AWS - Sample Exam Questions

https://d1.awsstatic.com/training-and-certification/docs-ml/AWS-Certified-Machine-Learning-Specialty_Sample-Questions.pdf

Results: 6/10 correct => 60% on 10/18/2020

1) A machine learning team has several large CSV datasets in Amazon S3. Historically, models built with the Amazon SageMaker Linear Learner algorithm have taken hours to train on similar-sized datasets. The team's leaders need to accelerate the training process.

What can a machine learning specialist do to address this concern?

A) Use Amazon SageMaker Pipe mode.
B) Use Amazon Machine Learning to train the models.
C) Use Amazon Kinesis to stream the data to Amazon SageMaker.
D) Use AWS Glue to transform the CSV dataset to the JSON format.

=> A: Correct

1) A – Amazon SageMaker Pipe mode streams the data directly to the container, which improves the performance of training jobs. (Refer to this link for supporting information.) In Pipe mode, your training job streams data directly from Amazon S3. Streaming can provide faster start times for training jobs and better throughput. With Pipe mode, you also reduce the size of the Amazon EBS volumes for your training instances. B would not apply in this scenario. C is a streaming ingestion solution, but is not applicable in this scenario. D transforms the data structure.

2) A term frequency–inverse document frequency (tf–idf) matrix using both unigrams and bigrams is built from a text corpus consisting of the following two sentences:

1. Please call the number below.
2. Please do not call us.

What are the dimensions of the tf–idf matrix?

A) (2, 16)
B) (2, 8)
C) (2, 10)
D) (8, 10)

=> A: Correct

2) A – There are 2 sentences, 8 unique unigrams, and 8 unique bigrams, so the result would be (2,16). The phrases are "Please call the number below" and "Please do not call us." Each word individually (unigram) is "Please," "call," "the," "number," "below," "do," "not," and "us." The unique bigrams are "Please call," "call the," "the number," "number below," "Please do," "do not," "not call," and "call us." The tf–idf vectorizer is described at this link.

3) A company is setting up a system to manage all of the datasets it stores in Amazon S3. The company would like to automate running transformation jobs on the data and maintaining a catalog of the metadata concerning the datasets. The solution should require the least amount of setup and maintenance.

Which solution will allow the company to achieve its goals?

A) Create an Amazon EMR cluster with Apache Hive installed. Then, create a Hive metastore and a script to run transformation jobs on a schedule.
B) Create an AWS Glue crawler to populate the AWS Glue Data Catalog. Then, author an AWS Glue ETL job, and set up a schedule for data transformation jobs.
C) Create an Amazon EMR cluster with Apache Spark installed. Then, create an Apache Hive metastore and a script to run transformation jobs on a schedule.
D) Create an AWS Data Pipeline that transforms the data. Then, create an Apache Hive metastore and a script to run transformation jobs on a schedule.

=> B: Correct

3) B – AWS Glue is the correct answer because this option requires the least amount of setup and maintenance since it is serverless, and it does not require management of the infrastructure. Refer to this link for supporting information. A, C, and D are all solutions that can solve the problem, but require more steps for configuration, and require higher operational overhead to run and maintain.

4) A data scientist is working on optimizing a model during the training process by varying multiple parameters. The data scientist observes that, during multiple runs with identical parameters, the loss function converges to different, yet stable, values.

What should the data scientist do to improve the training process?

A) Increase the learning rate. Keep the batch size the same.
B) Reduce the batch size. Decrease the learning rate.
C) Keep the batch size the same. Decrease the learning rate.
D) Do not change the learning rate. Increase the batch size.

=> C: Incorrect (Correct is B)

4) **B** – It is most likely that the loss function is very curvy and has multiple local minima where the training is getting stuck. Decreasing the batch size would help the data scientist stochastically get out of the local minima saddles. Decreasing the learning rate would prevent overshooting the global loss function minimum. Refer to the paper at this link for an explanation.

https://arxiv.org/pdf/1609.04836.pdf
https://stats.stackexchange.com/questions/164876/tradeoff-batch-size-vs-number-of-iterations-to-train-a-neural-network
It has been observed in practice that when using a larger batch there is a significant degradation in the quality of the model, as measured by its ability to generalize
The lack of generalization ability is due to the fact that large-batch methods

tend to converge to *sharp minimizers* of the training function.
The size of the learning rate is limited mostly by factors like how curved the cost function is. You can think of gradient descent as making a linear approximation to the cost function, then moving downhill along that approximate cost. If the cost function is highly non-linear (highly curved) then the approximation will not be very good for very far, so only small step sizes are safe.

Learning rate: Since it influences to what extent newly acquired information overrides old information, it metaphorically represents t**he speed at which a machine learning model "learns"**

In setting a learning rate, there is a trade-off between the rate of convergence and overshooting. While the descent direction is usually determined from the gradient of the loss function, the learning rate determines how big a step is taken in that direction.[4] A too high learning rate will make the learning jump over minima but a too low learning rate will either take too long to converge or get stuck in an undesirable local minimum.[5]

The **batch size** defines the number of samples that will be propagated through the network.

5) A data scientist is evaluating different binary classification models. A false positive result is 5 times more expensive (from a business perspective) than a false negative result.

The models should be evaluated based on the following criteria:

1) **Must have a recall rate of at least 80%**
2) **Must have a false positive rate of 10% or less**
3) **Must minimize business costs**

After creating each binary classification model, the data scientist generates the corresponding confusion matrix.

Which confusion matrix represents the model that satisfies the requirements?

A) TN = 91, FP = 9
FN = 22, TP = 78
B) TN = 99, FP = 1
FN = 21, TP = 79
C) TN = 96, FP = 4
FN = 10, TP = 90
D) TN = 98, FP = 2
FN = 18, TP = 82

=> D: Correct

5) D – The following calculations are required:

TP = True Positive
FP = False Positive
FN = False Negative
TN = True Negative
FN = False Negative

Recall = TP / (TP + FN)

False Positive Rate (FPR) = FP / (FP + TN)

Cost = 5 * FP + FN

	A	B	C	D
Recall	78 / (78 + 22) = 0.78	79 / (79 + 21) = 0.79	90 / (90 + 10) = 0.9	82 / (82 + 18) = 0.82
False Positive Rate	9 / (9 + 91) = 0.09	1 / (1 + 99) = 0.01	4 / (4 + 96) = 0.04	2 / (2 + 98) = 0.02
Costs	5 * 9 + 22 = 67	5 * 1 + 21 = 26	5 * 4 + 10 = 30	5 * 2 + 18 = 28

Options C and D have a recall greater than 80% and an FPR less than 10%, but D is the most cost effective. For **supporting** information, refer to this link.

6) A data scientist uses logistic regression to build a fraud detection model. While the model accuracy is 99%, 90% of the fraud cases are not detected by the model.

What action will definitively help the model detect more than 10% of fraud cases?

A) Using undersampling to balance the dataset
B) Decreasing the class probability threshold
C) Using regularization to reduce overfitting
D) Using oversampling to balance the dataset

=> D: Incorrect (Correct is B)

6) B – Decreasing the class probability threshold makes the model more sensitive and, therefore, marks more cases as the positive class, which is fraud in this case. This will increase the likelihood of fraud detection. However, it comes at the price of lowering precision. This is covered in the Discussion section of the paper at this link.

https://academic.oup.com/bib/article/14/1/13/304457
When the class sizes are very different, most standard classification algorithms may favor the larger (majority) class resulting in poor accuracy in the minority class prediction.

https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/
https://www.jeremyjordan.me/imbalanced-data/

https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

7) A company is interested in building a fraud detection model. Currently, the data scientist does not have a sufficient amount of information due to the low number of fraud cases.

Which method is MOST likely to detect the GREATEST number of valid fraud cases?

A) Oversampling using bootstrapping
B) Undersampling
C) Oversampling using SMOTE
D) Class weight adjustment

=> C: Correct

7) C – With datasets that are not fully populated, the Synthetic Minority Over-sampling Technique (SMOTE) adds new information by adding synthetic data points to the minority class. This technique would be the most effective in this scenario. Refer to Section 4.2 at this link for supporting information.

8) A machine learning engineer is preparing a data frame for a supervised learning task with the Amazon SageMaker Linear Learner algorithm. The ML engineer notices the target label classes are highly imbalanced and multiple feature columns contain missing values. The proportion of missing values across the entire data frame is less than 5%.

What should the ML engineer do to minimize bias due to missing values?

A) Replace each missing value by the mean or median across non-missing values in same row.
B) Delete observations that contain missing values because these represent less than 5% of the data.
C) Replace each missing value by the mean or median across non-missing values in the same column.
D) For each feature, approximate the missing values using supervised learning based on other features.

=> B (or C if we worry about imbalanced datasets and want to retain more rows): Incorrect (Correct is D)

8) D – Use supervised learning to predict missing values based on the values of other features. Different supervised learning approaches might have different performances, but any properly implemented supervised learning approach should provide the same or better approximation than mean or median approximation, as proposed in responses A and C. Supervised learning applied to the imputation of missing values is an active field of research. Refer to this link for an example.

9) A company has collected customer comments on its products, rating them as safe or unsafe, using decision trees. The training dataset has the following features: id, date, full review, full review summary, and a binary safe/unsafe tag. During training, any data sample with missing features was dropped. In a few instances, the test set was found to be missing the full review text field.

For this use case, which is the most effective course of action to address test data samples with missing features?

A) Drop the test samples with missing full review text fields, and then run through the test set.
B) Copy the summary text fields and use them to fill in the missing full review text fields, and then run through the test set.
C) Use an algorithm that handles missing data better than decision trees.
D) Generate synthetic data to fill in the fields that are missing data, and then run through the test set.

=> B: Correct

9) B – In this case, a full review summary usually contains the most descriptive phrases of the entire review and is a valid stand-in for the missing full review text field. For supporting information, refer to page 1627 at this link, and this link and this link.

10) An insurance company needs to automate claim compliance reviews because human reviews are expensive and error-prone. The company has a large set of claims and a compliance label for each. Each claim consists of a few sentences in English, many of which contain complex related information. Management would like to use Amazon SageMaker built-in algorithms to design a machine learning supervised model that can be trained to read each claim and predict if the claim is compliant or not.

Which approach should be used to extract features from the claims to be used as inputs for the downstream supervised task?

A) Derive a dictionary of tokens from claims in the entire dataset. Apply one-hot encoding to tokens found in each claim of the training set. Send the derived features space as inputs to an Amazon SageMaker built-in supervised learning algorithm.
B) Apply Amazon SageMaker BlazingText in Word2Vec mode to claims in the training set. Send the derived features space as inputs for the downstream supervised task.
C) Apply Amazon SageMaker BlazingText in classification mode to labeled claims in the training set to derive features for the claims that correspond to the compliant and non-compliant labels, respectively.
D) Apply Amazon SageMaker Object2Vec to claims in the training set. Send the derived features space as inputs for the downstream supervised task.

=>C because we still need to classify compliant or not: Incorrect (Correct is D)

10) D – Amazon SageMaker Object2Vec generalizes the Word2Vec embedding technique for words to more complex objects, such as sentences and paragraphs. Since the supervised learning task is at the level of whole claims, for which there are labels, and no labels are available at the word level, Object2Vec needs be used instead of Word2Vec. For supporting information, refer to this link and this link.