Free Questions for MLS-C01 by vceexamstest

Shared by Hurst on 22-07-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

A pharmaceutical company performs periodic audits of clinical trial sites to quickly resolve critical findings. The company stores audit documents in text format. Auditors have requested help from a data science team to quickly analyze the documents. The auditors need to discover the 10 main topics within the documents to prioritize and distribute the review work among the auditing team members. Documents that describe adverse events must receive the highest priority.

A data scientist will use statistical modeling to discover abstract topics and to provide a list of the top words for each category to help the auditors assess the relevance of the topic.

Which algorithms are best suited to this scenario? (Choose two.)

Options:

A- Latent Dirichlet allocation (LDA)

B- Random Forest classifier

C- Neural topic modeling (NTM)

D- Linear support vector machine

E- Linear regression

Answer:

A, C

Explanation:

The algorithms that are best suited to this scenario are latent Dirichlet allocation (LDA) and neural topic modeling (NTM), as they are both unsupervised learning methods that can discover abstract topics from a collection of text documents.LDA and NTM can provide a list of the top words for each topic, as well as the topic distribution for each document, which can help the auditors assess the relevance and priority of the topic12.

The other options are not suitable because:

Option B: A random forest classifier is a supervised learning method that can perform classification or regression tasks by using an ensemble of decision trees.A random forest classifier is not suitable for discovering abstract topics from text documents, as it requires labeled data and predefined classes3.

Option D: A linear support vector machine is a supervised learning method that can perform classification or regression tasks by using a linear function that separates the data into different classes.A linear support vector machine is not suitable for discovering abstract topics from text documents, as it requires labeled data and predefined classes4.

Option E: A linear regression is a supervised learning method that can perform regression tasks by using a linear function that models the relationship between a dependent variable and one or more independent variables.A linear regression is not suitable for discovering abstract topics from text documents, as it requires labeled data and a continuous output variable5.

References:

1: Latent Dirichlet Allocation

2: Neural Topic Modeling

3: Random Forest Classifier

4: Linear Support Vector Machine

5: Linear Regression

Question 2

Question Type: MultipleChoice

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?

Options:

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

Answer:

B

Explanation:

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 Exam Guide

2: AWS Training - Machine Learning on AWS

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

4: Precision and recall

Question 3

Question Type: MultipleChoice

An automotive company uses computer vision in its autonomous cars. The company trained its object detection models successfully by using transfer learning from a convolutional neural network (CNN). The company trained the models by using PyTorch through the Amazon SageMaker SDK.

The vehicles have limited hardware and compute power. The company wants to optimize the model to reduce memory, battery, and hardware consumption without a significant sacrifice in accuracy.

Which solution will improve the computational efficiency of the models?

Options:

A- Use Amazon CloudWatch metrics to gain visibility into the SageMaker training weights, gradients, biases, and activation outputs. Compute the filter ranks based on the training information. Apply pruning to remove the low-ranking filters. Set new weights based on the pruned set of filters. Run a new training job with the pruned model.

B- Use Amazon SageMaker Ground Truth to build and run data labeling workflows. Collect a larger labeled dataset with the labelling workflows. Run a new training job that uses the new labeled data with previous training data.

C- Use Amazon SageMaker Debugger to gain visibility into the training weights, gradients, biases, and activation outputs. Compute the filter ranks based on the training information. Apply pruning to remove the low-ranking filters. Set the new weights based on the pruned set of filters. Run a new training job with the pruned model.

D- Use Amazon SageMaker Model Monitor to gain visibility into the ModelLatency metric and OverheadLatency metric of the model after the company deploys the model. Increase the model learning rate. Run a new training job.

Answer:

C

Explanation:

The solution C will improve the computational efficiency of the models because it uses Amazon SageMaker Debugger and pruning, which are techniques that can reduce the size and complexity of the convolutional neural network (CNN) models. The solution C involves the following steps:

Use Amazon SageMaker Debugger to gain visibility into the training weights, gradients, biases, and activation outputs. Amazon SageMaker Debugger is a service that can capture and analyze the tensors that are emitted during the training process of machine learning models. Amazon SageMaker Debugger can provide insights into the model performance, quality, and convergence.Amazon SageMaker Debugger can also help to identify and diagnose issues such as overfitting, underfitting, vanishing gradients, and exploding gradients1.

Compute the filter ranks based on the training information. Filter ranking is a technique that can measure the importance of each filter in a convolutional layer based on some criterion, such as the average percentage of zero activations or the L1-norm of the filter weights.Filter ranking can help to identify the filters that have little or no contribution to the model output, and thus can be removed without affecting the model accuracy2.

Apply pruning to remove the low-ranking filters. Pruning is a technique that can reduce the size and complexity of a neural network by removing the redundant or irrelevant parts of the network, such as neurons, connections, or filters.Pruning can help to improve the computational efficiency, memory usage, and inference speed of the model, as well as to prevent overfitting and improve generalization3.

Set the new weights based on the pruned set of filters. After pruning, the model will have a smaller and simpler architecture, with fewer filters in each convolutional layer.The new weights of the model can be set based on the pruned set of filters, either by initializing them randomly or by fine-tuning them from the original weights4.

Run a new training job with the pruned model. The pruned model can be trained again with the same or a different dataset, using the same or a different framework or algorithm. The new training job can use the same or a different configuration of Amazon SageMaker, such as the instance type, the hyperparameters, or the data ingestion mode.The new training job can also use Amazon SageMaker

Debugger to monitor and analyze the training process and the model quality5.

The other options are not suitable because:

Option A: Using Amazon CloudWatch metrics to gain visibility into the SageMaker training weights, gradients, biases, and activation outputs will not be as effective as using Amazon SageMaker Debugger. Amazon CloudWatch is a service that can monitor and observe the operational health and performance of AWS resources and applications. Amazon CloudWatch can provide metrics, alarms, dashboards, and logs for various AWS services, including Amazon SageMaker. However, Amazon CloudWatch does not provide the same level of granularity and detail as Amazon SageMaker Debugger for the tensors that are emitted during the training process of machine learning models.Amazon CloudWatch metrics are mainly focused on the resource utilization and the training progress, not on the model performance, quality, and convergence6.

Option B: Using Amazon SageMaker Ground Truth to build and run data labeling workflows and collecting a larger labeled dataset with the labeling workflows will not improve the computational efficiency of the models. Amazon SageMaker Ground Truth is a service that can create high-quality training datasets for machine learning by using human labelers. A larger labeled dataset can help to improve the model accuracy and generalization, but it will not reduce the memory, battery, and hardware consumption of the model.Moreover, a larger labeled dataset may increase the training time and cost of the model7.

Option D: Using Amazon SageMaker Model Monitor to gain visibility into the ModelLatency metric and OverheadLatency metric of the model after the company deploys the model and increasing the model learning rate will not improve the computational efficiency of the models. Amazon SageMaker Model Monitor is a service that can monitor and analyze the quality and performance of machine learning models that are deployed on Amazon SageMaker endpoints. The ModelLatency metric and the OverheadLatency metric can measure the inference latency of the model and the endpoint, respectively. However, these metrics do not provide any information about the training weights, gradients, biases, and activation outputs of the model, which are needed for pruning. Moreover, increasing the model learning rate will not reduce the size and complexity of the model, but it may affect the model convergence and accuracy.

References:

1: Amazon SageMaker Debugger

2: Pruning Convolutional Neural Networks for Resource Efficient Inference

3: Pruning Neural Networks: A Survey

4: Learning both Weights and Connections for Efficient Neural Networks

5: Amazon SageMaker Training Jobs

6: Amazon CloudWatch Metrics for Amazon SageMaker

7: Amazon SageMaker Ground Truth

: Amazon SageMaker Model Monitor

Question 4

Question Type: **MultipleChoice**

A data scientist is working on a forecast problem by using a dataset that consists of .csv files that are stored in Amazon S3. The files contain a timestamp variable in the following format:

March 1st, 2020, 08:14pm -

There is a hypothesis about seasonal differences in the dependent variable. This number could be higher or lower for weekdays because some days and hours present varying values, so the day of the week, month, or hour could be an important factor. As a result, the data scientist needs to transform the timestamp into weekdays, month, and day as three separate variables to conduct an analysis.

Which solution requires the LEAST operational overhead to create a new dataset with the added features?

Options:

A- Create an Amazon EMR cluster. Develop PySpark code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3.

B- Create a processing job in Amazon SageMaker. Develop Python code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3.

C- Create a new flow in Amazon SageMaker Data Wrangler. Import the S3 file, use the Featurize date/time transform to generate the new variables, and save the dataset as a new file in Amazon S3.

D- Create an AWS Glue job. Develop code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3.

Answer:

C

Explanation:

The solution C will create a new dataset with the added features with the least operational overhead because it uses Amazon SageMaker Data Wrangler, which is a service that simplifies the process of data preparation and feature engineering for machine learning. The solution C involves the following steps:

Create a new flow in Amazon SageMaker Data Wrangler. A flow is a visual representation of the data preparation steps that can be applied to one or more datasets.The data scientist can create a new flow in the Amazon SageMaker Studio interface and import the S3 file as a data source1.

Use the Featurize date/time transform to generate the new variables. Amazon SageMaker Data Wrangler provides a set of preconfigured transformations that can be applied to the data with a few clicks. The Featurize date/time transform can parse a date/time column and generate new columns for the year, month, day, hour, minute, second, day of week, and day of year.The data scientist can use this transform to create the new variables from the timestamp variable2.

Save the dataset as a new file in Amazon S3. Amazon SageMaker Data Wrangler can export the transformed dataset as a new file in Amazon S3, or as a feature store in Amazon SageMaker Feature Store.The data scientist can choose the output format and location of the new file3.

The other options are not suitable because:

Option A: Creating an Amazon EMR cluster and developing PySpark code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3 will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist will have to manage the Amazon EMR cluster, the PySpark application, and the data storage.Moreover, the data scientist will have to write custom code for the date/time parsing and feature generation, which may require more development effort and testing4.

Option B: Creating a processing job in Amazon SageMaker and developing Python code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3 will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist will have to manage the processing job, the Python code, and the data storage.Moreover, the data scientist will have to write custom code for the date/time parsing and feature generation, which may require more development effort and testing5.

Option D: Creating an AWS Glue job and developing code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3 will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist will have to manage the AWS Glue job, the code, and the data storage.Moreover, the data scientist will have to write custom code for the date/time parsing and feature generation, which may require more development effort and testing6.

References:

1: Amazon SageMaker Data Wrangler

2: Featurize Date/Time - Amazon SageMaker Data Wrangler

3: Exporting Data - Amazon SageMaker Data Wrangler

4: Amazon EMR

5: Processing Jobs - Amazon SageMaker

6: AWS Glue

Question 5

Question Type: MultipleChoice

A media company is building a computer vision model to analyze images that are on social medi

a. The model consists of CNNs that the company trained by using images that the company stores in Amazon S3. The company used an Amazon SageMaker training job in File mode with a single Amazon EC2 On-Demand Instance.

Every day, the company updates the model by using about 10,000 images that the company has collected in the last 24 hours. The company configures training with only one epoch. The company wants to speed up training and lower costs without the need to make any code changes.

Which solution will meet these requirements?

Options:

A- Instead of File mode, configure the SageMaker training job to use Pipe mode. Ingest the data from a pipe.

B- Instead Of File mode, configure the SageMaker training job to use FastFile mode with no Other changes.

C- Instead Of On-Demand Instances, configure the SageMaker training job to use Spot Instances. Make no Other changes.

D- Instead Of On-Demand Instances, configure the SageMaker training job to use Spot Instances. Implement model checkpoints.

Answer:

C

Explanation:

The solution C will meet the requirements because it uses Amazon SageMaker Spot Instances, which are unused EC2 instances that are available at up to 90% discount compared to On-Demand prices. Amazon SageMaker Spot Instances can speed up training and lower costs by taking advantage of the spare EC2 capacity. The company does not need to make any code changes to use Spot Instances, as it can simply enable the managed spot training option in the SageMaker training job configuration.The company also does not need to implement model checkpoints, as it is using only one epoch for training, which means the model will not resume from a previous state1.

The other options are not suitable because:

Option A: Configuring the SageMaker training job to use Pipe mode instead of File mode will not speed up training or lower costs significantly. Pipe mode is a data ingestion mode that streams data directly from S3 to the training algorithm, without copying the data to the local storage of the training instance. Pipe mode can reduce the startup time of the training job and the disk space usage, but it does not affect the computation time or the instance price.Moreover, Pipe mode may require some code changes to handle the streaming data, depending on the training algorithm2.

Option B: Configuring the SageMaker training job to use FastFile mode instead of File mode will not speed up training or lower costs significantly. FastFile mode is a data ingestion mode that copies data from S3 to the local storage of the training instance in parallel with the training process. FastFile mode can reduce the startup time of the training job and the disk space usage, but it does not affect the computation time or the instance price.Moreover, FastFile mode is only available for distributed training jobs that use multiple instances, which is not the case for the company3.

Option D: Configuring the SageMaker training job to use Spot Instances and implementing model checkpoints will not meet the requirements without the need to make any code changes. Model checkpoints are a feature that allows the training job to save the model state periodically to S3, and resume from the latest checkpoint if the training job is interrupted.Model checkpoints can help to avoid losing the training progress and ensure the model convergence, but they require some code changes to implement the checkpointing logic and the resuming logic4.

References:

1: Managed Spot Training - Amazon SageMaker

2: Pipe Mode - Amazon SageMaker

3: FastFile Mode - Amazon SageMaker

4: Checkpoints - Amazon SageMaker

Question 6

Question Type: **MultipleChoice**

A company wants to conduct targeted marketing to sell solar panels to homeowners. The company wants to use machine learning (ML) technologies to identify which houses already have solar panels. The company has collected 8,000 satellite images as training data and will use Amazon SageMaker Ground Truth to label the data.

The company has a small internal team that is working on the project. The internal team has no ML expertise and no ML experience.

Which solution will meet these requirements with the LEAST amount of effort from the internal team?

Options:

A- Set up a private workforce that consists of the internal team. Use the private workforce and the SageMaker Ground Truth active learning feature to label the data. Use Amazon Rekognition Custom Labels for model training and hosting.

B- Set up a private workforce that consists of the internal team. Use the private workforce to label the data. Use Amazon Rekognition Custom Labels for model training and hosting.

C- Set up a private workforce that consists of the internal team. Use the private workforce and the SageMaker Ground Truth active learning feature to label the data. Use the SageMaker Object Detection algorithm to train a model. Use SageMaker batch transform for inference.

D- Set up a public workforce. Use the public workforce to label the data. Use the SageMaker Object Detection algorithm to train a model. Use SageMaker batch transform for inference.

Answer:

A

Explanation:

The solution A will meet the requirements with the least amount of effort from the internal team because it uses Amazon SageMaker Ground Truth and Amazon Rekognition Custom Labels, which are fully managed services that can provide the desired functionality. The solution A involves the following steps:

Set up a private workforce that consists of the internal team. Use the private workforce and the SageMaker Ground Truth active learning feature to label the data. Amazon SageMaker Ground Truth is a service that can create high-quality training datasets for machine learning by using human labelers. A private workforce is a group of labelers that the company can manage and control. The internal team can use the private workforce to label the satellite images as having solar panels or not.The SageMaker Ground Truth active learning feature can reduce the labeling effort by using a machine learning model to automatically label the easy examples and only send the difficult ones to the human labelers1.

Use Amazon Rekognition Custom Labels for model training and hosting. Amazon Rekognition Custom Labels is a service that can train and deploy custom machine learning models for image analysis. Amazon Rekognition Custom Labels can use the labeled data from SageMaker Ground Truth to train a model that can detect solar panels in satellite images.Amazon Rekognition Custom Labels can also host the model and provide an API endpoint for inference2.

The other options are not suitable because:

Option B: Setting up a private workforce that consists of the internal team, using the private workforce to label the data, and using Amazon Rekognition Custom Labels for model training and hosting will incur more effort from the internal team than using SageMaker Ground Truth active learning feature.The internal team will have to label all the images manually, without the assistance of the machine learning model that can automate some of the labeling tasks1.

Option C: Setting up a private workforce that consists of the internal team, using the private workforce and the SageMaker Ground Truth active learning feature to label the data, using the SageMaker Object Detection algorithm to train a model, and using SageMaker batch transform for inference will incur more operational overhead than using Amazon Rekognition Custom Labels. The company will have to

manage the SageMaker training job, the model artifact, and the batch transform job.Moreover, SageMaker batch transform is not suitable for real-time inference, as it processes the data in batches and stores the results in Amazon S33.

Option D: Setting up a public workforce, using the public workforce to label the data, using the SageMaker Object Detection algorithm to train a model, and using SageMaker batch transform for inference will incur more operational overhead and cost than using a private workforce and Amazon Rekognition Custom Labels. A public workforce is a group of labelers from Amazon Mechanical Turk, a crowdsourcing marketplace. The company will have to pay the public workforce for each labeling task, and it may not have full control over the quality and security of the labeled data.The company will also have to manage the SageMaker training job, the model artifact, and the batch transform job, as explained in option C4.

References:

1: Amazon SageMaker Ground Truth

2: Amazon Rekognition Custom Labels

3: Amazon SageMaker Object Detection

4: Amazon Mechanical Turk

Question 7

Question Type: **MultipleChoice**

A data science team is working with a tabular dataset that the team stores in Amazon S3. The team wants to experiment with different feature transformations such as categorical feature encoding. Then the team wants to visualize the resulting distribution of the dataset. After the team finds an appropriate set of feature transformations, the team wants to automate the workflow for feature transformations.

Which solution will meet these requirements with the MOST operational efficiency?

Options:

A- Use Amazon SageMaker Data Wrangler preconfigured transformations to explore feature transformations. Use SageMaker Data Wrangler templates for visualization. Export the feature processing workflow to a SageMaker pipeline for automation.

B- Use an Amazon SageMaker notebook instance to experiment with different feature transformations. Save the transformations to Amazon S3. Use Amazon QuickSight for visualization. Package the feature processing steps into an AWS Lambda function for automation.

C- Use AWS Glue Studio with custom code to experiment with different feature transformations. Save the transformations to Amazon S3. Use Amazon QuickSight for visualization. Package the feature processing steps into an AWS Lambda function for automation.

D- Use Amazon SageMaker Data Wrangler preconfigured transformations to experiment with different feature transformations. Save the transformations to Amazon S3. Use Amazon QuickSight for visualzation. Package each feature transformation step into a separate AWS Lambda function. Use AWS Step Functions for workflow automation.

Answer:

A

Explanation:

The solution A will meet the requirements with the most operational efficiency because it uses Amazon SageMaker Data Wrangler, which is a service that simplifies the process of data preparation and feature engineering for machine learning. The solution A involves the following steps:

Use Amazon SageMaker Data Wrangler preconfigured transformations to explore feature transformations. Amazon SageMaker Data Wrangler provides a visual interface that allows data scientists to apply various transformations to their tabular data, such as encoding categorical features, scaling numerical features, imputing missing values, and more.Amazon SageMaker Data Wrangler also supports custom transformations using Python code or SQL queries1.

Use SageMaker Data Wrangler templates for visualization. Amazon SageMaker Data Wrangler also provides a set of templates that can generate visualizations of the data, such as histograms, scatter plots, box plots, and more.These visualizations can help data scientists to understand the distribution and characteristics of the data, and to compare the effects of different feature transformations1.

Export the feature processing workflow to a SageMaker pipeline for automation. Amazon SageMaker Data Wrangler can export the feature processing workflow as a SageMaker pipeline, which is a service that orchestrates and automates machine learning workflows. A SageMaker pipeline can run the feature processing steps as a preprocessing step, and then feed the output to a training step or an inference step.This can reduce the operational overhead of managing the feature processing workflow and ensure its consistency and reproducibility2.

The other options are not suitable because:

Option B: Using an Amazon SageMaker notebook instance to experiment with different feature transformations, saving the transformations to Amazon S3, using Amazon QuickSight for visualization, and packaging the feature processing steps into an AWS Lambda function for automation will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist

will have to write the code for the feature transformations, the data storage, the data visualization, and the Lambda function.Moreover, AWS Lambda has limitations on the execution time, memory size, and package size, which may not be sufficient for complex feature processing tasks3.

Option C: Using AWS Glue Studio with custom code to experiment with different feature transformations, saving the transformations to Amazon S3, using Amazon QuickSight for visualization, and packaging the feature processing steps into an AWS Lambda function for automation will incur more operational overhead than using Amazon SageMaker Data Wrangler. AWS Glue Studio is a visual interface that allows data engineers to create and run extract, transform, and load (ETL) jobs on AWS Glue. However, AWS Glue Studio does not provide preconfigured transformations or templates for feature engineering or data visualization. The data scientist will have to write custom code for these tasks, as well as for the Lambda function.Moreover, AWS Glue Studio is not integrated with SageMaker pipelines, and it may not be optimized for machine learning workflows4.

Option D: Using Amazon SageMaker Data Wrangler preconfigured transformations to experiment with different feature transformations, saving the transformations to Amazon S3, using Amazon QuickSight for visualization, packaging each feature transformation step into a separate AWS Lambda function, and using AWS Step Functions for workflow automation will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist will have to create and manage multiple AWS Lambda functions and AWS Step Functions, which can increase the complexity and cost of the solution.Moreover, AWS Lambda and AWS Step Functions may not be compatible with SageMaker pipelines, and they may not be optimized for machine learning workflows5.

References:

1: Amazon SageMaker Data Wrangler

2: Amazon SageMaker Pipelines

3: AWS Lambda

4: AWS Glue Studio

5: AWS Step Functions