MLS-C01 Q&As

AWS Certified Machine Learning - Specialty (MLS-C01)

Pass Amazon MLS-C01 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.certbus.com/aws-certified-machine-learning-specialty.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Amazon Official Exam Center

- **Instant Download** After Purchase
- **100% Money Back** Guarantee
- **365 Days** Free Update
- **800,000+** Satisfied Customers

QUESTION 1

A machine learning (ML) specialist is developing a deep learning sentiment analysis model that is based on data from movie reviews. After the ML specialist trains the model and reviews the model results on the validation set, the ML specialist discovers that the model is overfitting.

Which solutions will MOST improve the model generalization and reduce overfitting? (Choose three.)

A. Shuffle the dataset with a different seed.

B. Decrease the learning rate.

C. Increase the number of layers in the network.

D. Add L1 regularization and L2 regularization.

E. Add dropout.

F. Decrease the number of layers in the network.

Correct Answer: DEF

A: possible but unlikely for movie reviews

B: wrong https://www.google.com/url?sa=tandrct=jandq=andesrc=sandsource=webandcd=andcad=rjaanduact=8andved=2ahUKEwi31N_10eX9AhWYQ0EAHXDFCAwQFnoECA8QAwandurl=https%3A%2F%2Fdeepchecks.com%2Fquestion%2Fdoes-learningrate-affect-overfitting%2Fandusg=AOvVaw19RT-u_XyEe8FG_10R6aFC

C: wrong because would increase complexity and potentially overfitting

D: correct

E: correct

F: correct

QUESTION 2

A machine learning (ML) specialist wants to create a data preparation job that uses a PySpark script with complex window aggregation operations to create data for training and testing. The ML specialist needs to evaluate the impact of the number of features and the sample count on model performance.

Which approach should the ML specialist use to determine the ideal data transformations for the model?

A. Add an Amazon SageMaker Debugger hook to the script to capture key metrics. Run the script as an AWS Glue job.

B. Add an Amazon SageMaker Experiments tracker to the script to capture key metrics. Run the script as an AWS Glue job.

C. Add an Amazon SageMaker Debugger hook to the script to capture key parameters. Run the script as a SageMaker processing job.

D. Add an Amazon SageMaker Experiments tracker to the script to capture key parameters. Run the script as a

SageMaker processing job.

Correct Answer: D

https://docs.aws.amazon.com/sagemaker/latest/dg/experiments-create.html#:~:text=CreateTrainingJob-,Processing,-Processor.run

QUESTION 3

A manufacturing company asks its Machine Learning Specialist to develop a model that classifies defective parts into one of eight defect types. The company has provided roughly 100000 images per defect type for training During the injial training of the image classification model the Specialist notices that the validation accuracy is 80%, while the training accuracy is 90% It is known that human-level performance for this type of image classification is around 90%

What should the Specialist consider to fix this issue1?

A. A longer training time

B. Making the network larger

C. Using a different optimizer

D. Using some form of regularization

Correct Answer: D

QUESTION 4

A company wants to create a data repository in the AWS Cloud for machine learning (ML) projects. The company wants to use AWS to perform complete ML lifecycles and wants to use Amazon S3 for the data storage. All of the company\\'s

data currently resides on premises and is 40 in size.

The company wants a solution that can transfer and automatically update data between the on-premises object storage and Amazon S3. The solution must support encryption, scheduling, monitoring, and data integrity validation.

Which solution meets these requirements?

A. Use the S3 sync command to compare the source S3 bucket and the destination S3 bucket. Determine which source files do not exist in the destination S3 bucket and which source files were modified.

B. Use AWS Transfer for FTPS to transfer the files from the on-premises storage to Amazon S3.

C. Use AWS DataSync to make an initial copy of the entire dataset. Schedule subsequent incremental transfers of changing data until the final cutover from on premises to AWS.

D. Use S3 Batch Operations to pull data periodically from the on-premises storage. Enable S3 Versioning on the S3 bucket to protect against accidental overwrites.

Correct Answer: C

Configure DataSync to make an initial copy of your entire dataset, and schedule subsequent incremental transfers of changing data until the final cut-over from on-premises to AWS. Reference: https://aws.amazon.com/datasync/faqs/

QUESTION 5

A Machine Learning Specialist has built a model using Amazon SageMaker built-in algorithms and is not getting expected accurate results The Specialist wants to use hyperparameter optimization to increase the model\\'s accuracy Which method is the MOST repeatable and requires the LEAST amount of effort to achieve this?

A. Launch multiple training jobs in parallel with different hyperparameters

B. Create an AWS Step Functions workflow that monitors the accuracy in Amazon CloudWatch Logs and relaunches the training job with a defined list of hyperparameters

C. Create a hyperparameter tuning job and set the accuracy as an objective metric.

D. Create a random walk in the parameter space to iterate through a range of values that should be used for each individual hyperparameter

Correct Answer: C

QUESTION 6

A Machine Learning Specialist kicks off a hyperparameter tuning job for a tree-based ensemble model using Amazon SageMaker with Area Under the ROC Curve (AUC) as the objective metric This workflow will eventually be deployed in a pipeline that retrains and tunes hyperparameters each night to model click-through on data that goes stale every 24 hours

With the goal of decreasing the amount of time it takes to train these models, and ultimately to decrease costs, the Specialist wants to reconfigure the input hyperparameter range(s) Which visualization will accomplish this?

A. A histogram showing whether the most important input feature is Gaussian.

B. A scatter plot with points colored by target variable that uses (-Distributed Stochastic Neighbor Embedding (I-SNE) to visualize the large number of input variables in an easier-to-read dimension.

C. A scatter plot showing (he performance of the objective metric over each training iteration

D. A scatter plot showing the correlation between maximum tree depth and the objective metric.

Correct Answer: D

QUESTION 7

A Machine Learning Specialist was given a dataset consisting of unlabeled data The Specialist must create a model that can help the team classify the data into different buckets What model should be used to complete this work?

A. K-means clustering

B. Random Cut Forest (RCF)

C. XGBoost

D. BlazingText

Correct Answer: A

QUESTION 8

A chemical company has developed several machine learning (ML) solutions to identify chemical process abnormalities. The time series values of independent variables and the labels are available for the past 2 years and are sufficient to accurately model the problem.

The regular operation label is marked as 0. The abnormal operation label is marked as 1 . Process abnormalities have a significant negative effect on the companys profits. The company must avoid these abnormalities.

Which metrics will indicate an ML solution that will provide the GREATEST probability of detecting an abnormality?

A. Precision = 0.91 Recall = 0.6

B. Precision = 0.61 Recall = 0.98

C. Precision = 0.7 Recall = 0.9

D. Precision = 0.98 Recall = 0.8

Correct Answer: B

The metrics that will indicate an ML solution that will provide the greatest probability of detecting an abnormality are precision and recall. Precision is the ratio of true positives (TP) to the total number of predicted positives (TP + FP), where FP is false positives. Recall is the ratio of true positives (TP) to the total number of actual positives (TP + FN), where FN is false negatives. A high precision means that the ML solution has a low rate of false alarms, while a high recall means that the ML solution has a high rate of true detections. For the chemical company, the goal is to avoid process abnormalities, which are marked as 1 in the labels. Therefore, the company needs an ML solution that has a high recall for the positive class, meaning that it can detect most of the abnormalities and minimize the false negatives. Among the four options, option B has the highest recall for the positive class, which is 0.98. This means that the ML solution can detect 98% of the abnormalities and miss only 2%. Option B also has a reasonable precision for the positive class, which is 0.61. This means that the ML solution has a false alarm rate of 39%, which may be acceptable for the company, depending on the cost and benefit analysis. The other options have lower recall for the positive class, which means that they have higher false negative rates, which can be more detrimental for the company than false positive rates. References:

1: AWS Certified Machine Learning - Specialty uide

2: AWS Training - Machine Learning on AWS

3: AWS Whitepaper - An Overview of Machine Learning on AWS

4: Precision and recall

QUESTION 9

A Machine Learning Specialist is assigned to a Fraud Detection team and must tune an XGBoost model, which is working appropriately for test data. However, with unknown data, it is not working as expected. The existing parameters are provided as follows.

```
param = {
      'eta': 0.05,  # the training step for each iteration
      'silent': 1,  # logging mode - quiet
      'n_estimators':2000,
      'max_depth':30,
      'min_child_weight': 3,
      'gamma': 0,
      'subsample': 0.8,
      'objective': 'multi:softprob',  # error evaluation for multiclass training
      'num_class': 201}  # the number of classes that exist in this dataset
num_round = 60  # the number of training iterations
```

Which parameter tuning guidelines should the Specialist follow to avoid overfitting?

A. Increase the max_depth parameter value.

B. Lower the max_depth parameter value.

C. Update the objective to binary:logistic.

D. Lower the min_child_weight parameter value.

Correct Answer: B

QUESTION 10

A data scientist has developed a machine learning translation model for English to Japanese by using Amazon SageMaker\\'s built-in seq2seq algorithm with 500,000 aligned sentence pairs. While testing with sample sentences, the data scientist finds that the translation quality is reasonable for an example as short as five words. However, the quality becomes unacceptable if the sentence is 100 words long.

Which action will resolve the problem?

A. Change preprocessing to use n-grams.

B. Add more nodes to the recurrent neural network (RNN) than the largest sentence\\'s word count.

C. Adjust hyperparameters related to the attention mechanism.

D. Choose a different weight initialization type.

Correct Answer: C

QUESTION 11

An Machine Learning Specialist discover the following statistics while experimenting on a model.

Experiment 1
Baseline model
Train error = 5%
Test error = 16%

Experiment 2
The Specialist added more layers and neurons to the model and received the following results:
Train error = 5.2%
Test error = 15.7%

Experiment 3
The Specialist reverted back to the original number of neurons form Experiment 1 and implemented regularization in the neural network, which yielded the following results:
Train error = 4.7%
Test error = 9.5%

What can the Specialist from the experiments?

A. The model In Experiment 1 had a high variance error lhat was reduced in Experiment 3 by regularization Experiment 2 shows that there is minimal bias error in Experiment 1

B. The model in Experiment 1 had a high bias error that was reduced in Experiment 3 by regularization Experiment 2 shows that there is minimal variance error in Experiment 1

C. The model in Experiment 1 had a high bias error and a high variance error that were reduced in Experiment 3 by regularization Experiment 2 shows thai high bias cannot be reduced by increasing layers and neurons in the model

D. The model in Experiment 1 had a high random noise error that was reduced in Expenment 3 by regularization Expenment 2 shows that random noise cannot be reduced by increasing layers and neurons in the model

Correct Answer: C

QUESTION 12

A Machine Learning Specialist is building a supervised model that will evaluate customers\\' satisfaction with their mobile phone service based on recent usage The model\\'s output should infer whether or not a customer is likely to switch to a competitor in the next 30 days.

Which of the following modeling techniques should the Specialist use1?

A. Time-series prediction

B. Anomaly detection

C. Binary classification

D. Regression

Correct Answer: C

QUESTION 13

A healthcare company wants to create a machine learning (ML) model to predict patient outcomes. A data science team developed an ML model by using a custom ML library. The company wants to use Amazon SageMaker to train this

model. The data science team creates a custom SageMaker image to train the model. When the team tries to launch the custom image in SageMaker Studio, the data scientists encounter an error within the application.

Which service can the data scientists use to access the logs for this error?

A. Amazon S3

B. Amazon Elastic Block Store (Amazon EBS)

C. AWS CloudTrail

D. Amazon CloudWatch

Correct Answer: A

QUESTION 14

A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data.

The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards.

Which solution should the Data Scientist build to satisfy the requirements?

A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.

B. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and writes the data to a processed data location in Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.

C. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and inserts it into an Amazon RDS PostgreSQL database. Have the Analysts query and run dashboards from the RDS database.

D. Use Amazon Kinesis Data Analytics to ingest the streaming data and perform real-time SQL queries to convert the records to Apache Parquet before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.

Correct Answer: A

Firehose does integrate with GLue data catalog and it also "Buffers" the data .

"When Kinesis Data Firehose processes incoming events and converts the data to Parquet, it needs to know which schema to apply." This is achived by glue data catalog and athena and it works on real-time data ingest.See link below.

https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/

QUESTION 15

A company sells thousands of products on a public website and wants to automatically identify products with potential durability problems. The company has 1.000 reviews with date, star rating, review text, review summary, and customer email fields, but many reviews are incomplete and have empty fields. Each review has already been labeled with the correct durability result.

A machine learning specialist must train a model to identify reviews expressing concerns over product durability. The first model needs to be trained and ready to review in 2 days.

What is the MOST direct approach to solve this problem within 2 days?

A. Train a custom classifier by using Amazon Comprehend.

B. Build a recurrent neural network (RNN) in Amazon SageMaker by using Gluon and Apache MXNet.

C. Train a built-in BlazingText model using Word2Vec mode in Amazon SageMaker.

D. Use a built-in seq2seq model in Amazon SageMaker.

Correct Answer: A