Data Description We demonstrate Twinify on a COVID-19 related data set [1], to which we refer as *Einstein COVID-19* from now on. The original data consists of 5 644 samples with 110 features. This data set has previously been used by Souza *et al.* [2] in a classification task to predict if a patient is infected with the SARS-Cov-2 virus using 44 data features as predictors. We aim to replicate this analysis on synthetic twin of the relevant features generated by Twinify.

Modelling We use Twinify's automatic modelling capability on our concrete example. Figure 1 shows an excerpt of the model text file: We choose the Normal distribution as feature distribution for continuous features (e.g. the number of leukocytes in the blood sample), the Bernoulli distribution for binary features (such as the COVID-19 test result) and a Poisson distribution for the ordinal age quantile.

The original data has large amount of missing values, for example $\sim 89\%$ of the *Leukocytes* feature's entries are missing. Twinify's automatic modelling handles these using the approach described in the `README`. Finally, we note that the binary features are represented as strings in the data (e.g. `detected` and `not detected` for the *Rhinovirus/Enterovirus* test result) which Twinify automatically encodes.

```
Patient age quantile:   Poisson
Leukocytes:             Normal
Platelets:              Normal
Monocytes:              Normal
Patient addmited to regular ward (1=yes, 0=no): Bernoulli
Influenza B:            Bernoulli
Rhinovirus/Enterovirus: Bernoulli
SARS-Cov-2 exam result: Bernoulli
```

Figure 1: Excerpt of the `model.txt` for the running example on the *Einstein COVID-19* data set. Each line contains the exact label of a feature column in the data table and the feature distribution assigned to it. See `models/model.txt` for the full file.

Learning a Classifier From the Synthetic Twin Following the earlier work by Souza *et al.* [2] we train a gradient boosting method (GBM) classifier to predict if a patient has a SARS-Cov-2 infection based on the 44 clinical predictors. We sample 20 000 instances from the trained model and train the GBM classifier using these synthetic twin data. We evaluate the performance on the same test split of the original data set used in the original analysis [2].

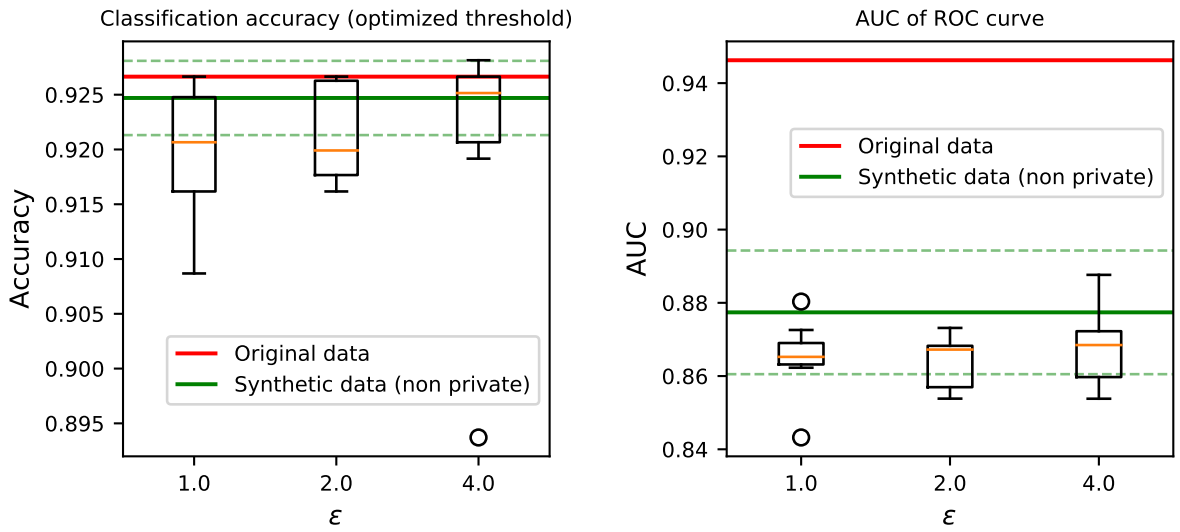
The GBM classifier assigns a probability for a SARS-Cov-2 infection to each testing point. For binary classification we need to specify a decision threshold for these probabilities. After learning the classifier, we optimise this decision threshold, again using only synthetic data, to maximize the sum of the classifier's sensitivity and specificity. To compare, we also train the same classifier using the original data as well as a non-private synthetic twin data set.

Figure 2a shows that the classifier trained with synthetic data achieves high accuracy, similar to the one trained on original data, with a reasonable level of privacy ($\epsilon = 2$). Besides the classification accuracy, the classifier trained with synthetic data is reasonably well calibrated as shown in Figure 2b.

How to Reproduce An explanation of all contained scripts and how to run them is given in the separate `README` contained in the example folder (`examples/covid19_analysis/README.md`).

References

[1] Hospital Israelita Albert Einstein. Diagnosis of COVID-19 and its clinical spectrum. *Kaggle*, 2020. `https://www.kaggle.com/einsteindata4u/covid19/`.



(a) (b)

Figure 2: Left: The GBM classifier trained with synthetic data achieves accuracy comparable to the classifier trained with original data with reasonable level of privacy ($\epsilon = 2$). Right: The classifier trained with synthetic data is reasonable balanced as characterised by the area under the ROC curve (AUC). On both sides the figure shows results of 10 independent runs (different seeds) for varying levels of privacy. The boxplots show the median (orange line inside the box) with the box extending from the first to the third quartile. The non-private comparison was also repeated 10 times, with the solid green line indicating the mean and dashed lines the standard deviation.

[2] Tharsis Souza, Gustavo Wenzel Sainatto, and Heli S. P. Souza. COVID-19 machine learning-based rapid diagnosis from common laboratory tests. *Towards Data Science*, 2020. `https://towardsdatascience.com/covid-19-machine-learning-based-rapid-diagnosis-from-common-laboratory-tests-afafa9178372`, Accessed: 2020-06-15.