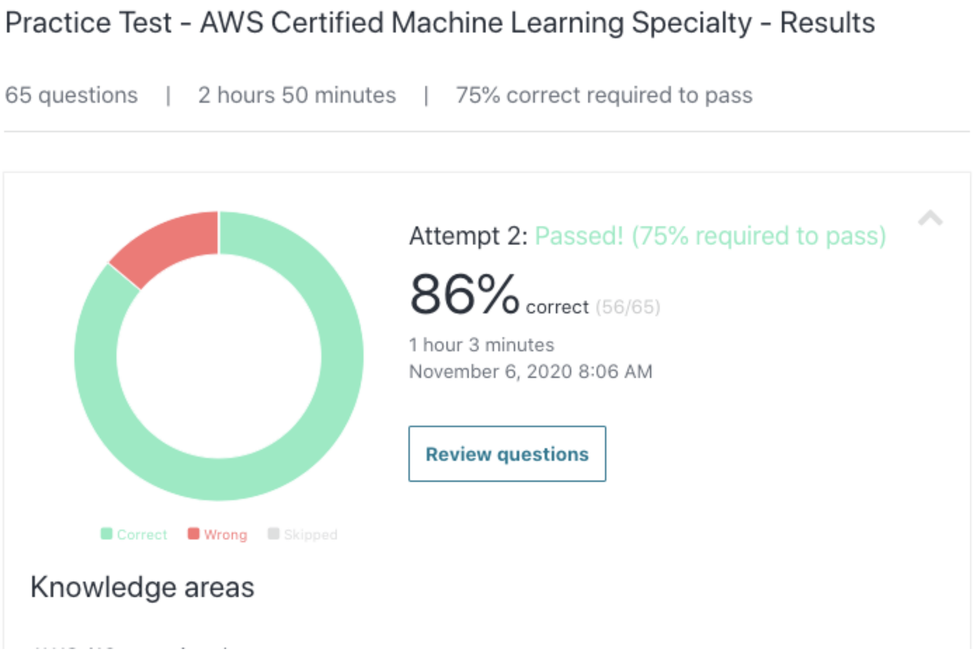
https://www.udemy.com/course/aws-machine-learning-a-complete-guide-with-python/learn/quiz/4774522/results?expanded=404584824#overview



Practice Test - AWS Certified Machine Learning Specialty - Results

65 questions | 2 hours 50 minutes | 75% correct required to pass

Attempt 2: Passed! (75% required to pass)

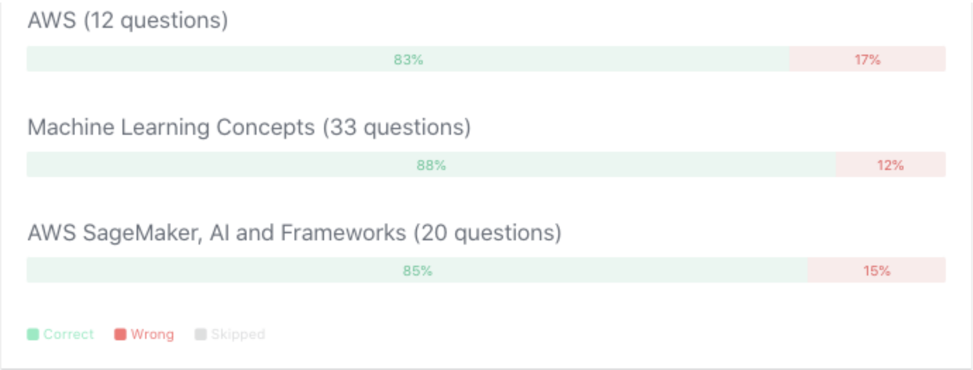
86% correct (56/65)

1 hour 3 minutes
November 6, 2020 8:06 AM

Review questions

🟩 Correct 🟥 Wrong ▨ Skipped

Knowledge areas



AWS (12 questions)

Correct	Wrong
83%	17%

Machine Learning Concepts (33 questions)

Correct	Wrong
88%	12%

AWS SageMaker, AI and Frameworks (20 questions)

Correct	Wrong
85%	15%

🟩 Correct 🟥 Wrong ▨ Skipped

Practice Test - AWS Certified Machine Learning Specialty - Results

Question 1: Correct

Which one of the services may be impacted when a single availability zone goes down in an AWS region?

-
- Artificial Intelligence Services like Rekognition
-
- S3
-
- SageMaker Endpoint with multiple instances
-
- SageMaker Endpoint with a single instance
    - **(Correct)**

Explanation

Each AWS region consists of three or more availability zones. Availability Zones are physically separate infrastructure. Among the choices presented, a SageMaker Endpoint that has only one instance to handle inference requests may be impacted if that instance is running in that Availability Zone. To improve Availability, for production workloads, you need to use at least two instances behind a SageMaker Endpoint – SageMaker will ensure that the instances are deployed in different availability zones. S3 automatically replicates data in three or more availability zones, and S3 can transparently handle availability zone failure. Managed AI Services like Rekognition is also multi-availability zone enabled and can handle availability zone failures automatically

Question 2: Correct

A machine learning specialist needs to come up with an approach to automatically summarize the content of large text documents. Which algorithm can be used for this use case?

-
- K-Means
-
- Seq2Seq
  - **(Correct)**
-
- LDA
-
- Random Cut Forest

Explanation

Seq2Seq algorithm is used for text summarization – It accepts a series of tokens as input and outputs another sequence of tokens. LDA is an unsupervised algorithm for topic modeling – it can generate probabilities of a document belonging to a number of specified topics. K-Means is a clustering algorithm that is used for identifying grouping within data. Random Cut Forest is used for detecting anomalous data points

Question 3: **Correct**

A team of students is building an application that can blur or remove unwanted objects in an image. Users can pick the objects on which action needs to be performed.

Which one of the AWS machine learning capabilities can you use for this?

-
- ImageClassification
-
- ObjectDetection
-
- Rekognition

-
- Semantic Segmentation
  - **(Correct)**

Explanation

The semantic Segmentation algorithm is useful for this use case – it can detect objects in an image, shape of each object along with location and pixels that are part of the object.

ImageClassification algorithm is used for classifying a whole image.

ObjectDetection is used for detecting objects and classifying them; you can also get a bounding box for each object – however, this algorithm does not gives you information about the exact shape of the object.

Rekognition AI Service can help you analyze images and videos; however, object shape detection is not one of the capabilities

Question 4: **Correct**

A machine learning specialist is using a SageMaker algorithm to train a model. The dataset is large, and the training job is distributed across multiple training instances. What mechanism does SageMaker provide to minimize temporary storage required in the training instance volumes?

-
- File Mode
-
- Explore compressed storage
-
- Pipe Mode
  - **(Correct)**
-

- SageMaker does not copy data to local instance volumes – all data resides in S3

Explanation

In Pipe Mode, training job streams data from S3 to your training instance.

Streaming can provide faster start times and better throughput. It also reduces the storage needed on your training instances as you need storage only for the final model artifacts.

In File mode, training job copies entire data from S3 to your training instance volumes.

So, you would need to allocate enough disk space in your training instances to store your full training dataset and for the final model artifacts

Question 5: Correct

A data scientist has a large dataset that needs to be trained on the AWS SageMaker service. The training algorithm is optimized for GPU processing and can benefit from substantial speed-up when trained on instances with GPUs. Which instance family can you use for a training job for the best performance?

-
- Compute Optimized family
-
- Memory-Optimized family
-
- Accelerated Computing family
    - **(Correct)**
-
- General Purpose family

Explanation

Accelerated computing family (P and G type instances) come with GPUs, and these are ideal for algorithms that are optimized for GPUs.

General Purpose family are some of the lowest cost instances and offer balanced performance and memory configuration (T and M type instances).

Compute Optimized family comes with the latest generation CPUs and is a higher performance system. These are suitable for CPU intensive model training and hosting (C type instances).

Memory-optimized family are optimized for workloads that process large datasets in memory (R type instances).

Besides, the sagemaker also has Elastic Inference Acceleration (partial GPUs) that provides fractional GPU capacity at a fraction of the cost of accelerated computing family.

Elastic inference Acceleration is suitable for inference workloads that can benefit from GPUs and can be easily added to other instance families.

Question 6: **Correct**

A company has received an email from a customer with product feedback. Feedback is in an unknown language, and the company’s product team has requested a German version of the email.

What steps are needed to accomplish this?

-
- Translate to English with source language set to auto-detect and then translate the output to German

-
- Transcribe to English, Translate to German
-
- Translate to German with source language set to auto-detect
  - **(Correct)**
-
- Transcribe to German with Source language set to auto-detect

Explanation

Translate to German with Source Language set to auto-detect. Translate service can auto-detect source language and convert it to a target language. However, both source and target languages must be on the supported list.
https://docs.aws.amazon.com/translate/latest/dg/how-it-works.html

Question 7: **Incorrect**

A data scientist is exploring the use of the XGBoost algorithm for a regression problem.

The dataset consists of numeric features.

Some of the features are highly correlated, and almost all the features are on different orders of magnitude.

What data-transformation is required to train on XGBoost?

-
- Data transformation is not needed for this dataset
  - **(Correct)**
-
- Normalization
-
- Remove one feature from every highly correlated feature pairs
-
- Scaling

    - **(Incorrect)**

Explanation

Decision Tree-based algorithms like XGBoost automatically handles correlated features, numeric features on a different scale, and numeric-categorical variables. Other algorithms like a neural network and the linear model would require features on a similar scale and range, and you need to keep only one feature in every highly correlated feature pairs and one-hot encode categorical features.

Question 8: Correct

When training a deep learning model, if you increase the batch size, you should also

-
- Increase the learning rate
    - **(Correct)**
-
- Decrease the learning rate
-
- Keep the learning rate same as batch size
-
- Learning rate and batch size are independent of each other

Explanation

As described in the below article, batch size and learning rate should be adjusted by the same factor. In a deep learning network, the loss curve is very complex and has several local minima. Imagine you need to find the deepest point in a large land area, and you don't know where the deepest point is or how deep it is. You would come across smaller valleys (local minima), and we don't want to conclude a small valley as the deepest point incorrectly. Our goal is for the optimization algorithm to explore different areas to find the deepest valley (global minima). Large batch sizes appear to cause the model to get stuck in local minima whereas smaller batch sizes make the algorithm jump out of local minima

and go towards global minima. To minimize the effect of large batches, when increasing batch size, you also need to increase the learning rate by the same factor. Similarly, if you decrease the batch size, you need to reduce the learning rate by the same factor.
Reference:
https://aws.amazon.com/blogs/machine-learning/the-importance-of-hyperparameter-tuning-for-scaling-deep-learning-training-to-multiple-gpus/

Question 9: Correct

An organization is consolidating data in S3, and data scientists need access to this data for initial exploration. They are well versed in SQL and would prefer to access the data in S3 using SQL. Which of these options provides the lowest cost without requiring to provision any servers?

-
- EMR Hive
-
- Athena
    - **(Correct)**
-
- Redshift Spectrum
-
- EMR Spark

Explanation

With Athena, you can query the data in S3 using SQL. You can either create the table structure in Glue Catalog or let the Glue Crawler collect the metadata and create the table. Athena automatically provisions the resources required for running queries. Redshift Spectrum also provides a capability similar to Athena; however, the query is executed in your Cluster. So, you would need to provision the servers. EMR Hive and Spark are also good options, but it would require provisioning your cluster, and you would also need to figure out how to load the data to the cluster.

Question 10: Incorrect

A utility company wants to forecast water consumption per household. The historical data set contains the following attributes:

* **Year - Numeric**

* **Month - Numeric**

* **Floor Size SqFt – numeric**

* **Lot Size SqFt - numeric**

* **Number of Bathrooms – numeric**

* **Lawn – categorical with values YES or NO**

* **Consumption – numeric (target)**

To train using XGBoost, what data transformation step do you need to perform?

-
- Transform non-numeric categories to equivalent numeric categories
  - **(Correct)**
-
- One-Hot encode categorical features
  - **(Incorrect)**
-
- Scale all numeric features to similar range and scale
-
- Normalize all numeric features

Explanation

XGBoost requires all numeric features. Tree-based algorithms can handle features with different scales. It also handles numeric categorical features (does not require one-hot encoding). However, XGBoost also supports One-hot encoded features. For a binary feature like Lawn (yes or no), only label encoding is needed (i.e., convert to 0 and 1). You should not perform one-hot encoding on binary features. For non-binary categorical features, you can test with label encoding first and then optionally, test performance with one-hot encoding.

Question 11: Correct

The human level error rate is 2%, and the model training error rate is 8%. What steps can you take to optimize the model? (Choose Three)

-
- Build a more complex model
  - **(Correct)**
-
- Increase regularization
-
- New neural network architecture
  - **(Correct)**
-
- Train longer
  - **(Correct)**

Explanation

Since the gap between human-level performance and training error is large, the model is underfitting. When a model underfits, it is not learning from training data. To fix the high training error, you can increase the model complexity, train the model longer (more epochs), and use a different network architecture. However, increasing regularization would reduce the model complexity (by suppressing the importance of features), and it will underfit more. Reference: NIPS 2016 tutorial: Nuts and bolts of building AI applications by Dr. Andrew Ng

Question 12: Correct

You are using AWS provided services for maintaining metadata about your data files stored in S3. The incoming files to S3 have additional attributes that are collected, and they are not showing up in the metadata. What is the recommended approach to address this issue?

-
- Ensure Athena queries are scheduled to run periodically to update metadata
-
- Create a new table in the Glue Catalog to capture the changes
-
- Configure the Lambda function to monitor S3 and to capture the metadata changes
-
- Ensure Glue Crawlers are configured as a scheduled job to scan the files and update metadata
    - **(Correct)**

Explanation

Glue Crawler is used for automated collection and maintenance of metadata in the Glue Catalog. You would need to configure the Crawler to periodically scan the source data to detect any structural changes in the files and keep the metadata in sync with data. With Athena service, you can query files in S3 using SQL – however, the metadata about the files needs to be created in Glue Catalog first. Lambda function is used for do-it-yourself catalog management. This requires more effort when compared to using Glue Crawler

Question 13: Incorrect

For a binary classification problem, the cost of misclassifying a positive sample is three times more than the cost of misclassifying a negative example.

Which model has the lowest cost with at least 60% recall?

Model 1 – TP: 10, FN: 5, TN: 25, FP: 10

Model 2 – TP: 5, FN: 10, TN: 20, FP: 15

Model 3 – TP: 1, FN: 14, TN: 30, FP: 5

Model 4 – TP: 9, FN: 6, TN: 20, FP: 15

-
- Model 1
  - **(Correct)**
-
- Model 3
-
- Model 2
-
- Model 4
  - **(Incorrect)**

Explanation

Recall needs to be at least 60%. Recall = TP/(TP+FN).

Model 1 recall is over 0.6, and Model 2 recall is 0.3; model 3 recall is 1/15 and is a small value and Model 4 recall is 0.6. So, the answer has to be either Model 1 or Model 4.

The cost of misclassifying a positive sample is three times more than misclassifying a negative sample.

Total Cost = 3 * FN + 1 * FP

Model 1 cost = 3*5 + 10 = 25. Model 4 cost is = 3 * 6 + 1 * 15 = 18 + 15 = 33.

So, Model 1 has the lowest cost while meeting 0.6 recall.

Question 14: Correct

The training error is low, but the test error is high. Among the choices presented, which one of these options can correct the issue? (Choose Three)

-
- Decrease regularization
-
- Increase the number of epochs
-
- Train with more data
  - **(Correct)**
-
- New neural network architecture
  - **(Correct)**
-
- Increase Regularization
  - **(Correct)**

Explanation

The training error is low, and the test error is high. So, the model has a variance problem. The model is overfitting the training data, or the data distribution between test and train data sets is different. You can train with more data to handle issues with different data distribution between train and test data sets. More data is also useful if the model is not detecting all the patterns. If you suspect overfitting, you can also increase regularization to simplify the model. Finally, you can also try different neural network architecture (for example reduce number of neurons, change the number of hidden layers and so forth). However, decreasing regularization would cause the model to overfit more. You can reduce the number of epochs to minimize variance. This would prevent the model from memorizing too much about training data.

Question 15: Correct

You are training a model to predict the probability of leaving the mobile operator. You would like to assess the quality of the metrics at various cut-off thresholds.

Which metric gives you insight into the model performance over a range of tradeoffs between true positive rate and false-positive rate?

-
- F1 Score
-
- Squared Error
-
- Accuracy
-
- ROC AUC Metric
  - **(Correct)**

Explanation

Receiver Operating Characteristic (ROC) curve compares the true positive rate and the false positive rate at different thresholds. AUC metrics measure the area formed by such a curve, and the ROC AUC is used for summarizing the model performance with a range of tradeoffs

Question 16: Correct

Your company has a portfolio of machine learning models that are used by web applications and mobile apps. What is the best mechanism to integrate machine learning models with your application? The solution also needs to scale on demand.

-
- API Gateway, Lambda, SageMaker Endpoint with Auto Scaling
  - **(Correct)**
-
- Host your models on EC2 web server instances, and load balance using Elastic Load Balancing. Setup autoscaling to scale web servers
-
- Use Lambda function to invoke machine learning models and invoke the Lambda function from the client application
-
- Invoke Machine Learning model endpoint from your Client application

Explanation

To streamline access to the backend services, you can API Gateway. You can use API Gateway as a Gatekeeper to ensure only authorized users and applications have access to the services. Configure API Gateway to invoke the Lambda function. Lambda function, in turn, invokes SageMaker endpoint. You need to configure Autoscaling to ensure SageMaker endpoint scales on demand. This approach also makes it easy to try different versions of Machine Learning Models without requiring code changes in the client application.

Question 17: **Correct**

An organization has human experts who perform manual classification of products by visual inspection. A Machine Learning specialist is building a classification system to match human-level performance. When reviewing the error rate of humans, the specialist observes the following:

Newly trained employees had a misclassification error rate of 5%, Experienced employee had an error rate of 2.5%, and when a team of experienced employees worked together, they had a misclassification rate of 1%.

What should be considered as human-level performance?

-
- 1%
    - **(Correct)**
-
- 5%
-
- 2.5%
-
- Average of the error rates

Explanation

1% should be used as the human-level performance and it is a good proxy for Bayes optimal error (theoretical best possible error rate).

Reference: NIPS 2016 tutorial: Nuts and bolts of building AI applications by Dr. Andrew Ng (Starting at 1:24:00) https://www.youtube.com/watch?v=wjqaz6m42wU

Question 18: Correct

What privileges does a newly created Identity and Access Management (IAM) user have? This User does not have any policy attached and does not belong to any IAM Groups.

-
- Read-only access to all resources in your account
-
- Read-only access in the region where IAM user was created
-
- Read-Write access to all resources in your account
-
- User cannot access AWS resources until explicit allow access is granted
    - **(Correct)**

Explanation

When you create a new IAM user without attaching any policies, the user is not allowed access to any AWS resource. User needs to be granted permissions by assigning policies or by adding them to a Group with necessary permissions. IAM is a global resource – when you create a policy, role, user, or group, they can granted permission to AWS resources in any region.

Question 19: Correct

You are using SageMaker Automatic Hyperparameter tuning to search for optimal parameters for a learning algorithm.

What are the best practices when running a hyperparameter tuning job? (Choose three)

-
- Use Linear Scaling for hyperparameter that spans several orders of magnitude
-
- Use Logarithmic Scaling for hyperparameter that spans several orders of magnitude
  - **(Correct)**
-
- Use fewer concurrent tuning jobs
  - **(Correct)**
-
- Configure the tuning job to explore all hyperparameters supported by the algorithm
-
- Configure the tuning job to search a smaller number of hyperparameters
  - **(Correct)**

Explanation

“you can simultaneously use up to 20 variables in a hyperparameter tuning job, limiting your search to a much smaller number is likely to give better results”

“a tuning job improves only through successive rounds of experiments. Typically, running one training job at a time achieves the best results with the least amount of compute time”

“Choose logarithmic scaling when you are searching a range that spans several orders of magnitude”

“you specify a range of values between .0001 and 1.0 for the learning_rate hyperparameter, searching uniformly on a logarithmic scale gives you a better sample of the entire range than searching on a linear scale would, because searching on a linear scale would, on average, devote 90 percent of your training budget to only the values between .1 and 1.0, leaving only 10 percent of your training budget for the values between .0001 and .1” https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-define-ranges.html

https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-considerations.html

Question 20: Correct

A model has the following errors: Training Error is 2%, Test Error is 5%. The benchmark is human-level performance, and the human error is 1%.

The model is:

-
- Performing close to human-level performance
-
- Normal
-
- Underfitting
-
- Overfitting
    - **(Correct)**

Explanation

Here, the model training error is comparable to human-level performance. So, the model is doing well with training data. However, it is not generalizing well for unseen data, and the test error is much larger. So, it is showing signs of overfitting. (memorized too much about training data)

Question 21: Correct

A dataset consists of following features along with the type of values it can contain

*** DayOfWeek – Sunday, Monday, Tuesday and so forth**

*** Holiday – True or False**

*** Temperature – in Fahrenheit**

*** Humidity – 0 to 100**

*** Precipitation – 0 to 100**

*** Windspeed – 0 to 150**

*** Pollen – 0 to 1**

*** AirQuality – Good, Bad**

AirQuality is the label

The Machine Learning Analyst is planning to compare a variety of algorithms and would like to reuse the same transformed dataset for training and testing.

What data transformation is recommended? (Select Three)

-
- Scale Temperature, Humidity, Precipitation, Windspeed, Pollen features
  - **(Correct)**
-
- Transform using Principal Component Analysis
-
- Use numeric data without any transformation, and one hot encode categorical features
-
- Label encode AirQuality and Holiday features
  - **(Correct)**
-
- One-Hot encode Day of Week

- **(Correct)**

Explanation

The categorical features need to be one-hot encoded for algorithms like the linear model and neural network.

XGBoost and tree-based algorithms can work with numeric categorical features as well as one-hot encoded categorical features (one-hot encoding is not required; however, XGBoost can handle one-hot encoded data).

For binary feature like holiday, you need to convert to numeric value using label-encoding.

For numeric features, convert them to a similar range and scale.

AirQuality is the label, and the model needs to learn to predict one of two outcomes.

Since the label contains text, it needs to be converted to a numeric value and we can do that using label encoding.

Note: For multi-class problem, neural networks require one-hot encoding of labels

Question 22: Correct

A company has several audio files that must be converted to other languages.

What is the best way to complete this task?

-
- Transcribe, Polly, Translate
-
- Translate
-
- Translate, Polly
-
- Transcribe, Translate, Polly
    - **(Correct)**

Explanation

Transcribe, Translate, Polly – Translation step requires text data. So, the first step is to transcribe the text from audio and then translate the text. Finally, to convert to speech, use Polly

Question 23: Correct

A manufacturing company has a collection of images that contains examples of normal and defective products. These images need to be manually labeled by human experts for model training, and they need a solution to manage the workflow to distribute images among human experts for manual labeling.

What capability can you for this?

-
- SageMaker GroundTruth
    - **(Correct)**
-
- Rekognition
-
- ImageClassification
-
- SageMaker Neo

Explanation

SageMaker GroundTruth service provides two capabilities to manage labeling process – Automatic Labeling can learn from examples that you provide and label all instances.

Manual labeling uses Mechanical Turk service to distribute the task across human labelers, and GroundTruth provides the capability to manage the entire workflow.

Neo is used for deploying your Machine Learning algorithm anywhere in the Cloud and at Edge Locations – it is a cross-compilation capability to compile your Machine Learning Algorithm to run on specified hardware.

Rekognition is not used for manual labeling even though you can use this service to train to classify images by providing labeled data.

ImageClassification also expects labeled data as input. This question is about how to create labeled data

Question 24: Correct

You have launched an EC2 instance using Deep Learning AMI. Under AWS Shared Responsibility Model, who is responsible for applying critical security patches on EC2 instances?

-
- EC2 Support
-
- AMI Provider
-
- AWS
-
- Customer
    - **(Correct)**

Explanation

For Infrastructure as a Service (IaaS) products like EC2, the customer who launched the instance is responsible for adequately patching the instance. AWS is responsible for keeping AMI up-to-date. Once the EC2 instance is launched, only the customer can patch the instance. Reference: Security is Job Zero https://youtu.be/T7MnJOfOVcY

Question 25: Correct

A binary classifier metrics for validation data has the following values:

TP: 8, FN: 2, TN: 3, FP: 5

What is the Precision for this model?

-
- 0.8
-
- 0.6
  - **(Correct)**
-
- 0.5
-
- 0.3

Explanation

Precision = TP / (TP + FP)

Question 26: Incorrect

An Auto Show organizer wants to detect celebrities who are among the audience. The event center has several cameras that are recording the event live. What combination of service and order of processing can help achieve this task?

-
- Kinesis Video Streams, Amazon Rekognition, Amazon Data Stream
  - **(Correct)**
-
- Kinesis Firehose, Kinesis Analytics, and Amazon Rekognition
-
- Kinesis Data Streams, Amazon Rekognition, Kinesis Video Stream
  - **(Incorrect)**
-
- Kinesis Firehose, Lambda, and Amazon Rekognition

Explanation

Use Kinesis Video Streams to capture the video feed from cameras. Rekognition service can directly consume Kinesis Video Streams, and you can configure Rekognition service to detect celebrities. The output of streaming analysis is stored in a Kinesis Data Stream.

Reference: https://aws.amazon.com/blogs/machine-learning/easily-perform-facial-analysis-on-live-feeds-by-creating-a-serverless-video-analytics-environment-with-amazon-rekognition-video-and-amazon-kinesis-video-streams/

For offline video analysis (video stored in s3), you need to start a job and once the job completes, it will notify using SNS (notification service). You can then pick up the results. Or, you can periodically poll by calling GetCelebrityRecognition.

Working with stored videos: https://docs.aws.amazon.com/rekognition/latest/dg/video.html

Question 27: Incorrect

A highly unbalanced dataset has 95% normal data and 5% positive data. What is a good performance metric to use for assessing the quality of the model?

-
- F1 Score
    - **(Correct)**
-
- Accuracy
-
- Recall
-
- Precision
    - **(Incorrect)**

Explanation

Accuracy is not a useful metric for skewed data sets. Recall on its own is not enough as the model that predicts everything as positive will have a very high Recall. Precision on its own is not enough as the model can have very high precision even if it predicts only one positive correctly and misclassifies everything as negative. F1 Score is a useful metric as it considers both recall and precision. Another metric that you can use is the ROC AUC score.

Question 28: Incorrect

You are using SageMaker's Automatic Hyperparameter tuning to find an optimal set of parameters for a deep learning network. You are using the Bayesian search with a maximum number of training jobs set to 100. What is the recommended amount of concurrent tuning jobs that you can run for the best results?

-
- 4
    - **(Incorrect)**
-
- 1

    - (Correct)
-
- 32
-
- 100

Explanation

“Running more hyperparameter tuning jobs concurrently gets more work done quickly, but a tuning job improves only through successive rounds of experiments. Typically, running one training job at a time achieves the best results with the least amount of compute time.”

https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-considerations.html

Question 29: Correct

You are using CSV formatted files to train on SageMaker’s built-in XGBoost algorithm.

SageMaker expects your training and validation to follow this convention:

-
- CSV must have column headers with the target variable in the first column
-
- CSV must not have a column header record. Target variable must be the last column
-
- CSV must not have a column header record. Target variable must be the first column
    - (Correct)
-

- CSV must have column headers and target variable must be the last column

Explanation

With CSV format, SageMaker XGBoost expects the target variable in the first column and without a column header

Question 30: **Correct**

When you increase the mini-batch size, for every iteration of the training set, the weights of features are adjusted

-
- Less often
    - **(Correct)**
-
- Weight adjustment depends on the number of examples
-
- Weight adjustment is not dependent on mini-batch size
-
- More often

Explanation

Weights are adjusted based on error observed in a mini-batch. The training set is divided into mini-batches, and when you increase the number of examples in a mini-batch, you have fewer mini-batches. And this would result in less-frequent weight adjustment.

Question 31: **Correct**

A startup is analyzing social media trends with data stored in S3. For analysis, it is common to access a subset of attributes across a large number of records. Which of these formats can lower the cost of storage while improving query performance?

-
- JSON
-
- CSV
-
- Avro
-
- Parquet
    - **(Correct)**

Explanation

Parquet is a columnar storage format that transparently compresses data. It is a very efficient format for querying a subset of columns across a large number of records. Avro is a suitable binary format that uses row storage and optimized for use cases that need to access the entire row. JSON and CSV are text formats that use Row storage

Question 32: Correct

An online marketplace wants to help customers make an informed choice when purchasing products. They would like to present the most positive and most critical customer reviews side-by-side in the product summary page.

Which capability can you use for this purpose?

-
- Custom Classification with Comprehend
-
- Rekognition
-
- Textract
-
- Sentiment Analysis with Comprehend
    - **(Correct)**

Explanation

Sentiment Analysis with Comprehend. Sentiment analysis can evaluate text based on the content and provide a confidence score for Positive, Negative, Neutral, Mixed. You could use this to assess reviews based on the sentiment and shortlist strong positive and strong negative reviews.

Question 33: Correct

A labeled dataset contains a lot of duplicate examples. How should you handle duplicate data?

-
- Ensure there are no duplicates
  - **(Correct)**
-
- Ensure all duplicates are in train data
-
- Ensure all duplicates are in test data
-
- Ensure data is shuffled before creating train and test set

Explanation

Duplicates can accidentally leak into validation and test sets when you split your data. This can cause artificially better performance on validation and test sets. You should clean up the data so that all examples are distinct.

Question 34: Correct

Which activation function would you use in the output layer for a Multi-class Classification neural network that predicts a single label from a set of possible labels?

-

- Softmax
    - **(Correct)**
-
- ReLU
-
- None
-
- Sigmoid

Explanation

Softmax activation is used for predicting a single label from a set of possible labels. Softmax returns the probability for each label, and the sum of all probabilities adds up to a 1. The class with the highest probability is used as the final class for the example.

Question 35: Correct

A binary classifier metrics for validation data has the following values:

TP: 8, FN: 2, TN: 3, FP: 5

What is the Recall for this model?

-
- 0.8
    - **(Correct)**
-
- 0.5
-
- 0.6
-
- 0.3

Explanation

Recall or true positive rate = TP/(TP+FN)

Question 36: Correct

Under the AWS Shared Responsibility Model, the customer is responsible for which of these tasks?

-
- Configuring Access to S3 bucket based on job role
    - **(Correct)**
-
- Virtualization infrastructure
-
- Patching Host Operating System
-
- Physical security of hardware

Explanation

Under the shared responsibility model, data security is the responsibility of the customer. AWS provides capabilities to manage data security; however, it is up to the customer to take advantage of security capabilities based on their individual needs. Physical infrastructure, Facilities, Host Computers (underlying physical servers on which virtual instances run), Network infrastructure are all responsibilities of AWS.

Patching Host Operating System is AWS responsibility - here, host refers to the Physical server.

Patching Guest/Instance Operating System is Customer responsibility - here, instance refers to the virtual instance that customer created.

Additional reading and references:
https://wa.aws.amazon.com/wat.concept.shared-resp-model.en.html,

Security is Job Zero https://youtu.be/T7MnJOfOVcY

Question 37: Correct

A grocery store has a robust online presence. The store wants to improve product recommendations using machine learning and suggest products that are purchased together.

Which of these algorithms can be used for this requirement?

-
- Comprehend
-
- BlazingText
-
- Factorization Machines
  - **(Correct)**
-
- DeepAR

Explanation

Factorization Machines algorithm is used for building recommender systems and for collaborative filtering.

Collaborative filtering algorithms learn the likelihood of a customer purchasing a product based on other customer purchase behavior.

BlazingText is used for text analysis and classification problems.

Comprehend is used for natural language processing and not suitable for this use case.

DeepAR is used for time series forecasting.

Question 38: Correct

You are using a lambda function to invoke SageMaker Endpoints. This function can accept a batch of records as input and returns the list of predicted values. You are testing a new model that requires compute-intensive pre-processing of incoming data. You want to use a higher-performing instance for your lambda function. What option does AWS provide to improve performance?

-
- Increase allocated vCPU
-
- Use a compute-optimized instance
-
- Increase timeout
-
- Increase allocated memory
    - **(Correct)**

Explanation

With Lambda, you must choose the amount of memory needed to execute your function. Based on the memory configuration, proportional CPU capacity is allocated. You can also increase the timeout for up to a maximum of 15 minutes.

Question 39: Correct

You have a dense dataset with 1000s of features. You are using a custom training algorithm that has difficulty handling large datasets; you would like to reduce this dataset to a few important features.

The transformed dataset needs to retain as much information as possible from the original dataset.

What approach can you use for this problem?

-
- Store data in Parquet format
-
- Compress using GZIP algorithm
-
- Use algorithms like Factorization Machines that are optimized for very large datasets
-
- Reduce Dimension using Principal Component Analysis
    - **(Correct)**

Explanation

Principal Component Analysis (PCA) is a dimensionality reduction technique – it works by capturing information contained in the original dataset using far fewer features known as components. The newly generated features (component) can then be used for model training. The one drawback with PCA is the newly generated components cannot be mapped to real-world features as there is no easy way to figure how each feature contributes to a component. You also need to standardize or normalize data before you perform PCA.

The factorization algorithm works very well with large sparse datasets. Since this is a dense dataset, this option is not valid.

GZIP compression merely reduces the storage needed – it does not reduce the number of features.

Parquet format is an efficient binary storage format for columnar access – it is useful for scenarios where you want to extract only some columns from 1000s of columns

Question 40: Correct

Your company uses S3 for storing data collected from a variety of sources. The users are asking for a feature similar to a trash can or recycle bin. Deleted files

should be available for restore for up to 30 days.   How would you implement this? (Choose Two)

-
- Enable Lifecycle Policies on the bucket
    - **(Correct)**
-
- Enable Cross-Region Replication and restore objects from the replicated site
-
- Move the deleted object to a temporary bucket and use it for restoring
-
- Enable Versioning on the bucket
    - **(Correct)**

Explanation

You can enable S3 versioning to keep the older version of the objects. You can create life cycle policies to remove the older version after 30 days. Cross-region can help in protecting against accidental deletion and disaster recovery by keeping a copy of data in a different region. But it is more expensive as a full copy of your bucket is maintained in another region. Moving the deleted objects to another bucket is unnecessary and requires other components.

Question 41: Correct

A dataset contains a large number of features. You would like the algorithm to aggressively prune features that are not relevant. What hyperparameter can you use for this?

-
- L1 Regularization
    - **(Correct)**
-
- Either L1 or L2 Regularization
-
- L2 Regularization
-

- Learning Rate

Explanation

Regularization is used to control how a feature can influence the outcome.

When the model overfits, you can increase regularization to reduce the relative weight of each feature.

Similarly, when a model underfits, you can reduce regularization to allow features to assist in predicting the outcome more actively.

L1 Regularization works by eliminating features that are not important.

L2 Regularization keeps all the features but simply assigns a very small weight to features that are not important

Question 42: Correct

A customer has 1000s of documents, and they would like to create a summary of each document.

Which of these services is best suited for this requirement?

-
- Comprehend
    - **(Correct)**
-
- Textract
-
- Rekognition Text Extraction

-
- Transcribe

Explanation

Comprehend. With Comprehend, you can analyze a document to extract entities and key phrases along with confidence scores. You can use it to provide a summary of the talking points of a document. Seq2Seq algorithm supported by SageMaker is another algorithm that you could for this purpose.

Question 43: Correct

You are using unigram text transformation to convert words to the frequency of occurrence. There are two sentences in the text.

“this is working - not disappointed”

“this is not working - disappointed”

How many features would the transformed dataset have?

-
- 5
    - **(Correct)**
-
- 8
-
- 10
-
- 6

Explanation

With unigram transformation, each unique word is a feature. There are five unique words: disappointed, is, not, this, working. With bigram transformation, you need to include consecutive two-word combinations like “this is”, “is working” and so forth.

Question 44: Correct

You are working on a model to differentiate positive and negative classes – the dataset that was provided to you is highly unbalanced. 99% of the data is normal, with only 1% positive. What steps can you go through to handle this unbalanced dataset? (select two)

-
- Oversample positive data using techniques like SMOTE
  - **(Correct)**
-
- Oversample by duplicating positive data
-
- Use ROC AUC as a metric for the unbalanced dataset
-
- Collect more positive samples
  - **(Correct)**
-
- Use Accuracy as a measure for the unbalanced dataset

Explanation

For the unbalanced dataset, accuracy is not a good measure as a model that predicts all instances as normal will be 99% accurate. ROC AUC metric considers True Positive Rate and False positive Rates at all possible cutoff thresholds. It is a useful metric for binary classifiers. However, they are not suitable for highly imbalanced datasets as ROC curve considers only true positive and false positive rates. ROC does not account for negatives and does not measure the performance well. Instead precision-recall curve is used for imbalanced datasets. Oversampling by duplicating data is not going to improve quality as it does not add any new patterns that algorithms can learn. Synthetic Minority Over-sampling Technique (SMOTE) provides a mechanism for artificial data generation that has shown to improve the accuracy of unbalanced classifiers. To generate data similar to an existing instance, SMOTE uses the nearest neighbors, and generate synthetic data along the lines

that connect the neighbors. This method ensures data is close to other similar instances.
Reference (ROC Imbalance):
https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/
https://towardsdatascience.com/sampling-techniques-for-extremely-imbalanced-data-281cc01da0a8

Question 45: Correct

Training data has values for all features. With Test data, some of the features have missing values. If you build a neural network with training data and use test data to verify performance, how would the neural network behave?

-
- The network would not learn from missing values
    - **(Correct)**
-
- The network would automatically learn insights about missing values
-
- Behavior depends on the number of layers
-
- The response depends on activation function

Explanation

The system would not learn insights from missing values. You would need to create new examples in training data with missing values so that the model can learn to ignore missing values

Question 46: Incorrect

Your legal department has asked your team to ensure that historical manufacturing data are not deleted or tampered for a 5-year period. Your team is currently using Glacier for long term storage. What option would you pick to enforce this policy?

-

- Replicate Data to another read-only bucket
-
- Implement IAM Access Policy to remove delete access or modify access
  - **(Incorrect)**
-
- Use Vault Lock to implement write once, read many type policies
  - **(Correct)**
-
- Enforce controls like these at the application level

Explanation

Vault Lock allows you to set immutable policies to enforce compliance controls. With the IAM Access policy, you can define who has access to storage and type of access. However, the IAM policy on its own is not sufficient for compliance-related controls as someone could change the policy to grant write permissions

Question 47: Correct

You need to read the CSV files in S3, transform the content to Parquet format, and store the processed data back in S3. Which of these options is recommended for this solution?

-
- Use Kinesis Datastreams for collecting the data from S3 and use built-in transformation to store the results in Parquet format
-
- Use Kinesis Firehose for reading the data from S3 and use built-in transformation to store the results in Parquet format
-
- Use Glue ETL to run Spark ETL scripts and configure it as a scheduled job
  - **(Correct)**
-
- Configure S3 to invoke Lambda function when a new file is added, perform the transformation in Lambda, and store the results back in S3

Explanation

Glue ETL provides an easy option to automatically generate ETL scripts and run the script as a scheduled job. Glue ETL provisions required Spark infrastructure to run the job and automatically terminates the environment after the job is completed.

A solution involving Kinesis Firehose requires an additional component to read data from S3 and add it firehose stream. For large files, you would also need to chunk into many messages when adding to the firehose.

Question 48: Correct

You want to test new values for hyperparameters for an algorithm. At what point in the model lifecycle can you change hyperparameters?

-
- Training
  - **(Correct)**
-
- Validation
-
- Hosting
-
- Testing

Explanation

Hyperparameters are used when training the model – These parameters control how a model learns. Once a model is trained, you cannot change the hyperparameters – you need to retrain it

Question 49: Correct

You need to configure the SageMaker Endpoint to Scale on demand. Based on load testing, you have determined that one instance can handle 150 requests per second. Assume a safety factor of 0.5.

What value do you need to set for SageMakerVariantInvocationsPerInstance to trigger auto-scaling action?

Note: SageMakerVariantInvocationsPerInstance is a per minute metric.

-
- 2,250
-
- 18,000
-
- 9,000
-
- 4,500
  - **(Correct)**

Explanation

SageMakerVariantInvocationsPerInstance is a per minute metric that you can monitor with CloudWatch to trigger Auto Scaling actions. When this value exceeds 4500, Autoscaling needs to add a server to handle the increased workload.

SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60 = 150 * 0.5 * 60 = 4500

Question 50: Correct

You are building a neural network for image analysis – What type of network would you use?

-

- Recurrent Neural Network
-
- Convolutional Neural Network
    - **(Correct)**
-
- Try different neural network architectures
-
- General Purpose Neural Network

Explanation

CNN's are ideal for image and video analysis applications – it considers a pixel and surround pixels to identify patterns. RNNs are used for applications where the predicted value depends on previously seen values – time-series forecasting, speech recognition and so forth. The general-purpose neural network considers each pixel as a separate feature and may require a very complex network for image analysis applications – whereas a simple CNN can easily outperform a general-purpose neural network for visual analysis application.

Question 51: Correct

A team of machine learning experts is building a speech recognition system that can work in a noisy factory environment. The dataset consists of 10,000 hours of clean speech data and another dataset with 100 hours of noisy speech data recorded inside the factory.

How do you define training, validation, and test set? (Select Two)

-
- Use 10,000 hours of clean speech data for training the model. Divide 100 hours of noisy data into validation and test sets. Optimize the model to improve validation performance and perform the final test using the test set
    - **(Correct)**
-
- Split the 10,000 hours of clean speech data into training and validation set. Divide 100 hours of noisy speech data, add some to the validation set and keep the rest in the test set

    - **(Correct)**
-
- Split the 10,000 hours of clean speech data into training and validation sets. Optimize the model to improve validation performance. Use 100 hours of noisy data for final testing
-
- Use 100 hours of noisy data for training and split the general speech data for validation and testing

Explanation

The objective of this model is to recognize speech in a noisy environment. Since there is very little noisy data available when compared to clean data, one approach that can be used is to train the model on clean data, split the noisy data into validation and test set. Use the noisy validation data to tune the model performance and perform the final check with test data.

Another option is to split the clean speech data into training and validation sets. Add some of the noisy data to the validation dataset and keep the remaining noisy data for the test set.

If you keep split the clean data into training and validation sets and tune model based on validation performance, this model only performs well with clean data and would perform poorly with noisy test data. That is because the distribution of clean and noisy data is different.

Just training on 100 hours of noisy data may not be enough for this use case. Reference: NIPS 2016 tutorial: Nuts and bolts of building AI applications by Dr. Andrew Ng

Question 52: Correct

When training a deep learning network, what is the impact of using smaller mini-batch sizes?

-
- It can help optimization algorithm jump local minima and explore other areas for global minima
    - **(Correct)**
-
- It will make smoother and more gradual adjustments to the weight
-
- Smaller mini-batch will force the algorithm to converge and get stuck in local minima
-
- Optimization algorithm uses all samples for every weight adjustment

Explanation

Mini-batch has the effect of making more substantial changes to weight as it uses a smaller set of samples to determine the gradient. In a deep learning network, the loss curve is very complex with multiple local minima. To prevent the optimizer from getting stuck in local minima, you can reduce the batch size to jump over local minima

Question 53: Incorrect

Which of these services require you to select an AWS region when using it (choose three)?

-
- S3
    - **(Correct)**
-
- CloudWatch
    - **(Correct)**
-
- SageMaker
    - **(Correct)**
-
- IAM
    - **(Incorrect)**

Explanation

IAM is a global resource, and any policy or user or group or role that you create are available across all regions. With SageMaker, you need to pick a region to launch notebook instances, or for training and hosting models. S3 requires you to specify a region to create a bucket. CloudWatch is a repository of all metrics for monitoring resources in the region

Question 54: Incorrect

You have a collection of documents that has text about a variety of different topics: animals, plants, transportation, travel, food, and so forth. You want to train an algorithm to categorize the documents into one of the above categories.

Which of these algorithms can you use for this requirement?

-
- LDA
-
- Seq2Seq
  - **(Incorrect)**
-
- Comprehend
  - **(Correct)**
-
- Neural Topic Modeling (NTM)

Explanation

LDA and NTM are used for topic modeling; however, they are unsupervised and generally used in exploratory setting for understanding data.

You have the flexibility to specify the number of topics – however, the algorithms automatically assign topics – it may not match with what we consider as topics: travel, food, transportation, and so forth. It will automatically generate appropriate topics.

For example, LDA/NTM may come with a topic that groups travel and food together.

For this problem, Comprehend service can be used to train a classifier that can map text content to a topic. Seq2Seq is used for translation, summarization and so forth

Question 55: Correct

A machine learning specialist needs to get inference for the entire dataset that is stored in S3. The Machine Learning Model was trained on SageMaker.

Which of these options provides a managed infrastructure that is cost-effective for large scale inference?

-
- S3 Analytics
-
- Autoscaling
-
- SageMaker Batch Transform
    - **(Correct)**
-
- SageMaker Endpoint

Explanation

Batch Transform is a cost-effective way to get large scale inference using SageMaker. Batch transform is ideal for situations where you don't need a persistent real-time endpoint, scenarios where you don't need sub-second latency performance. SageMaker manages all resources required for batch transform. SageMaker Endpoint is used for real-time inference. Autoscaling allows you to maintain capacity, handle instance failures and scale based on workload. S3 Analytics is used for analyzing storage access patterns, which in turn can help you to transition data to the right storage class in S3.

Question 56: Correct

An organization is using TensorFlow Machine Learning Framework for building models and would like to migrate the machine learning infrastructure to AWS.

Which one of these options takes the least effort to train, host, and manage TensorFlow models in AWS?

-
- Launch EC2 instance, download and install required Machine Learning Frameworks
-
- Built custom docker image that conforms to SageMaker specification to develop and host models using SageMaker infrastructure
-
- Use pre-built TensorFlow docker images provided by SageMaker to train and host models on SageMaker infrastructure
  - **(Correct)**
-
- Launch EC2 instance with Deep Learning AMIs

Explanation

SageMaker provides pre-built TensorFlow docker images that you can use to train, and host models on SageMaker managed infrastructure efficiently.

Deep Learning AMIs are another option, and you can use it to launch desired EC2 instances pre-configured with necessary tools. However, this requires you to manage and patch EC2 instances.

Launching desired EC2 instances and installing machine learning frameworks is another option – however, you need to manage and patch EC2 instances, and besides, you need to validate and patch ML framework.

Custom Docker images are required when we need to deploy custom models or use a machine learning framework not supported by SageMaker

Question 57: Correct

A Machine Learning Expert is working on a time series forecasting problem to predict future demand for products. The dataset consists of two years' worth of historical data. What is the recommended way to split the training and test set?

-
- Order data by time and set aside first 80% for training and the remaining 20% for testing
    - **(Correct)**
-
- Shuffle data and perform a random split to keep 80% for training and 20% for testing
-
- Split data in such a way that first 80% of the days in a month are part of the training set and the remaining 20% of each month is set aside in the test set
-
- Split data into 80% for training and 20% for testing

Explanation

For time-series forecasting, our objective is to predict the values in the future. To get a realistic assessment of model performance, you need to split the dataset based on time. Set aside the first 70-80% for training and keep the most recent data (toward the end) for testing the accuracy of predictions. Random shuffling is not recommended for time series forecasting

Question 58: Correct

For a regression problem, which of these algorithms cap the output to a range of values seen in the training set? (Choose two)

-
- linear regression
-
- neural network
-
- decision tree

  - **(Correct)**
-
- xgboost
  - **(Correct)**

Explanation

Tree-based algorithms like decision tree, random forest, and xgboost have a lower and upper bound it can predict for regression. The lower and upper bound is determined based on the range of values seen during training.

Question 59: **Correct**

A data scientist is working on a problem to classify incoming data into one of five categories: Good, DefectA, DefectB, DefectC, and DefectD. The dataset consists of primarily numeric features, and some of the samples have missing values for features. This missing values in features can help predict the defect class.

How do you train the model to learn from missing values?

-
- Do nothing – algorithms can handle missing values if you provide examples in the training set
-
- Add substitute variables for each feature – when the feature has a missing value for a sample, set the substitute variable to 1 for that feature, and when the feature has a valid value, set the variable to 0
  - **(Correct)**
-
- Replace missing values with the average value for that feature
-
- Replace missing values with 0

Explanation

Substitute variables are Boolean features that capture if a feature contains a missing value for the sample. This allows the algorithm to learn from missing values

https://docs.aws.amazon.com/machine-learning/latest/dg/data-insights.html#missing-values

Question 60: Correct

You have a requirement to convert temperature from Celsius to Fahrenheit. You have a dataset of a few hundred rows that contain examples of Celsius and equivalent Fahrenheit. These are results observed using different approaches.

Which option would you pick?

-
- When using Linear Regression algorithm, it easily handles this dataset with very low RMSE error on the validation dataset
-
- Instead of using Machine Learning, implement the logic in code as the conversion logic is simple
    - **(Correct)**
-
- When using XGBoost Regression algorithm, it easily handles this dataset with very low RMSE error on the validation dataset
-
- Use either Linear Regression or XGBoost

Explanation

This problem is an example of where you don't want to use Machine Learning. Temperature conversion is easily doable with a simple formula, whereas implementing as an ML solution is more complicated and has a lot of overhead in terms of model training, validation, hosting, and ongoing maintenance.

Question 61: Correct

A binary classifier metrics for validation data has the following values:

TP: 8, FN: 2, TN: 4, FP: 5

How many positive and negative samples are there in the validation dataset?

-
- Positive: 13, Negative: 6
-
- Positive: 6, Negative: 13
-
- Positive: 10, Negative: 9
  - **(Correct)**
-
- Positive: 12, Negative: 7

Explanation

Positive = True Positive + False Negative

Negative = True Negative + False Positive

Question 62: Correct

You are working on developing a solution to identify specific breeds of cats and dogs from an image. The dataset you have is small. You noticed that an existing image classification neural network that was trained on a large dataset has an excellent ability to classify images. You would like to reuse the network to make it work for the new problem. What steps can you take to accomplish this?

-
- Use Transfer learning and remove the first hidden layer of image classification model and retrain the model
-

- Retrain the image classification model with new data
-
- Use Transfer learning by removing the output layer of the image classification model, reinitialize the weights of all layers and retrain the model
-
- Use Transfer learning by removing the output layer of the image classification model, reinitialize the weights of last hidden layer and retrain the model
    - **(Correct)**

Explanation

Transfer learning is an approach of reusing a model that works well for a similar problem. With neural networks and deep learning, some domains like speech recognition, image recognition, and so forth require an extensive dataset for the algorithm to learn all patterns. You can reuse these models for more specialized tasks by using transfer learning. Commonly, with transfer learning, you remove the output layer of one model and feed the hidden layer to a different set of neurons that assess performance with the new dataset. The algorithm is now retrained to learn new patterns and adjust the weight. You can start by randomly initializing the weights of the last layer of the existing, and for more complex use cases, you may need to random initialize of weights of the final few layers. Reference: NIPS 2016 tutorial: Nuts and bolts of building AI applications by Dr. Andrew Ng

Question 63: Correct

A customer is using Polly to generate audio for text. However, Polly is not pronouncing some of the words correctly. What option would help you control the speech output?

-
- Use correct Region and Language
-
- Use real-time streaming for highest quality output
-
- Use batch streaming for highest quality outputs
-
- Use Speech Synthesis Markup Language
    - **(Correct)**

Explanation

With Polly, you can use Speech Synthesis Markup Language (SSML) to “control aspects of speech, such as pronunciation, volume, pitch, speed rate, etc.” Reference: https://aws.amazon.com/polly/

Question 64: Correct

You are developing a deep learning network for converting speech to text. The dataset has recordings of 1,000 individuals, with everyone providing five different audio files along with the transcribed text.   (for a total 5,000 audio samples). The trained model must generalize well for new individuals. How would you use this data for developing a model?

-
- Randomly split data between training and test set
-
- Ensure some individuals are only in the test set – use the remaining data for training and validation
    - **(Correct)**
-
- For each individual, keep four audio files in the training set and one in the test set
-
- For each individual, keep three audio files in the training set, one in validation set and one in the test set

Explanation

The objective is to ensure the model generalizes well for unheard voices.   So, the test set should not contain any individuals from the training or validation set. If we have the same individuals in the training and test set – the model may memorize voice for that individual and may artificially show improved performance. Reference: NIPS 2016 tutorial: Nuts and bolts of building AI applications by Dr. Andrew Ng.

Question 65: Correct

You are exploring different parameters for tuning the model. What dataset should you use to guide with this tuning exercise?

-
- Train
-
- Test
-
- Validation
    - **(Correct)**
-
- Use a random sample from train, validation and test sets

Explanation

Tune model using validation data. To prevent the model from overfitting the validation data, you need to plan to do a final check with an unseen test data