newData.ai
data-driven success starts here

Exploring Synthetic Data

by Taymour | December 27, 2022

Synthetic data is artificially generated data used for the development of software, testing, and training machine learning models. In some cases, it has advantages over real data, including the ability to generate large amounts of data quickly, control over the data's characteristics, and the ability to create data for rare or specific scenarios. In this blog, we will discuss five common use cases for synthetic data in more detail, and explore the benefits and considerations of using synthetic data for each of these use cases.

1. Data Augmentation

One common example application for synthetic data is data augmentation, or the process of generating additional data that is similar to an existing data set. This can be especially useful when the real data set is small or when it is difficult to obtain additional real data. By generating synthetic data that is similar to the real data, organizations can augment their data sets and improve the accuracy and performance of their machine learning models.

Synthetic data is a cost-efficient and time-saving solution for data augmentation, as it can be generated quickly and in large quantities at a fraction of the price of working with real data. This can be especially useful when the real data set is small or expensive to acquire, as it can be difficult to obtain enough real data to train or test a machine learning model. Synthetic data can also be generated to special characteristics, such as specific distributions or patterns, which can be useful for testing or training a model.

Another advantage of using synthetic data for data augmentation is that it can be used to create data for rare or specific scenarios. For example, if a machine learning model is being developed to predict

the likelihood of a rare event occurring, it may be difficult to obtain enough real data to train the model. Synthetic data can be used to create additional data for this rare event, which can improve the accuracy of the model.

There are also privacy and security benefits to using synthetic data for data augmentation. Synthetic data can be used to protect the privacy and security of real data by generating data that is similar to the real data, but does not contain any sensitive or personally identifiable information. This can be useful for organizations that need to handle large amounts of sensitive data but want to protect the privacy of individuals.

While synthetic data can be a useful tool for data augmentation, it's important to keep in mind that it may not always be as representative of the real world as real data. This can be especially true if the synthetic data is not generated correctly or does not accurately reflect the types of data the model will encounter in practice. To ensure the best results, it's important to carefully consider the characteristics of the synthetic data and ensure that it is as representative of real data as possible.

2. Training Machine Learning Models

Synthetic data can be used to accelerate development projects by simulating various datasets and prototyping applications more quickly than using real-world data. It can be particularly useful in machine learning for training models or assisting in decision making when real data is not available or is insufficient.This can be useful for scenarios that are rare or hard to replicate in the real world. Synthetic data can be a valuable resource for training machine learning models due to its ability to be quickly and easily generated in large quantities. When real data is scarce or difficult to obtain, it can be challenging to gather enough data to effectively train a model. Synthetic data can be tailored to have specific characteristics, such as specific distributions or patterns, which can be useful for training purposes. Additionally, synthetic data can be generated to include specific scenarios or edge cases, which can help improve the robustness and reliability of the model and enhance its performance in real-world situations.

Synthetic data can be used to speed up development projects in a variety of ways. First, it can simulate complex datasets that would otherwise be difficult or impossible to replicate in the real world. This allows organizations to prototype applications and collect insights faster and more accurately than using real-world data. Additionally, synthetic data can help train and test machine learning algorithms with little or no bias. Synthetic datasets can be created to simulate any type of data needed for practice, training, or experimentation. This helps organizations ensure they are

prepared to handle real-world data without worrying about accessing confidential information.

However, it's important to be mindful of the risk of overfitting, as a model that is overly reliant on synthetic data may not perform well on real-world data. To avoid overfitting, it's important to ensure that the synthetic data is representative of the real-world data and to use a diverse and representative training dataset. Overfitting is a common issue in machine learning that occurs when a model becomes too closely tied to the training data and is not able to generalize well to new, unseen data. One way that synthetic data can contribute to overfitting is if it is not representative of the real-world data that the model will encounter, such as if it has specific patterns or distributions that do not reflect the real-world data.

Overfitting is a common issue in both machine learning and statistical models that occurs when a model becomes too closely tied to the training data and is not able to generalize well to new, unseen data. It may be more of a concern in machine learning models for a few reasons. One reason is that machine learning models often have a larger number of parameters that can be adjusted, increasing the risk of overfitting. Statistical models typically have a smaller number of parameters, which can make them less prone to overfitting. Another reason is that machine learning models are often trained on larger datasets, which can also increase the risk of overfitting. When a model is trained on a large dataset, it may pick up on subtle patterns that may not be representative of the overall population, leading it to become overly reliant on those patterns and not generalize well to new data. To prevent overfitting, it is important to use diverse and representative training datasets, carefully consider the characteristics of the synthetic data, and use appropriate evaluation metrics and techniques.

3. Testing machine learning models

Another common use case for synthetic data is testing machine learning models, or using synthetic data to evaluate the performance of a model. This can be especially useful when real data is not representative of the types of data the model is expected to encounter in the real world. By using synthetic data to test machine learning models, organizations can improve the accuracy and performance of their models and make data-driven decisions by generating data expected to be encountered in the real world. This can help ensure that the model is robust and reliable, and can help improve the model's performance in the real world.

There are several ways that a machine learning model can be tested for overfitting. One common method is to use cross-validation. In cross-validation, the training dataset is split into multiple smaller

datasets, and the model is trained and evaluated on each of these datasets. This allows the model to be evaluated on data that it has not seen during training, which can help identify if the model is overfitting. If the model performs well on the training data but poorly on the validation data, it may be a sign of overfitting.

Another method for testing for overfitting is to use a holdout dataset. In this approach, a portion of the training dataset is set aside and not used for training. The model is trained on the remaining data and then evaluated on the holdout dataset. If the model performs well on the training data but poorly on the holdout dataset, it may be overfitting.

Regularization techniques, such as L2 regularization, can also be used to help prevent overfitting by limiting the complexity of the model. Regularization techniques can be applied during the training process, and their effectiveness can be evaluated by comparing the performance of the model on the training and validation datasets.

Another way to test for overfitting is to compare the performance of the model on the training and test datasets. If the model performs significantly better on the training data than on the test data, it may be overfitting.

4. Protecting Privacy and security

Another common use case for synthetic data is protecting the privacy and security of real data, or generating data that is similar to the real data but does not contain any sensitive or personally identifiable information. This can be especially useful for organizations that need to handle large amounts of sensitive data but want to protect the privacy of individuals. By using synthetic data to protect privacy and security, organizations can improve the accuracy and reliability of their models and ensure compliance with privacy regulations. The synthetic data does not contain any sensitive or personally identifiable information. When migrating from one platform or database to another, synthetic datasets can be used in place of sensitive real customer information, protecting customer privacy during transitional periods. This is useful for organizations that are required to comply with privacy regulations, such as the General Data Protection Regulation (GDPR) in the European Union.

Real data that is private or sensitive is often subject to "permissible use" restrictions, which limit the ways in which the data can be accessed and used. Permissible use restrictions are designed to protect the privacy and security of the data and ensure that it is only used for specific, authorized purposes. These restrictions can include limitations on who has access to the data, how the data can

be used, and what types of analysis or processing can be performed on the data. Permissible use restrictions are typically outlined in the terms of service or data use agreements that govern access to the data.

Permissible use restrictions can be a challenge for organizations that rely on private data for their work, as they may limit the ways in which the data can be used and analyzed. For example, an organization may only be permitted to use private data for a specific research project, or it may be required to delete the data after a certain period of time. In some cases, these restrictions can make it difficult or impossible for organizations to use the data in the way they need to.

One way to overcome these restrictions is to use synthetic data, which can be created to be similar to real data but does not contain any sensitive or personally identifiable information. By using synthetic data, organizations can gain access to data that is similar to real data, but is not subject to the same permissible use restrictions. This can be especially useful for organizations that need to perform complex analysis or processing on data but are limited by permissible use restrictions. Synthetic data can provide a way for organizations to gain the insights they need without having to worry about accessing real data that is subject to these restrictions.

Synthetic data enables overseas data analysts to analyze data that may not be possible to transfer outside the US due to regulations. In some cases, real data may be subject to strict regulations that prevent it from being transferred across borders, making it difficult or impossible for overseas data analysts to access the data they need to perform their work. By using synthetic data, data analysts can gain access to data that is similar to the real data, but does not contain any sensitive or personally identifiable information. This can be especially useful for organizations that need to collaborate with data analysts located in different countries and are subject to data transfer restrictions. Synthetic data can provide a way for overseas data analysts to perform their work without having to worry about accessing real data that is subject to regulatory constraints.

5. Data Exploration and Analysis

Synthetic data can be extremely helpful for data exploration and analysis, as it can be tailored to have specific distributions or patterns and scenarios that would not be available in real-world data. It can help reveal trends and patterns that may not be apparent in real data.

The newly created fake data can be customized to exhibit particular distributions and patterns, as well as represent specific scenarios or edge cases. This allows organizations to gain insights and ask

questions that may not have been possible with real-world data alone. Additionally, this type of fake data is ideal for situations where using actual customer information is not possible or desirable. These features make synthetic data a powerful tool for data exploration and analysis, allowing data scientists to gain insights and ask questions that may not have been possible with real data alone. Additionally, synthetic data can be a valuable resource when real data is scarce or difficult to obtain, allowing data scientists to explore and analyze data trends and patterns even when real data is not available.

This allows data scientists to gain insights and ask questions that may not have been possible if they had only relied on real-world data, as well as explore and analyze data trends and patterns in cases where obtaining real data is difficult or impossible. In addition, synthetic data can provide a safer alternative when working with sensitive customer information, ensuring privacy while still providing the necessary insights.

In conclusion, synthetic data is a valuable tool that can be used in a variety of applications, including data augmentation, training machine learning models, testing and evaluating machine learning models, creating simulations, and data privacy and security. Synthetic data has several advantages over real data, including the ability to be quickly and easily generated in large quantities, control over the data's characteristics, and the ability to create data for rare or specific scenarios. However, it is important to carefully consider the characteristics of the synthetic data and ensure that it is representative of the real-world data as much as possible to prevent overfitting and ensure the best results. Synthetic data can be a powerful resource for organizations looking to accelerate development projects, improve machine learning model performance, and protect the privacy and security of sensitive data.
