Cloud Guru - 5 - Modeling Quiz

QUESTION 1

You are working on a model that tries to predict the future revenue of select companies based on 50 years of historic data from public financial filings. What might be a strategy to determine if the model is reasonably accurate?

- [x] Use a set of the historic data as testing data to back-test the model and compare results to actual historical results.
- [ ] Randomize the training data and reserve 20% as a validation set after the training process is completed.
- [ ] Use Random Cut Forest to remove any outliers and rerun the algorithm on the last 20% of the data.
- [ ] Use a softmax function to invert the historical data then run the validation job from most recent to earliest history.

Good work!

Time-series data should be training and validated in order as it has, by definition, time as a potential influencing feature. A common method to validate time-series data is backtesting, or replaying the historical data as if it were new data, and then evaluating the model on how successful it predicted the historic values.

QUESTION 2

We are using a CSV dataset for unsupervised learning that does not include a target value. How should we indicate this for training data as it sits on S3?

- CSV data format should not be used for unsupervised learning algorithms.
- ✓ Include label_size=0 appended to the Content-Type key.
- Enable pipe mode when we initiate the training run.
- SageMaker will automatically detect the data format for supervised learning algorithms.
- Include a reserved word metadata key of "ColumnCount" for the S3 file and set it to the number of columns.

Good work!

To run unsupervised learning algorithms that don't have a target, specify the number of label columns in the content type. For example, in this case 'text/csv;label_size=0'

QUESTION 3

You have been provided with a cleansed CSV dataset you will be using for a linear regression model. Of these tasks, which might you do next?

Choose 2

- [ ] Perform one-hot encoding on the softmax results.
- [x] Split the data into testing and training datasets.
- [ ] Run a Peterman distribution on the data to sort it properly for linear regression.
- [x] Run a randomization process on the data.

Good work!

When given a dataset, we should randomize it before separating it into testing and training sets. However, if it is a time-series dataset, you could just split the data into testing and training datasets without randomization. The one-hot encoding and Peterman distribution are nonsense answers.

QUESTION 4

We are running a training job over and over again using slightly different, very large datasets as an experiment. Training is taking a very long time with your I/O-bound training algorithm and you want to improve training performance. What might you consider?

Choose 2

- [x] Make use of pipe mode to stream data directly from S3.
- [ ] Convert the data format to an Integer32 tensor.
- [ ] Use the SageMaker console to change your training job instance type from an ml.c5.xlarge to a r5.xlarge.
- [x] Convert the data format to protobuf recordIO format.
- [ ] Make use of file mode to stream data directly from S3.

Good work!

The combination of using the protobuf recordIO format and pipe mode will result in improved performance for I/O-bound algorithms because the data can be streamed directly from S3 versus having to be first copied to the instance locally.

QUESTION 5

You want to be sure to use the most stable version of a training container. How do you ensure this?

- Use the ECR repository located in US-EAST-2.
- Use the path to the global container repository.
- Use the :latest tag when specifying the ECR container path.
- ✓ Use the :1 tag when specifying the ECR container path.

Good work!

When specifying a training or inference container, use the :1 tag at the end of the path to use the stable version. If you want the latest version, use :latest but that might not be backward compatible.

When you issue a CreateModel API call using a built-in algorithm, which of the following actions would be next?

- SageMaker provisions an EMR cluster and prepares a Spark script for the training job.
- Sagemaker provisions an EC2 instances using the appropriate AMI for the algorithm selected from the regional container registry.
- SageMaker launches an appropriate training container from the algorithm selected from the regional container repository.
- SageMaker launches an appropriate inference container for the algorithm selected from the global container repository.
- ✓ SageMaker launches an appropriate inference container for the algorithm selected from the regional container repository.
- Sagemaker provisions an EC2 instances using the appropriate AMI for the algorithm selected from the global container registry.

Good work!

CreateModel API call is used to launch an inference container. When using the built-in algorithms, SageMaker will automatically reference the current stable version of the container.

QUESTION 7

Which of the following mean that our algorithm predicted false but the real outcome was true?

- True Positive
- False Negative
- False Affirmative
- False Positive
- True Negative

Good work!

A false negative is when the model predicts a false result but the real outcome was true.

QUESTION 8

We are using a k-fold method of cross-validation for our linear regression model. What outcome will indicate that our training data is not biased?

- Bias is not a concern with linear regression problems as the error function resolves this.
- Each subsequent k-fold validation round has a decreasing error rate over the one prior.
- Each subsequent k-fold validation round has an increasing accuracy rate over the one prior.
- ✓ All k-fold validation rounds have roughly the same error rate.
- K-fold is not appropriate for us with linear regression problems.

Good work!

When using a k-fold cross validation method, we want to see that all k-groups have close to the same error rate. Otherwise, this may indicate that the data was not properly randomized before the training process.

QUESTION 9

You have launched a training job but it fails after a few minutes. What is the first thing you should do for troubleshooting?

- Submit the job with AWS X-Ray enabled for additional debug information.
- Ensure that your instance type is large enough and resubmit the job in a different region.
- ✓ Go to CloudWatch logs and try to identify the error in the logs for your job.
- Check to see that your Notebook instance has the proper permissions to access the input files on S3.
- Go to CloudTrail logs and try to identify the error in the logs for your job.

Good work!

All errors in a training job will be logged in CloudWatch, so that should be your first stop to determine the cause of what the failure might be.

QUESTION 10

We are designing a binary classification model that tries to predict whether a customer is likely to respond to a direct mailing of our catalog. Because it is expensive to print and mail our catalog, we want to only send to customers where we have a high degree of certainty they will buy something. When considering if the customer will buy something, what outcome would we want to minimize in a confusion matrix?

- [x] False Positive
- [ ] True Negative
- [ ] False Negative
- [ ] True Positive
- [ ] False Affirmative

Good work!

We would want to minimize the occurrence of False Positives. This would mean that our model predicted that the customer would buy something but the actual outcome was that the customer did not buy anything.

You are consulting for a mountain climbing gear manufacturer and have been asked to design a machine learning approach for predicting the strength of a new line of climbing ropes. Which approach might you choose?

- ✗ You would approach the problem as a linear regression problem to predict the tensile strength of the rope based on other ropes.
- You would choose a binary classification approach to determine if the rope will fail or not.
- ✓ You would recommend they do not use a machine learning model.
- You would choose a simulation-based reinforcement learning approach.
- You would choose a multi-class classification approach to classify the rope into an appropriate price range.

Sorry!

Correct Answer

We take care not to assume every problem is a machine learning problem. In this case we can test the strength of a rope through physical tests, so creating a machine learning problem does not make sense.

QUESTION 12

Your company currently has a large on-prem Hadoop cluster that contains data you would like to use for a training job. Your cluster is equipped with Mahout, Flume, Hive, Spark, and Ganglia. How might you most efficiently use this data?

- [x] Ensure that Spark is supported on your Hadoop cluster and leverage the SageMaker Spark library.
- [ ] Use Mahout on the Hadoop Cluster to preprocess the data into a format that is compatible with SageMaker. Export the data with Flume to the local storage of the training container and launch the training job.
- [ ] Using EMR, create a Scala script to export the data to an HDFS volume. Copy that data over to an EBS volume where it can be read by the SageMaker training containers.
- [ ] Use Data Pipeline to make a copy of the data in Spark DataFrame format. Upload the data to S3 where it can be accessed by the SageMaker training jobs.

Good work!

If the Hadoop cluster has Spark, you can use the SageMaker Spark Library to convert Spark DataFrame format into protobuf and load onto S3. From there, you can use SageMaker as normal.