Synthetic Tabular Data Generation

Using Transformers (NVIDIA)

Duration: 08 hours

In this course you'll learn the end-to-end development workflow for generating synthetic data using Transformers, including data preprocessing, model pre-training, fine-tuning, inference, and evaluation

Course Prerequisites:

- Competency in the Python 3 programming language
- Basic understanding of Machine Learning and Deep Learning concepts and pipelines
- Experience building Machine Learning models with Tabular data
- Basic understanding of language modelling and Transformers

About this Course

Synthetic data generation (SDG) is a data augmentation technique necessary for increasing the robustness of models by supplying training data. With advancements in pre-trained Transformers, data scientists across all industries are learning to use them to generate synthetic training data for downstream predictive tasks. In this course, you'll explore the use of Transformers for synthetic tabular data generation. We will use credit card transactions data and the Megatron framework for the course, but this technique is broadly applicable to tabular data in general.

Learning Objectives

By participating in this course, you will:

- Learn how synthetic data can improve model performance.
- Learn to use Transformers for Synthetic Data Generation.
- Go through the end-to-end development workflow for generating synthetic data using Transformers, including data preprocessing, model pre-training, fine-tuning, inference, and evaluation.

Topics Covered

Upon completion, you will have a basic understanding of how to generate synthetic tabular data for downstream predictive tasks

Course Outline

- Learn how synthetic data can improve model performance.

- Learn to use Transformers for Synthetic Data Generation.
- Go through the end-to-end development workflow for generating synthetic data using Transformers, including data preprocessing, model pre-training, fine-tuning, inference, and evaluation.

Upon completion, you will have a basic understanding of how to use transformer-based models to generate synthetic tabular data for downstream predictive tasks