Metrics for datasets comparison

Adrien Pavao

April 2018

1 Metrics

Privacy and resemblance with Minimum Distance Accumulation

1. Compute the distance of the nearest neighbor from the other distribution for each point

2. Compute the graph: a distance θ on x axis and the number of points with a minimum distance smaller than θ on y axis.

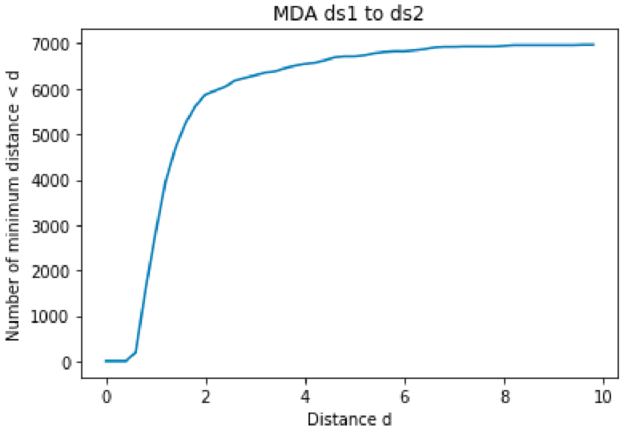
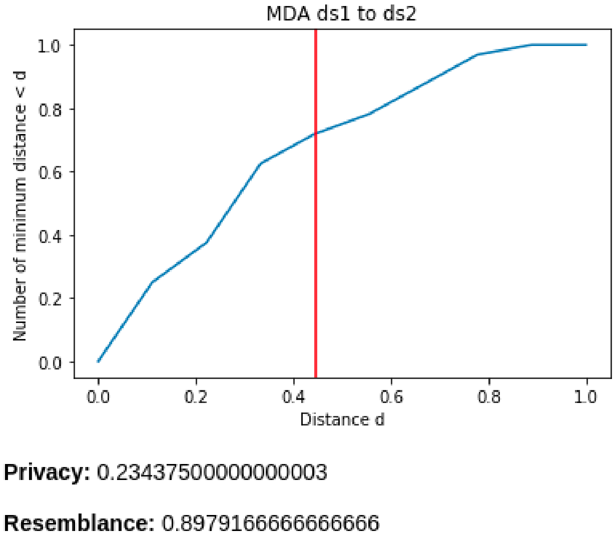


Figure 1: Example of MDA curve

3. Define a threshold distance for privacy/resemblance. Under the threshold, we want points to be as less as possible similar to respect privacy. Over the threshold, we want points to be as more as possible similar to maximize resemblance. The metrics are the areas under the curve on the left and on the right of the threshold.



Privacy: 0.23437500000000003

Resemblance: 0.8979166666666666

Figure 2: Example of MDA curve with threshold, normalization and areas under curve

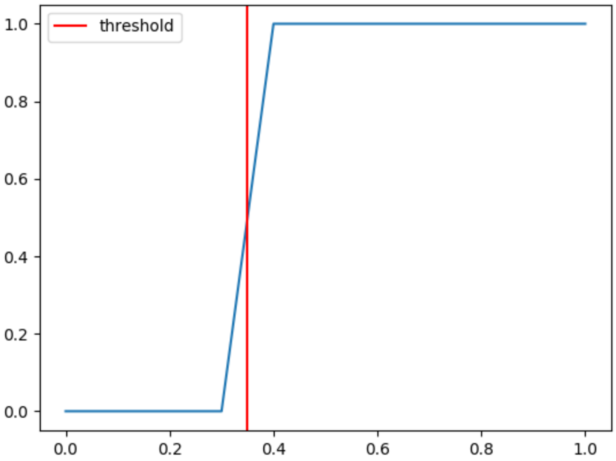


Figure 3: Perfect MDA curve: privacy = 0 and resemblance = 1

Distances and accumulation are normalized to have more robust metrics. The areas are normalized with the threshold to be within 0 and 1.

2 Baseline generator

Values imputations using random forests

1. For each variable (feature) x_i, train a random forest regressor (or classifier for categorical variables) to predict x_i given all the other variables $X_{/x_i}$.

2. Copy the data in a new matrix.

3. Go though the values of the new matrix and, with a probability p, replace it with the output of the imputation model of the variable. The predictions are always done with the original data.

Why random forest?

Random forest is a good idea for a baseline method because:

- It does not extrapolate values: it will only predict values within the range of original data
- It computes fast

One problem may be that the values predicted will be too smooth without the noise.

3 Test of metrics and generator

We compute the privacy and resemblance of original data and generated data for different values of p. Bigger is p, smaller should be the privacy and the resemblance.

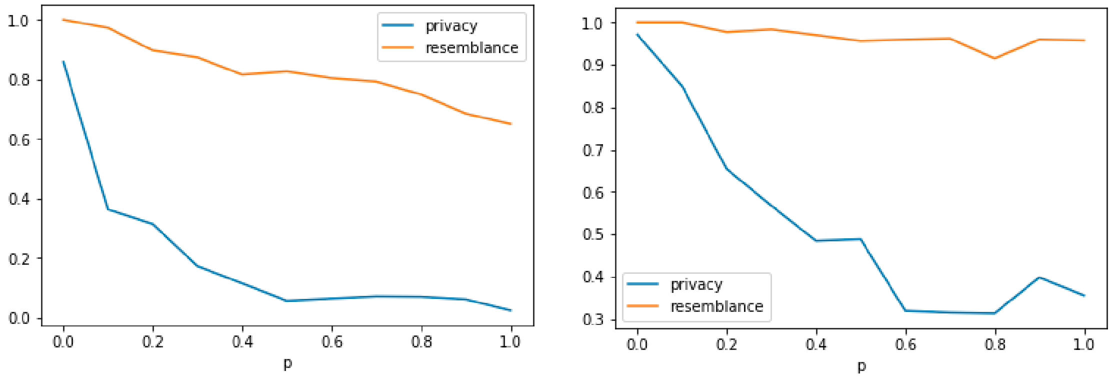


Figure 4: Metrics by p, threshold=0.1 (left), threshold=0.5 (right)

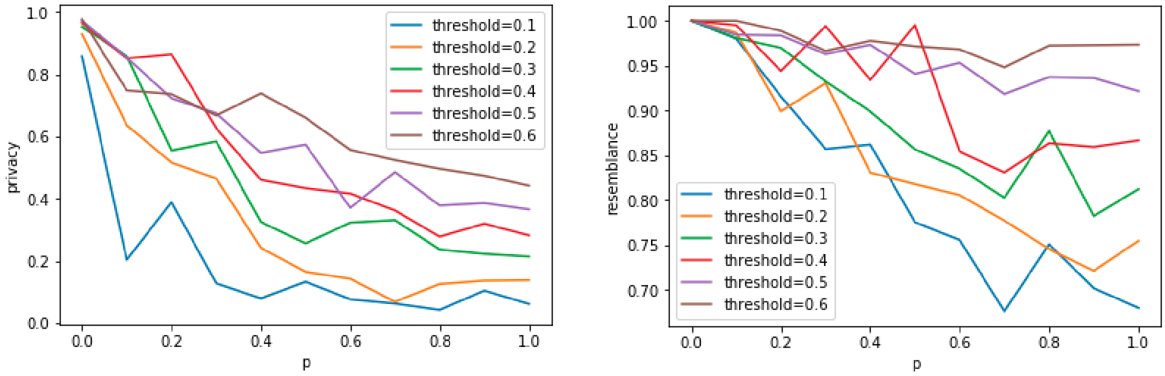


Figure 5: Privacy and resemblance by threshold