AWS ML – Q & A

1. What is the purpose of regularization in machine learning, and what are some common regularization techniques?
   - Regularization is used to decrease the complexity of the model in order to prevent overfitting, and common techniques include L1 and L2 regularization.

2. You are working on a project where you need to visualize and analyze large amounts of customer transaction data to identify trends and patterns. Which AWS service can you use to perform complex data analysis and visualization tasks efficiently and securely?
   - Amazon QuickSight

3. Which of the following AWS services provides a web-based notebook interface that can be used for data exploration, analysis, and machine learning model building?
   - Amazon SageMaker Notebooks

4. You are designing a machine learning application that requires deploying Docker containers in a serverless environment. The application needs to scale up and down automatically based on the incoming traffic. Which of the following AWS services is best suited for this requirement?
   - AWS Fargate
     - AWS Fargate is a serverless compute engine for containers that allows you to run Docker containers without having to manage servers or clusters. Fargate provides a serverless platform for deploying and scaling containers, where you only pay for the resources that you use. Fargate also integrates with Amazon ECS, which is a fully-managed container orchestration service. With Fargate and ECS, you can easily deploy and manage your containerized applications, and scale them up and down automatically based on incoming traffic.

5. Your company has a streaming data pipeline that processes data using Amazon Kinesis Data Analytics. The pipeline processes a high volume of data and requires high throughput and low latency. Which of the following configurations would you recommend for the Kinesis Data Analytics application?
   - Use multiple input streams and multiple output streams
     - When processing a high volume of streaming data with Amazon Kinesis Data Analytics, it is essential to configure the application for high throughput and low latency. Using multiple input streams and multiple output streams can help achieve this goal by distributing the processing load across multiple processing units.
     - **Use a combination of Kinesis Data Streams and Kinesis Data Firehose instead of Kinesis Data Analytics**
       - using multiple input streams and multiple output streams, is the recommended configuration for processing high

volumes of streaming data with Amazon Kinesis Data Analytics. It helps distribute the processing load and can provide high throughput and low latency.

6. Which of the following **Amazon Elastic MapReduce (EMR)** cluster configurations would be most suitable for a machine learning (ML) workload that requires high memory resources and distributed processing across a large number of instances?
   - A cluster with all instances running on r5d.4xlarge with EBS-optimized instances and a single master node running on an m5.2xlarge instance

   Explanation

   For a machine learning workload that requires high memory resources and distributed processing across a large number of instances, a cluster configuration with all instances running on r5d.4xlarge with EBS-optimized instances and a single master node running on an m5.2xlarge instance would be the most suitable option. The r5d.4xlarge instance type is optimized for memory-intensive workloads, with 16 vCPUs, 32 GiB of memory, and up to 3.5 Gbps of dedicated EBS bandwidth. The EBS-optimized instances would provide dedicated bandwidth between EC2 instances and EBS volumes, which is important for I/O-intensive workloads. Additionally, using a single master node running on an m5.2xlarge instance would ensure that the master node has enough resources to manage the cluster while minimizing costs.

7. A retail company wants to collect customer transaction data from multiple sources, such as point-of-sale systems and online purchases, for analysis. Which AWS service is best suited for this use case?
   - Amazon Kinesis Data Streams

   Explanation

   Amazon Kinesis Data Streams is a fully managed real-time data streaming service that allows you to collect and process large amounts of data from multiple sources. In this scenario, Kinesis Data Streams can be used to collect transaction data from various sources such as point-of-sale systems and online purchases. Kinesis Data Streams can then be used to process this data in real-time and feed it into an analytics solution or data store.

8. You are building a machine learning application that requires you to process data in real-time and store the results in an S3 bucket. You have decided to use AWS Lambda to run your machine learning application. However, you are concerned about the security of your Lambda function and the data it processes. Which of the following options would be the best way to secure your Lambda function and the data it processes?
   - Use the VPC endpoint to isolate your Lambda function and restrict access to it using security groups

   Explanation

   Using the VPC endpoint to isolate your Lambda function and restrict access to it using security groups can help ensure that only authorized traffic is allowed to reach your function and that your function can only access authorized resources. This can help to prevent unauthorized access to your function and the data it processes.

9. Which of the following statements is true regarding Amazon Translate and its use in machine learning applications?
   - Amazon Translate is a cloud-based neural machine translation service that provides real-time translation of text from one language to another.

10. You are designing a serverless machine learning application that will be deployed on AWS Lambda. The application will require access to a large amount of data stored in an S3 bucket. What is the most efficient way to provide access to the S3 data from Lambda?
    - Use IAM roles to grant access to the S3 bucket

    Explanation

    The most efficient and secure way to provide access to S3 data from Lambda is to use IAM roles. IAM roles allow you to grant permissions to AWS services, such as Lambda, without requiring the use of access keys. By creating an IAM role with permissions to access the S3 bucket and then assigning that role to the Lambda function, you can ensure that the function has the necessary permissions to access the data without the need to manage access keys or credentials.

11. A company wants to detect faces of celebrities in their social media campaign images to better understand the effectiveness of their influencer marketing strategy. Which feature of Amazon Rekognition would be most useful for this task?
    - Celebrity recognition

12. You are developing a machine learning application that requires a high-performance file system that can be accessed by multiple EC2 instances simultaneously. Which of the following Amazon FSx options would be the best choice for your application?
    - Amazon FSx for Lustre - Amazon FSx for Lustre is a high-performance file system optimized for compute-intensive workloads, making it ideal for machine learning applications. It provides sub-millisecond access to petabyte-scale file systems, enabling multiple EC2 instances to read and write data simultaneously. Additionally, it provides data compression and encryption, as well as the ability to export data to Amazon S3 for long-term storage.

13. You are designing a machine learning solution on AWS that requires role-based access control for your AWS resources. Which of the following statements regarding AWS Identity and Access Management (IAM) is correct?
    - IAM is a service that provides secure and controlled access to your AWS resources using roles, policies, and permissions.
      - AWS Identity and Access Management (IAM) is a service that enables you to manage access to your AWS resources securely.

IAM enables you to create and manage AWS users, groups, and roles, and to assign permissions to those entities to access AWS resources. IAM is a critical component of AWS security, providing granular control over who can access your AWS resources and what actions they can perform. By using IAM, you can ensure that your AWS resources are secure and that only authorized users have access to them.

14. What is the most appropriate AWS service for deploying and scaling a real-time machine learning model that is built using TensorFlow and requires low-latency inference, and why?
    - Amazon SageMaker because it offers a fully-managed service for building, training, and deploying machine learning models at scale, with support for TensorFlow and low-latency inference using the SageMaker Neo runtime.

15. Which AWS service provides automated anomaly detection and forecasting capabilities for operational metrics and logs, allowing you to set up alerts and take automated actions when anomalies are detected?
    - Amazon CloudWatch

16. Which of the following instance types is optimized for compute-intensive workloads and provides the lowest cost per vCPU in Amazon EC2?
    - C5

17. What is the maximum size of a single record that can be sent to Amazon Kinesis Data Streams?
    - 512 KB

18. You are training a machine learning model on a large dataset with high-dimensional features. The training process is taking a long time, and you want to speed it up by using multiple GPUs. Which of the following AWS services would you use for distributed model training?
    - Amazon SageMaker

19. A company wants to develop a system for transcribing audio recordings of customer service calls to improve their customer service. They plan to use Amazon Transcribe for this purpose. Which of the following is a feature of Amazon Transcribe that can help the company achieve their goal?
    - Speaker diarization
        - Amazon Transcribe is a speech recognition service that can be used to convert speech to text. It provides several features that can help the company transcribe their customer service calls. One of these features is speaker diarization, which can automatically identify and separate different speakers in a conversation. This can be useful in customer service calls where there are multiple speakers, such as the customer and the representative. By separating the speakers, the company can

analyze the conversation more easily and gain insights into the interactions between the customer and the representative.

20. Which of the following statements accurately describe Amazon Textract?
    - Amazon Textract is a fully-managed machine learning service that automatically extracts text and data from scanned documents.

21. Which of the following statements accurately describe the features of Amazon S3 lifecycle policies for machine learning workloads?
    - Amazon S3 lifecycle policies allow you to automate the transition of objects between storage classes based on their age, reducing storage costs for machine learning workloads.

22. Which of the following strategies can be used to ensure efficient operational performance of a deployed machine learning model on AWS?
    - Implementing a horizontal scaling strategy by increasing the number of EC2 instances in an Auto Scaling group.

23. Which AWS service can be used to orchestrate and automate ETL workflows for large datasets?
    - AWS Glue

24. Which of the following statements is true about using Amazon Elastic Kubernetes Service (EKS) for deploying machine learning models?
    - EKS supports GPU instances, which are ideal for training and deploying machine learning models
        - Amazon EKS provides GPU instances that can be used to train and deploy machine learning models. EKS supports various types of GPU instances such as P3, P2, and G4. These GPU instances are ideal for machine learning workloads as they offer high-performance computing capabilities and accelerate the training process for deep learning models. In addition, EKS also supports CPU instances that can be used for machine learning workloads that do not require GPU acceleration.

25. You are deploying a Kubernetes cluster on Amazon EKS and you want to use the AWS Key Management Service (KMS) to encrypt your Kubernetes secrets. Which of the following statements is true about the use of KMS with Amazon EKS?
    - You need to manually configure KMS to encrypt Kubernetes secrets on Amazon EKS

26. Which of the following techniques is used in natural language processing (NLP) to represent words as dense vectors of real numbers, also known as word embeddings?
    - Word2Vec

27. An AWS DeepLens device has been deployed in a factory for object detection and recognition purposes. The factory produces a variety of goods with

different shapes and sizes, and the objects need to be identified in real-time. The factory has a mix of fluorescent and incandescent lighting, which can cause variations in the images captured by the camera. What AWS DeepLens feature could be used to adjust the image quality and improve the accuracy of the object detection models in this scenario?

- Image preprocessing using AWS Lambda functions
  - In the given scenario, the factory has a mix of fluorescent and incandescent lighting, which can cause variations in the images captured by the camera. These variations can affect the accuracy of the object detection models. To improve the accuracy of the models, image preprocessing techniques can be used. AWS Lambda functions can be used to adjust the image quality, such as reducing noise and adjusting contrast, before the images are fed into the object detection models.

28. Which AWS service provides a fully managed, real-time streaming data ingestion and processing solution that can be used for machine learning applications that require low-latency data processing?
    - AWS IoT Analytics
    - Amazon Kinesis Data Analytics
      - **AWS IoT Analytics** is a fully managed service that provides real-time streaming data ingestion and processing capabilities for Internet of Things (IoT) applications. It supports custom data transformation, enrichment, and filtering, and it integrates with other AWS services such as Amazon S3 and Amazon Elasticsearch. It is an ideal service for machine learning applications that require low-latency data processing.
      - **Amazon Kinesis Data Streams** is a fully managed, scalable, and highly available service that can be used for real-time streaming data ingestion and processing. It can be used for custom data transformation, processing, and analysis, and it can integrate with other AWS services such as Amazon S3, Amazon Redshift, and Amazon Elasticsearch. It is an ideal service for machine learning applications that require real-time data processing and analysis.
      - **AWS AppSync** is a managed service that provides GraphQL APIs for building scalable and secure applications, but it is not designed for data ingestion or collection.
      - **Amazon Kinesis Data Analytics** is a service for analyzing real-time streaming data, but it is not designed for data ingestion or collection.
      - **Amazon API Gateway** is a fully managed service for creating, deploying, and managing APIs, but it is not designed for data ingestion or collection.

29. A data science team is developing a machine learning application using Amazon ECS for container orchestration. The application needs to process large amounts of data and requires scalable compute capacity. The team has decided to use Spot Instances to save costs. Which of the following statements is true regarding the use of Spot Instances in Amazon ECS?
    - Spot instances can only be used in Amazon ECS if the instances are launched as part of an Auto Scaling group

30. You need to monitor the performance of your EC2 instances and ensure that your applications are running smoothly. Which of the following statements about Amazon CloudWatch is true?
    - Amazon CloudWatch can monitor CPU utilization, disk I/O, and network traffic for EC2 instances.

31. A company wants to run large-scale, batch computing workloads using AWS Batch. They want to ensure that their jobs are optimized for cost and performance. Which of the following options would best achieve their goals?
    - Use AWS Batch managed compute environments to automatically provision and manage compute resources for the jobs. Use On-Demand instances for the compute environment, and configure the job definitions to use Spot Instances for the actual job execution.

32. You are working on an ETL (Extract, Transform, Load) job using AWS Glue, and you want to write the transformed data to an S3 bucket. Which of the following are valid ways to accomplish this? Choose the correct option.
    - Use the *write_dynamic_frame* method of a Glue *DynamicFrame* object to write data to S3.

33. A media company wants to use Amazon Polly to create audio content for their news website. They want to use a voice that sounds like a specific newscaster who has a unique voice. Which of the following options would be the best solution for this requirement?
    - Use Amazon Polly's Neural Text-to-Speech (NTTS) technology with a custom voice.

34. Which of the following is not a feature of Amazon Fraud Detector?
    - Providing a platform for virtual reality development.
        - Amazon Fraud Detector is a fully managed service that makes it easy to identify potentially fraudulent activities by providing a real-time risk score for each transaction or entity. It allows you to create custom models by ingesting data from various sources such as Amazon S3, Amazon SageMaker, Amazon RDS, and more. The service also provides automated feedback loops to help you continuously improve your models. However, providing a platform for virtual reality development is not a feature of Amazon Fraud Detector. This feature is not related to fraud detection or risk assessment and is not mentioned in the documentation for Amazon Fraud Detector.

35. A team of developers is building a machine learning application and wants to store Docker images containing the application code and dependencies. The application uses GPU-based machine learning models and requires a custom deep learning framework. The team has decided to use Amazon Elastic Container Registry (Amazon ECR) to store their Docker images. Which of the following statements is true regarding the use of Amazon ECR for this purpose?

- Amazon ECR is a fully managed container registry service that allows users to store, manage, and deploy Docker images. It supports GPU-based workloads and custom deep learning frameworks, and provides integrations with Amazon ECS, Amazon EKS, and AWS Batch.

36. Which AWS service can be used to collect, process, and store log data, providing real-time data streaming and search capabilities, and integration with other AWS services?
    - Amazon CloudWatch Logs

37. Which of the following natural language processing (NLP) techniques is best suited for extracting semantic meaning from text data?
    - Latent Dirichlet Allocation (LDA)
        - **Latent Dirichlet Allocation (LDA)** is a probabilistic topic modeling technique that is widely used in natural language processing (NLP) to identify the topics present in a large corpus of text data. It is used to discover hidden topics in text documents and to infer the probability distribution of words in those topics. LDA is particularly useful for extracting semantic meaning from unstructured text data because it can identify the underlying themes and concepts in a large corpus of text without relying on explicit semantic or syntactic information.

38. A company is using Amazon Redshift for their data warehousing needs. They have created a new Redshift cluster with 4 nodes and 2 slices per node. They want to ensure that their workload is spread evenly across all nodes and slices to achieve maximum performance. Which of the following options should they use to achieve this goal?
    - Use ROUND ROBIN distribution style on their tables
        - When using Amazon Redshift, choosing the right distribution style is crucial for achieving optimal performance. The distribution style determines how data is distributed across the nodes of the cluster. A ROUND ROBIN distribution style distributes the rows of a table evenly across all of the slices in the cluster. This means that each slice will contain an equal portion of the data and workload will be spread evenly across all nodes and slices. This is the best option to achieve maximum performance when working with a cluster with multiple nodes and slices.

39. A company has millions of sensor data records that need to be preprocessed before training a machine learning model. Which AWS service can be used to preprocess this data?
    - AWS Glue

40. You are designing a real-time data streaming pipeline with Amazon Kinesis Data Firehose. You need to transform incoming data before it is stored in Amazon S3. Which of the following services can you use to perform the data transformation?

- AWS Lambda

41. You are working on a project to forecast sales for a retail company. You have historical sales data for the past three years and want to use Amazon Forecast to build a forecasting model. Which of the following is a valid approach to prepare the data for use with Amazon Forecast?
    - Convert the data into a time series dataset and upload it to Amazon S3, with each row containing a timestamp, item ID, and the corresponding sales value.

42. A manufacturing company wants to build an ML model that can predict equipment failure based on sensor data from their production line. Which AWS ML application service would be most appropriate for this use case, considering the need for real-time predictions, scalability, and the ability to handle streaming data?
    - Amazon Kinesis Data Analytics

43. You are building a machine learning model to predict the likelihood of a customer buying a product based on their past purchase history and demographic information. As part of the model evaluation process, you want to visualize the distribution of the target variable and the relationship between the target variable and the input features. Which of the following AWS services would you use for data analysis and visualization?
    - Amazon QuickSight

44. You need to deploy a machine learning model for batch inference on AWS. The input data for the model is stored in an Amazon S3 bucket, and the output needs to be stored in a different S3 bucket. Which of the following AWS services would be the best fit for this use case?
    - Amazon SageMaker Batch Transform

45. Which of the following AWS services can be used for data ingestion/collection in a machine learning pipeline?
    - Amazon Kinesis Data Firehose
        - **Amazon Kinesis Data Firehose** is an AWS service that can be used to capture, transform, and load streaming data into data lakes, data stores, and analytics tools. It can be used to collect and ingest large amounts of data from various sources such as sensors, clickstreams, and social media feeds. Kinesis Data Firehose can also transform and pre-process data using AWS Lambda functions before storing it in S3, Redshift, or Elasticsearch.

46. You are a data engineer responsible for designing a data pipeline to process and transform large datasets in AWS. Which of the following statements is true regarding AWS Data Pipeline?
    - AWS Data Pipeline is a fully managed service that allows you to create complex workflows for data processing and transformation using a graphical interface.

47. A company wants to build a chatbot using Amazon Lex to automate their customer support process. They want to use multiple channels such as web, mobile, and messaging platforms to interact with customers. What is the most suitable integration option for them to implement the chatbot?
    - Amazon Connect Integration
        - For a company that wants to build a chatbot using Amazon Lex and use it across multiple channels such as web, mobile, and messaging platforms, the best integration option is Amazon Connect. Amazon Connect is an omnichannel cloud contact centre that provides a seamless experience for customers across voice and chat. It integrates with Amazon Lex to provide intelligent chatbot experiences, allowing customers to interact with the bot through various channels such as voice, chat, and messaging

48. Which of the following statements accurately describes Amazon Elastic File System (Amazon EFS)?
    - Amazon EFS provides a scalable and fully managed file storage service that is accessible from on-premises and cloud-based resources via the Network File System (NFS) protocol.

49. A company is looking to build a dashboard to visualize real-time insights for their e-commerce platform. They want to use Amazon QuickSight as the BI tool to visualize data from multiple sources, including Amazon Redshift, Amazon RDS, and Amazon S3. Which of the following options describes the best way to achieve this?
    - Use AWS Glue to crawl and catalog the data sources, and then create a Glue ETL job to transform and load the data into a single data store such as Amazon Redshift or Amazon RDS. Connect QuickSight to the data store and create the dashboard. Share the dashboard with the users.

50. Which of the following machine learning models is supported by Amazon Machine Learning (Amazon ML)?
    - Random Forest(RF)
        - Amazon Machine Learning supports two types of models: regression and binary classification. The algorithms used for these models are linear regression, logistic regression, and random forests. Random forests are a popular choice for building machine learning models because they can handle complex data sets and can be used for both classification and regression tasks. CNNs and RNNs are neural network architectures used for tasks such as image recognition and natural language processing, and are not supported by Amazon ML. LSTM is a type of RNN architecture that is also not supported. SVMs are a popular choice for classification tasks, but they are not supported by Amazon ML

51. Which AWS machine learning (ML) application service should be used to train and deploy a natural language processing (NLP) model for sentiment analysis of social media data, given the requirement of minimizing training and deployment costs?
    - Amazon Comprehend - is the AWS ML application service that provides pre-built APIs for natural language processing tasks, such as sentiment analysis, entity recognition, and topic modeling. It can be used to analyze large volumes of text data, including social media posts, and can detect sentiment at scale with a high degree of accuracy. It also supports custom classification models for specific use cases, and the costs of training and deployment are comparatively lower than other options.

52. Which of the following is not a feature of Amazon Comprehend?
    - Object Detection

53. Which of the following techniques can help address the problem of overfitting when training a machine learning model?
    - Implementing early stopping during training

54. Which of the following statements is true about the AWS Deep Learning AMIs (DLAMI)?
    - The AWS Deep Learning AMIs come with pre-installed deep learning frameworks and libraries, making it easier to develop and deploy deep learning models.

55. Which of the following AWS services can be used for ingesting and collecting large amounts of data for machine learning purposes? (select 2)
    - Amazon S3
    - Amazon Kinesis

56. You are responsible for monitoring and analyzing the activities of multiple AWS accounts in your organization. Which of the following statements about AWS CloudTrail is true?
    - AWS CloudTrail stores log files in an S3 bucket in the same region as the trail.
        - AWS CloudTrail is a service that enables governance, compliance, operational auditing, and risk auditing of your AWS account. CloudTrail logs all management events for AWS services and resources in the account. The logs are delivered to an S3 bucket that you specify and can be used for various purposes, such as security analysis, resource change tracking, and operational troubleshooting.

57. What is the maximum number of VPCs that can be created per AWS account, and what is the default limit of subnets per VPC in the Amazon VPC service?
    - The maximum number of VPCs that can be created per AWS account is 5, and the default limit of subnets per VPC is 5.

58. A healthcare provider wants to build an ML model that can predict patient outcomes based on their medical history and demographic information. Which AWS ML application service would be the best fit for this use case, given the need for interpretability, explainability, and compliance with regulatory requirements?
    - Amazon HealthLake
        - **Amazon HealthLake** is an AWS ML application service that provides a HIPAA-eligible data lake for healthcare and life sciences organizations. It allows healthcare providers to securely store, transform, and analyze health data using machine learning algorithms, while also complying with regulatory requirements for privacy and security. It also provides features for data standardization, entity recognition, and natural language processing, which can help with interpreting and explaining the predictions made by an ML model.

59. Which AWS service can be used to collect and process large amounts of data from multiple sources, providing real-time data streaming, data transformation, and integration with AWS Machine Learning services?
    - Amazon Kinesis Data Analytics

60. Which of the following is a benefit of using a cloud-based notebook environment, such as Amazon SageMaker Notebook, over a local notebook environment?
    - Better performance and scalability.

61. Which of the following services on AWS provides a fully-managed ETL solution for processing large datasets at scale?
    - AWS Glue

62. Which of the following is a feature of Amazon S3?
    - Ability to host static websites

63. You are building a deep learning model to classify images in an ecommerce platform, and you want to train your model using Amazon SageMaker. You have a large dataset of images stored in an Amazon S3 bucket, and you want to use Amazon SageMaker's built-in algorithms to train your model. However, you need to preprocess the images before training by resizing them to a specific size and converting them to grayscale.Which of the following approaches should you use to preprocess the images using Amazon SageMaker?
    - Use Amazon SageMaker’s built-in transform jobs to resize and convert the images to grayscale before training, and specify the transformed data location as the input to the training job.

64. Which of the following statements regarding the integration of Notebooks and IDEs with AWS services is true?

- SageMaker Notebook instances are pre-configured with the latest version of JupyterLab.
    - **AWS SageMaker** provides multiple options for integrated development environments, including Jupyter Notebook, JupyterLab, and the recently introduced SageMaker Studio. SageMaker Notebook instances are pre-configured with the latest version of JupyterLab, an interactive development environment (IDE) that enables users to create and run Jupyter notebooks. SageMaker Studio is a web-based IDE that provides access to pre-configured Jupyter and PyCharm, and it includes several tools for building, training, and deploying machine learning models.

65. Which of the following is true about AWS IoT Greengrass?
    - AWS IoT Greengrass enables local execution of AWS Lambda funcitons, AWS IoT Device Shadow, and AWS IoT Device Gateway on IoT devices even when they are not connected to the cloud.
        - AWS IoT Greengrass enables IoT devices to run AWS Lambda functions, AWS IoT Device Shadow, and AWS IoT Device Gateway locally, even when the device is not connected to the cloud. This enables edge computing and allows for faster response times and reduced data transfer to the cloud. AWS IoT Greengrass supports multiple messaging protocols including MQTT, HTTP, and WebSocket. Devices do not require a constant internet connection to operate with AWS IoT Greengrass. Additionally, AWS IoT Greengrass does support containerization of applications.