the key algebraic property to compute directly this stable probability vector ?

Social tagging

Remember about tagging of items (books, pictures etc…)
Try to adapt the non-negative matrix factorization technique to find groups of tags that frequently occur together. How can this be useful ?