Generating Tabular Data using GANs with Differential Privacy
Microdata Protection

02/12/2021
Who we are

ITINNOVATION@ECB

Giacomo Astolfi
Giacomo.Astolfi@ecb.europa.eu

David Köpfer
David.Kopfer@ecb.europa.eu
Imagine we have a data set of many items

Categories: shapes, colors

Numeric: Surface, thickness

... but it is really Confidential !!!

... but we would like to use it anyways for:

- machine learning
- In the test environment
- DB tests
- realistic visualizations
- examples for external tenders
Idea: Let’s randomize

Let’s just randomly draw:

- **shape**
- **color**
- **Surface**
- **thickness**

Fake

Confidential
Problem: loosing **Structure**

- There is **Structure**
- Fake

[Please select]
[Please select]
Idea: Masking

Masking

- Masks the original records
- Protects names and labels
- Can be “unmasked”

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>City</th>
<th>Loan</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>25</td>
<td>HE</td>
<td>Yes</td>
<td>2000</td>
</tr>
<tr>
<td>John</td>
<td>25</td>
<td>77</td>
<td>1</td>
<td>2000</td>
</tr>
</tbody>
</table>
Idea: Data Generation!

Masking
- Masks the original records
- Protects names and labels
- Can be “unmasked”

Data Generation
- Generates new records
- Protects true information
- Can only reveal averages

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>City</th>
<th>Loan</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>25</td>
<td>HE</td>
<td>Yes</td>
<td>2000</td>
</tr>
</tbody>
</table>

Name Age City Loan Income
John 25 77 1 2000
Idea: **Generative Adversarial Neural Nets (GANN)**

- **Generator**
  - random numbers (inspiration)

- **Discriminator**
  - Real
  - Fake

Iterate till both are really good

* ! Confidential! *
Result: GANN

- Generator:
  - has learned the **Structure** of the data
  - easily transferable
  - low computational cost for generation
  - can easily generate large volumes of data

- Discriminator:
  - can pick up very subtle details about the data
  - can be used for **outlier** or **fraud** detection
What does Structure look like?
What does Learning look like?

Epoch: 2

Training iteration

Original Data

Fake data

age

20 40 60 80 100
What does Learning look like?

Training iteration
Differential Privacy

I. ensure that a defined level of “privacy” is kept
II. privatized data cannot be deanonymized without additional information
III. privacy is conserved through many operations
IV. …
Differential Privacy

I. ensure that a defined level of “privacy” is kept
II. privatized data cannot be deanonymized without additional information
III. privacy is conserved through many operations
IV. …

Mount Math

Is X’s data there?
Differential Privacy

I. ensure that a defined level of “privacy” is kept
II. privatized data cannot be deanonymized without additional information
III. privacy is conserved through many operations
IV. …
Intuition: Differential Privacy

amount of real data exposure

Generator

Discriminator

! Confidential!
CTGAN + DPGAN

- Anonymize the preprocessing step of CTGAN
- CTGAN model to generate the data
- DPGAN to add Privacy to a GAN

Result: training this DPCTGAN produces a (private) Generator of data
Differential Privacy

Non-Private CTGAN (no noise)  Private CTGAN (high noise)
Differential Privacy

Non-Private CTGAN (no noise)

Private CTGAN (high noise)
We now have...

- A Data Generator architecture that works with Tabular data and once trained can be moved anywhere
- A tunable approach that allows us to choose the tradeoff between Privacy and Utility of the data we aim to generate

<table>
<thead>
<tr>
<th>Method</th>
<th>Noise</th>
<th>Utility</th>
<th>Detection</th>
<th>Privacy</th>
<th>Epsilon ((\varepsilon))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Classification</td>
<td>Regression</td>
<td>Numerical</td>
<td>Categorical</td>
</tr>
<tr>
<td>Identity</td>
<td>N/A</td>
<td>0.78</td>
<td>0.014</td>
<td>1</td>
<td>0.26</td>
</tr>
<tr>
<td>CTGAN</td>
<td>N/A</td>
<td>0.59</td>
<td>-0.08</td>
<td>0.7</td>
<td>0.32</td>
</tr>
<tr>
<td>DPCTGAN 0.00001</td>
<td>0.575</td>
<td>-0.11</td>
<td>0.65</td>
<td>0.51</td>
<td>0.72</td>
</tr>
<tr>
<td>DPCTGAN 0.001</td>
<td>0.565</td>
<td>-0.23</td>
<td>0.61</td>
<td>0.56</td>
<td>0.74</td>
</tr>
<tr>
<td>DPCTGAN 0.1</td>
<td>0.6</td>
<td>-2.85</td>
<td>0.34</td>
<td>0.59</td>
<td>0.84</td>
</tr>
<tr>
<td>DPCTGAN 1</td>
<td>0.58</td>
<td>-12.5</td>
<td>0.15</td>
<td>0.78</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Thank you for your Attention

ITINNOVATION@ECB

Jack
Giacomo Astolfi
Giacomo.Astolfi@ecb.europa.eu

John
David Köpfer
David.Kopfer@ecb.europa.eu