RML - HW - Fairness and Privacy

Théo DI PIAZZA

December 2022

1 Simpson's Paradox

In this exercice, we'll work on the Simpson's Paradox. The Simpson paradox a statistical paradox in which a phenomenon observed in several groups is reversed when the groups are combined.

1.1 Question 1

The pourcentage of men admitted is given by:

$$\frac{0.62 * 825 + 0.63 * 560 + ... + 0.06 * 373}{825 + 560 + ... + 373} = 0.45$$

The pourcentage of women admitted is given by:

$$\frac{0.82 * 108 + 0.68 * 25 + ... + 0.07 * 341}{825 + 560 + ... + 373} = 0.30$$

The percentage of men admitted is significantly higher than the percentage of women admitted: +15%. At first sight, this may be seen as biased in favour of men. However, before making this conclusion, it is necessary to look in more detail at the data available.

1.2 Question 2

Comparing the percentages of admissions by department, we observe that they are very close for men and women. In fact, the difference does not seem very significant. On the contrary, for department A, there is even a clear difference in the admission percentage in favour of women.
From this point of view, we could then consider that the admission percentages for each department are not biased. For department A, the trend would even be reversed, and we could possibly speak of a bias in favour of women.

This phenomenon illustrates Simpson's paradox very well. When the data is aggregated, the observations are no longer the same. For example, this could be explained by the fact that the number of male and female applicants is not the same: 2691 men vs. 1835 women.

2 Some fairness criteria

2.1 Question 1

R a binary classifier and that there are only two groups.
If R satisfies separation, it means that true positive rates and the same false positive rates in all groups must be the same.

2.2 Question 2

The point which corresponds to te classifier satisfying the separation criteria is the point that corresponds to the intersection of the 2 ROC curves of the 2 groups. With the example bellow, it's the intersection of the A and B curve.

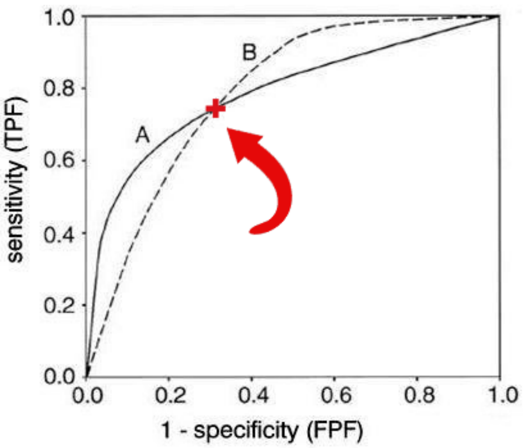

Figure 1: Point that corresponds to the intersection of the 2 ROC curves

2.3 Question 3

If R satisfies sufficiency, $Y \perp\!\!\!\perp A|R$.
In terms of probabilities, it means that $P\{Y = 1|\ l(R) = r, A = a\}$ is the same for groups a and b, with

$$l(R) = P\{Y = 1|\ R = r, A = a\}$$

Then, it comes that:

$$\begin{aligned} r &= P\{Y = 1|\ l(R) = r, A = a\} \\ &= P\{Y = 1|\ R \in l^{-1}(r),\ A = a\} \\ &= P\{Y = 1|\ R \in l^{-1}(r),\ A = b\} \\ &= P\{Y = 1|\ l(R) = r,\ A = b\} \end{aligned}$$

Hence, we showed that : if R satisfies sufficiency, there exists a function l such that $l(R)$ satisfies calibration by group.

2.4 Question 4

In this question, it's needed to show that if A is not independent of Y and R is not independent of Y , then independence and separation cannot both hold.
To answer this question, the contrapositive form will be use. Hence, it's needed to show that if independence and separation both hold, the, A is independent of Y or R is independent of Y.

To start:

$$P\{R = r|A = a\} = \sum_y P\{R = r, Y = y|A = a\}$$

$$= \sum_y P\{R = r|A = a, Y = y\}P\{Y = y|A = a\}$$

$$= \sum_y P\{R = r|Y = y\}P\{Y = y|A = a\} = P\{R = r\}$$

We also have that:

$$P\{R = r\} = \sum_y P\{R = r, Y = y\}$$

$$= \sum_y P\{R = r|Y = y\}P\{Y = y\}$$

With the 2 expressions of above, we have that:

$$\sum_y P\{R = r|Y = y\}P\{Y = y|A = a\} = \sum_y P\{R = r|Y = y\}P\{Y = y\}$$

$$\iff P\{R = r|Y = 0\}P\{Y = 0|A = a\} + P\{R = r|Y = 1\}P\{Y = 1|A = a\}$$
$$= P\{R = r|Y = 0\}P\{Y = 0\} + P\{R = r|Y = 1\}P\{Y = 1\}$$

$$\iff P\{R = r|Y = 0\}P\{Y = 0|A = a\} + (1 - P\{Y = 0|A = a\}P\{R = r|Y = 1\})$$
$$= P\{Y = 0\}P\{R = r|Y = 0\} + (1 - P\{Y = 0\})P\{R = r|Y = 1\}$$

And the equality of above holds if:

$$P\{R = r|Y = 0\} = P\{R = r|Y = 1\} \implies R \perp\!\!\!\perp Y$$

or if:

$$\forall a, P\{Y = 0\} = P\{Y = 0|A = a\} \implies Y \perp\!\!\!\perp A$$

3 Plausible deniability example

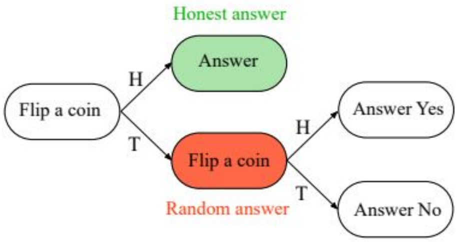

Figure 2: Illustration of exercice 3

3.1 Question 1

To answer this question, let's introduce 2 variables : A for the response where $A \in \{\text{Yes, No}\}$ and T the true response where $T \in \{\text{Yes, No}\}$.

For a given person, we have that

$$P\{A = Yes \mid T = Yes\} = 3/4$$

since $P\{\text{Heads first toss}\} + P\{\text{Tails first toss}\}P\{\text{Heads second toss}\} = 1/2 + (1/2)^2 = 3/4$

And :

$$P\{A = Yes \mid T = No\} = 1/4$$

since $P\{\text{Tails first toss}\}P\{\text{Heads second toss}\} = (1/2)^2 = 1/4$

So finally, it comes that :

$$\frac{P\{A = Yes \mid T = Yes\}}{P\{A = Yes \mid T = No\}}$$

$$= \frac{P\{A = No \mid T = No\}}{P\{A = No \mid T = Yes\}} = \frac{\frac{3}{4}}{\frac{1}{4}} = 3$$

Hence, the protocole is $ln(3)$-DP.

3.2 Question 2

One solution would be to modify the probabilities of the tree in order to obtain a protocol ϵ-DP.

4 Mechanisms

4.1 Question 1

Recall: the p.d.f. of a random variable sampled according to Lap(a) is equal to $f(x) = \frac{1}{2a}e^{-\frac{|x|}{a}}$
Let's show that $M_{Lap}(A, D, \epsilon, \delta) = A(D) + Z$ with $Z \sim Lap(\frac{\Delta_1(A)}{\epsilon})$ is ϵ-DP.

For this, we'll consider two neightbouring databases D_1 and D_2 which only differ from one element : z_{diff}.
Then, we have that: $M_{Lap}(A, D_1, \epsilon, \delta) = A(D_1) + Z$ and $M_{Lap}(A, D_2, \epsilon, \delta) = A(D_2) + Z$.

$$\frac{f_1(z_{diff})}{f_2(z_{diff})} = \frac{\prod_{i=1}^{d} \frac{\epsilon}{\Delta_1(A)} e - \frac{\epsilon}{\Delta_1(A)} |A(D_1)_i - (z_{diff})_i|}{\prod_{i=1}^{d} \frac{\epsilon}{\Delta_1(A)} e - \frac{\epsilon}{\Delta_1(A)} |A(D_2)_i - (z_{diff})_i|}$$

$$= \frac{\prod_{i=1}^{d} e^{-\frac{\epsilon}{\Delta_1(A)} |A(D_1)_i - (z_{diff})_i|}}{\prod_{i=1}^{d} e^{-\frac{\epsilon}{\Delta_1(A)} |A(D_2)_i - (z_{diff})_i|}}$$

$$= \prod_{i=1}^{d} e^{\frac{-\epsilon(|A(D_1)_i - (z_{diff})_i)| - |A(D_2)_i - (z_{diff})_i)|)}{\Delta_1(A)}}$$

$$\leq \prod_{i=1}^{d} e^{\frac{-\epsilon |A(D_2)_i - A(D_1)_i|}{\Delta_1(A)}}$$

$$= e^{\frac{-\epsilon \sum_{i=1}^{d} |A(D_2)_i - A(D_1)_i|}{\Delta_1(A)}}$$

$$\leq e^{\frac{\epsilon ||A(D_2)_i - A(D_1)_i||_1}{\Delta_1(A)}} \leq e^{\epsilon}$$

Hence, $M_{Lap}(A, D, \epsilon, \delta)$ is $\epsilon - DP$.

4.2 Question 2

Recall: to get differential privacy for a process $Proc$, we must have for any databases D_1 and D_2 which are neightbours, and for any output y that:

$$\frac{P\{Proc(D_1) = y\}}{P\{Proc(D_2) = y\}} \leq \epsilon$$

The problem with this definition is that it must hold for any output y. Thus, if there is the slightest possibility that this inequality does not hold, then the property is no longer valid: it is therefore a worst case property.

To relax this, the (ϵ, δ)-DP could be used instead of the ϵ-DP.
The definition is: For all $\epsilon \geq 0$, a randomized algorithm A is (ϵ, δ)-differentially private if, $\forall S \subset Im(A)$ and $\forall$ D_1 and D_2 datasets such as $d(D_1, D_2) = 1$ we have:

$$P(A(D_1) \in S) \geq P(A(D_2) \in S)e^{\epsilon} + \delta$$

δ can be interpreted as the the probability of privacy leakage.

4.3 Question 3

$$\mathbb{E}[||M_{Lap}(A, D, \epsilon, \delta) - A(D)||_1] = \mathbb{E}[||Z||_1] = \sum_{i=1}^{d} \mathbb{E}[|Z_i|]$$

$$\text{Where } \mathbb{E}[|Z_i|] = \int_{-\infty}^{+\infty} |z_i| f(z_i)\, dz_i$$

$$= \int_{-\infty}^{+\infty} |z_i| \frac{\epsilon exp(\frac{-\epsilon|z_i|}{\Delta_1(A)})}{2\Delta_1(A)}\, dz_i \ = \frac{\epsilon}{2\Delta_1(A)} \int_{-\infty}^{+\infty} |z_i| exp(\frac{-\epsilon|z_i|}{\Delta_1(A)})$$

$$= \frac{\epsilon}{2\Delta_1(A)}(\int_{-\infty}^{0} |z_i| exp(\frac{-\epsilon|z_i|}{\Delta_1(A)})dz_i + \int_{0}^{+\infty} |z_i| exp(\frac{-\epsilon|z_i|}{\Delta_1(A)})dz_i)$$

$$= \frac{\epsilon}{2\Delta_1(A)}(\int_{-\infty}^{0} -z_i exp(\frac{-\epsilon|z_i|}{\Delta_1(A)})dz_i + \int_{0}^{+\infty} z_i exp(\frac{-\epsilon|z_i|}{\Delta_1(A)})dz_i)$$

$$= \frac{\epsilon}{2\Delta_1(A)}(-\int_{+\infty}^{0} x_i exp(\frac{-\epsilon|x_i|}{\Delta_1(A)})dz_i + \int_{0}^{+\infty} z_i exp(\frac{-\epsilon|z_i|}{\Delta_1(A)})dz_i)$$

$$= \frac{2\epsilon}{2\Delta_1(A)} \int_{0}^{+\infty} z_i exp(\frac{-\epsilon|z_i|}{\Delta_1(A)})dz_i$$

$$= \frac{2\epsilon}{2\Delta_1(A)} \int_{0}^{+\infty} \frac{\Delta_1(A)x_i}{\epsilon} exp(-|x_i|) \frac{\Delta_1(A)dx_i}{\epsilon}$$

$$= \frac{\Delta_1(A)}{\epsilon} \int_{0}^{+\infty} x_i exp(-|x_i|)dx_i = \frac{\Delta_1(A)}{\epsilon} \int_{0}^{+\infty} |x_i| exp(-|x_i|)dx_i = \frac{\Delta_1(A)}{\epsilon}$$

$$\text{Hence, } \mathbb{E}[||M_{Lap}(A, D, \epsilon, \delta) - A(D)||_1] = d\frac{\Delta_1(A)}{\epsilon}$$

4.4 Question 4

To simplify the calculations, we define $a = \frac{\Delta_1(A)}{\epsilon}$

$$P\{||M_{Lap}(A, D, \epsilon, \delta) - A(D)||_\infty \leq \alpha\} = P\{||Z||_\infty \leq \alpha\} \geq 1 - \beta$$

$$\iff P\{max_i|Z_i| \leq \alpha\} \geq 1 - \beta$$

$$\iff \prod_{i=1}^{d} P\{|Z_i| \leq \alpha\} = \prod_{i=1}^{d} P\{-\alpha \leq Z_i \leq \alpha\} \geq 1 - \beta$$

$$\iff \prod_{i=1}^{d} \int_{-\alpha}^{\alpha} f(z_i)dz_i = \prod_{i=1}^{d} \int_{-\alpha}^{\alpha} \frac{1}{2a} exp(-\frac{|z_i|}{a})dz_i \geq 1 - \beta$$

$$\iff \prod_{i=1}^{d}(\int_{-\alpha}^{0}\frac{1}{2a}exp(\frac{z_i}{a})dz_i + \int_{0}^{\alpha}\frac{1}{2a}exp(-\frac{z_i}{a})dz_i) \geq 1-\beta$$

$$\iff \prod_{i=1}^{d}([\frac{1}{2}exp(\frac{z_i}{a})]_{-\alpha}^{0}) + ([\frac{-1}{2}exp(\frac{-z_i}{a})]_{0}^{\alpha}) \geq 1-\beta$$

$$\iff \prod_{i=1}^{d}(1-exp(\frac{-\alpha}{a})) = (1-exp(\frac{-\alpha}{a}))^{d} \geq 1-\beta$$

$$\iff \alpha \geq -\frac{\Delta_1(A)}{\epsilon}ln(1-(1-\beta)^{\frac{1}{d}})$$

4.5 Question 5

To simplify calculations, let's define : $\sigma^2 = \frac{2ln(\frac{2}{\delta}\Delta_2(A)^2)}{\epsilon^2}$

$$P\{||M_{Lap}(A,D,\epsilon,\delta)-A(D)||_{\infty} \leq \alpha\} = P\{||Z||_{\infty} \leq \alpha\} \geq 1-\beta$$

$$\iff P\{max_i|Z_i| \leq \alpha\} \geq 1-\beta$$

$$\iff \prod_{i=1}^{d}P\{|Z_i| \leq \alpha\} = \prod_{i=1}^{d}P\{-\alpha \leq Z_i \leq \alpha\} \geq 1-\beta$$

$$\iff \prod_{i=1}^{d}(F_{Z_i}(\alpha) - F_{Z_i}(-\alpha)\} \geq 1-\beta$$

$$\iff \prod_{i=1}^{d}\frac{1}{2}(erf(\frac{\alpha}{\sigma\sqrt{2}}) - erf(\frac{-\alpha}{\sigma\sqrt{2}}))\} = \prod_{i=1}^{d}erf(\frac{\alpha}{\sigma\sqrt{2}})\} \geq 1-\beta$$

$$\iff (erf(\frac{\alpha}{\sigma\sqrt{2}}))^{d} \geq 1-\beta$$

$$\iff \alpha \geq \sigma\sqrt{2}erf^{-1}((1-\beta)^{\frac{1}{d}})$$

4.6 Question 6

Regarding to Question 4 and 5, a major strength of the Gaussian mechanism is that it allows adding much less noise.

End of RML Homewok - Fairness. Thank you for reading ! - Théo Di Piazza

theo.dipiazza@gmail.com