SageMaker Tools

SageMaker is expanding rapidly with the addition of several tools!

But do remember that, like any other AWS service, not all these tools will gain traction and usage. Usually, new services will come in, and existing services will also silently disappear. For now, getting basic familiarity is sufficient for the exam.

You may see a couple of questions on the below topics.

SageMaker Studio

SageMaker Studio is an AWS developed fully integrated development environment (IDE) for Machine Learning

The Studio extends the standard Jupyter notebook instances (referred to as instance-based notebooks) in a few different ways

1. You can quickly launch Studio notebooks without requiring to provision an instance manually
2. Studio notebook startup-time is 5-10 times faster than instance-based notebooks
3. Flexibility to choose from an extensive collection of instance types
4. Each user gets an isolated home directory in EFS (Elastic File System)
5. The user home directory is automatically mounted into all studio notebooks that you use. So, you can access your files from any instance
6. You can easily share your notebook with your peers and colleagues
7. Studio is integrated with AWS Single Sign-On (SSO). You can use your corporate credentials to access your Studio notebook (no need for AWS user credentials)
8. Studio has a **JumpStart** capability that includes 150+ pre-trained open-source models from PyTorch Hub and TensorFlow Hub.
9. JumpStart also includes example solutions for image processing, natural language processing, demand forecasting, fraud detection, and so forth
10. Studio has visualization capabilities for SageMaker Clarify, SageMaker Experiments, Model Monitor, Debugger, Data Wrangler, Pipelines [Looks like AWS is moving towards using Studio as the primary IDE for interacting and integrating with other SageMaker tools]

SageMaker Debugger

Debugger help you track your training job and identify issues such as CPU, GPU, Disk IO, Network, Memory bottlenecks, Model issues such as Overfitting

Metrics are visualized using SageMaker Studio along with remediation advice

Data Wrangler

Data Wrangler is a data analysis and preparation tool for machine learning applications. You can think of this as an extract-transform-load (ETL) type tool.

“You can create a data flow to combine datasets from different data sources, identify the number and types of transformations you want to apply to datasets, and define a data prep workflow that can be easily integrated into an ML pipeline.”

1. You can import data from one or more sources such as S3, Athena, Redshift
2. Transform the data for your needs. For example, string, number and date formatting, feature engineering, categorical encoding, and so forth
3. Analyze features in your dataset with built-in visualization such as histogram, scatter plots, correlation
4. Export data

Auto Pilot

Before SageMaker was born, AWS had another Machine Learning capability simply named AWS Machine Learning Service that can automatically create machine learning models and tune them. One fine morning, AWS announced they were phasing out that product (I got a panic attack and had to rewrite the entire course – six months gone)

Auto Pilot is the latest incarnation of that tool!

You simply provide a tabular dataset and specify the target column to predict (regression, binary classification, and multi-class classification)

Auto Pilot will automatically explore the dataset, do feature engineering, and explore different solutions to find the best model

It will automatically try multiple algorithms such as linear models, XGBoost, and Deep Learning

You can directly deploy the model to production

https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-model-support-validation.html

https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-automate-model-development.html

Ground Truth

Ground Truth is an automatic labeling service.

Labeled data with correct answers are required for many supervised learning problems (classification, object detection, sentiments, and so forth)

However, labeling the dataset is very time consuming and labor-intensive

Ground Truth automates this process by using a combination of machine learning and human-in-the-loop workflow.

Ground Truth may sound similar to the Augmented AI capability that we saw in the previous lecture. Augmented AI is used for building a custom workflow for your machine learning solution. Whereas Ground Truth is explicitly intended for labeling

You need to provide sample data with correct labels

Ground Truth will use machine learning to learn from the sampled data and attempt to label the rest of the data

For low-confidence scores, Ground Truth will send the data to human labelers for review

The data labeled by humans are used to improve the model

This process repeats until all the raw data are labeled.

For human labelers, you can use:

- Mechanical Turk for crowdsourced labelers
- Private Workforce (your own employees) or
- AWS Marketplace labeling service providers

https://aws.amazon.com/sagemaker/groundtruth/faqs/

Distributed Training

When working with large datasets, you may run into a situation where the time to train a model is too long

One solution is to increase the number of CPUs and GPUs in the training instance

Another option is to scale by using multiple instances

AWS recommends that you try a larger instance before trying to increase the number of instances

If the problem requires multiple instances, you would need to ensure all the instances are in the same region and availability zone. This will ensure network latency is kept to a minimum as instances will frequently share the data and learned parameters.

SageMaker SDK ensures all training instances are deployed in the same availability zone

Many of the SageMaker built-in algorithms automatically support distributed training

If you write a custom algorithm, AWS recommends that you use SageMaker Distributed Data Parallel library and SageMaker Distributed Model Parallel library

You can approach distributed training in two ways: Data Parallel and Model Parallel

"Data parallel *is the most common approach to distributed training: You have a lot of data, batch it up, and send blocks of data to multiple CPUs or GPUs (nodes) to be processed by the neural network or ML algorithm, then combine the results"*

*"*A model parallel *approach is used with large models that won't fit in a node's memory in one piece; it breaks up the model and places different parts on different nodes. In this situation, you need to send your batches of data out to each node so that the data is processed on all parts of the model."*

https://docs.aws.amazon.com/sagemaker/latest/dg/distributed-training.html#distributed-training-scenarios

FAQ

Please review SageMaker FAQs before your exam: https://aws.amazon.com/sagemaker/faqs/