DSC-50 Synthetic data using generative models

2018-06-12

The project involves the generation of synthetic data using machine learning to replace real data for the purpose of data processing and, potentially, analysis. This is particularly useful in cases where the real data are sensitive (for example, microdata, medical records, defence data). Additionally, the methods developed as part of the project may be used for imputation.

Team members

- Ioannis Kaloskampis (Lead, point of contact)
- Chaitanya Joshi
- David Pugh
- Lanthao Benedikt
- Alex Noyvirt
- Louisa Nolan

The need

In our digital world, data are produced at an exponential rate. Various organisations such as government departments, banks and retailers etc. would like to exploit the so called "big data" to build statistical models to make accurate decisions and predict a number of important measures, such as the inflation rate and exchange rates. However, the raw data are often sensitive. In this project, we propose methods that generate synthetic data to replace the raw data for the purposes of processing and analysis.

Impact

The project will result in a safer, easier and faster way to share data between the Office for National Statistics (ONS) and the research communities in cases where the real data are sensitive. Additionally, It will make sharing data between the research communities and ONS easier and faster. Furthermore, the project is linked to several current ONS Data Science projects (such as Trade and Housing).

Data science

We investigate several state-of-the-art algorithms that are used to generate synthetic data such as generative adversarial networks (GANs), variational autoencoders (VAE) and autoregressive models. Additionally, since the project involves big data, we are particularly interested in the efficient implementation of the synthetic data generation algorithms using graphics processing units (GPUs).

Stakeholders

- ONS Methodology
- ONS Trade team
- United Nations global platform

Further information

Please contact datasciencecampus@ons.gov.uk for more information.

Updates

2019-02-22T14:41:44Z

The project reports are now published on the Data Science Campus website:

- Synthetic data for public good
- Synthetic data for public good and art
- Generative adversarial networks (GANs) for synthetic dataset generation with binary classes