Machine Learning Pipeline: Solutions to Assessments and Quizzes

Pre- and Post-Assessment

1 - Of the following, which is an example of machine learning? (Select TWO.) **[Correct Answer: B, C]**

A. Calculating the shortest route from current location to the destination

B. Optimizing product pricing based on real-time sales data

C. Sentiment analysis of text on product reviews

D. A loan approval system that classifies applicants entirely based on credit score

2 - Which of the following is an appropriate use case for unsupervised learning? **[Correct Answer: C]**

A. Partitioning an image of a street scene into multiple segments

B. Finding an optimal path out of a maze

C. Identifying clusters of housing sales based on related data points

D. Analyzing sentiment of social media posts

3 - A retail company wants to deploy a machine learning model to predict the demand for a product using sales data from the past 5 years.

What is the MOST efficient solution that the company should implement first? **[Correct Answer: A]**

A. Regression

B. Multi-class classification

C. Binary class classification

4 - In which phase of the ML pipeline do you analyze the business requirements and re-frame that information into a machine learning context. **[Correct Answer: A]**

A. Problem formulation

B. Model training

C. Deployment

D. Data preprocessing

5 - Which feature of Amazon SageMaker can you use to learn patterns in data? **[Correct Answer: B]**

A. SageMaker notebook instances

B. SageMaker training jobs

C. SageMaker hyperaparameter tuning

D. SageMaker endpoints

|  |
| --- |
| 2 |

6 - A retail company wants to start personalizing product recommendations to visitors of their website. They have historical data of what products the users have purchased and want to implement the system for new users, prior to them purchasing a product. What's one way of phrasing a machine learning problem for this situation? **[Correct Answer: D]**

A. Predict if a user will buy the product based on all the other products they bought

B. Classify the products to categories to recommend the user based on the most common category bought

C. Predict the next item that user will buy based on every the products all users bought

D. Predict the item the user will buy based on other users' purchase history and the product ratings and reviews

7 - A Data Scientist at a retail company is using Amazon SageMaker to classify social media posts that mention the company into one of two categories: Posts that require a response from the company, and posts that do not. The Data Scientist is using a training dataset of 10,000 posts, each of which contain the timestamp, author, and full text of each post. However, the Data Scientist is missing the target labels that are required for training.

Which approach can the Data Scientist take to create valid target label data? (Select TWO.) **[Correct Answer: A, C]**

A. Ask the social media handling team to review each post using Amazon SageMaker GroundTruth and provide the label

B. Use the sentiment analysis natural language processing library to determine whether a post requires a response

C. Use Amazon Mechanical Turk to publish Human Intelligence Tasks that ask Turk workers to label the posts

D. Use the a priori probability distribution of the two classes. Then, use Monte-Carlo simulation to generate the labels

E. Use K-Means to cluster posts into various groups, and pick the most frequent word in each group as its label

8 - An ad tech company is using an XGBoost model to classify its clickstream data. The company’s Data Scientist is asked to explain how the model works to a group of non-technical colleagues. **[Correct Answer: A]**

What is a simple explanation the Data Scientist can provide?

A. XGBoost is an Extreme Gradient Boosting algorithm that is optimized for boosted decision trees

B. XGBoost is a logistic regression algorithm to split each feature of the data and used for classification problem

C. XGBoost is a robust, flexible, scalable algorithm that uses linear regression and used for regression problems

D. XGBoost is an efficient and scalable neural network architecture

Double click to edit

10 - You are preprocessing a credit card fraud dataset that contains the column 'Amount'. You want to determine what the average amount is for the cases where there is fraud. Which basic descriptive statistic could you use to achieve this? **[Correct Answer: D]**

A. Mode

B. Median

C. Standard Deviation

|  |
| --- |
| 3 |

D. Mean

Double click to edit

12 - An ML Engineer at a real estate startup wants to use a new quantitative feature for an existing ML model that predicts housing prices. Before adding the feature to the cleaned dataset, the Engineer wants to visualize the feature in order to check for outliers and overall distribution and skewness of the feature.

What visualization technique should the ML Engineer use? (Select TWO.) **[Correct Answer: A, B]**

A. Box Plot

B. Histogram

C. Scatterplot

D. Heatmap

E. T-SNE

13 - When performing an exploratory data analysis on a dataset that will be used for a logistic regression model, you've realized that you have one categorical feature with missing values. You decided to deal with that by imputing new data for the missing values. You know that the feature has this distribution for the valid categories: 10% of missing values, 5% of 'A' value, 70% of 'B' value and 15% of 'C' value. **[Correct Answer: B]**

What would be the best approach for this imputation?

A. Impute 'A' for 50% of the missing values and 'B' for the rest.

B. Impute 'B' for all missing values

C. Impute 'B' for 50% of the missing values and 'C' for the rest.

D. Impute randomly one of the three values (A, B or C)

14 - A tech startup is building an image classification model. During the process, they copied over some of their validation data into their training examples, creating duplicate values in the training and validation subsets.

Which is a possible result of taking this approach? (Select TWO.) **[Correct Answer: A, C]**

A. The model may perform worse with the test dataset than with the validation dataset

B. This is a common practice in machine learning and will improve the overall performance of the model

C. This could lead to overfitting the model

D. This is a good way to increase the training dataset size and therefore strengthen the model's ability to generalize

15 - When updating your weights using the loss function, what dictates how much change the weights should have? **[Correct Answer: B]**

A. Batch size

B. Learning rate

C. Initial weights

|  |
| --- |
| 4 |

D. Bias term

16 - You are tasked to answer whether a company should email a particular customer based on past customer responses to a marketing campaign. You will create this machine learning model using XGBoost.

What Amazon SageMaker option should the company use to train their ML models that reduces the management and automates the pipeline for future retraining? **[Correct Answer: B]**

A. Create and train your XGBoost algorithm on your local laptop and then use an Amazon SageMaker endpoint to host the ML model.

B. Use Amazon in-built algorithms to run the training and inference jobs.

C. Use the Build Your Own Container (BYOC) Amazon SageMaker option. Create a new Docker container with the existing code. Register the container in Amazon Elastic Container Registry (ECR). Finally, run the training and inference jobs using this container.

D. Create a new Amazon SageMaker notebook instance. Copy the existing code into an Amazon SageMaker notebook. Then, run the pipeline from this notebook.

17 - A Data Scientist at a credit card company trained a classification model to predict fraud at the time of a transaction. The Data Scientist used a confusion matrix to evaluate the performance of the model.

Using the confusion matrix below, determine the percent of positive records that were classified correctly. Choose the answer that also labels this evaluation metric correctly. **[Correct Answer: A]**

                                True Positive True Negative

 Predicted Positive         100         90

 Predicted Negative         25              250

A. 80%; Recall

B. 52.6%; Recall

C. 80%; Precision

D. 52.6%; Precision

18 - A real estate company is building a linear regression model to predict housing prices for different cities in the US. Which of the following is NOT a good metric to measure performance of their regression model? **[Correct Answer: B]**

A. R-Squared value

B. F1 score

C. Mean-squared error

D. Mean absolute error

19 - A data scientist is trying to determine how a model is doing based on training evaluation. The train accuracy plateaus out at around 70% and the validation accuracy is 67%. How should the data scientist interpret these results? **[Correct Answer: A]**

|  |
| --- |
| 5 |

A. The model is underfit and needs more complexity

B. The model is overfit and needs less complexity

C. The model is generalizing well and isn't overfit or underfit

D. The model is overfit and underfit and needs more epochs

20 - Does model performance always increase as you add more features to your data?  **[Correct Answer: C]**

A.Yes, the model performance always increases

B. No, model performance always decreases

C. The model performance may increase, but it can lead to overfitting

D. The model performance may increase, but it can lead to underfitting

21 - A real estate startup wants to use ML to predict the value of homes in various cities. To do so, the startup's data science team is joining real estate price data with other variables such as weather, demographic, and standard of living data. However, the team is having problems with slow model convergence. Additionally, the model includes large weights for some features, which is causing degradation in model performance. What kind of feature engineering technique should the team use to more effectively prepare this data? **[Correct Answer: A]**

A. Standard scaler

B. Normalizer

C. Max absolute scaler

D. One hot encoder

22 - When preparing the dataset for your machine learning model, you should use one-hot encoding on what type of data? **[Correct Answer: C]**

A. Continuous

B. Ordinal

C. Nominal

D. Numerical

23 - A Data Scientist wants to tune the hyperparameters of a machine learning model to improve the model’s F1 score.

Which technique CANNOT be used in Amazon SageMaker to solve this problem? **[Correct Answer: A]**

A. Grid Search

B. Random Search

C. Bayesian optimization

24 - Which inference type should you use if you have requirements for latency and predictions that are based on dynamic features? **[Correct Answer: A]**

|  |
| --- |
| 6 |

A. Online inference

B. Batch inference

25 - A team of Data Scientists wants to use Amazon SageMaker training jobs to run two different versions of the same model in parallel to compare the long-term effectiveness of the different versions in reaching the related business outcome.

How should the team deploy these two model versions with minimum management? **[Correct Answer: B]**

A. Create a Lambda function that preprocesses the incoming data, calls a single Amazon SageMaker endpoints for the two models, and finally returns the prediction.

B. Create an endpoint configuration with production variants for the two models with equal weights.

C. Create an endpoint configuration with production variants for the two models with a weight ratio of 90:10.

D. Create a Lambda function that downloads the models from Amazon S3 and calculates and returns the predictions of the two models.

Quiz 1

1 - Which of the following is NOT an example of machine learning? **[Correct Answer: B]**

A. Estimating home prices based on attributes of the property

B. Sorting department store items by bar code

C. Identifying objects in an image based on training data

D. Recommending products based on users' buying habits

2 - Which of the following is most suitable for supervised learning? **[Correct Answer: A]**

A. Identifying birds in an image

D. Grouping people into smaller groups based on buying habits

C. Reducing the number of features in a data set

D. Identifying anomalies in your data to label credit card transactions as fraudulent

3 - A real estate company wants to provide its customers with a more accurate prediction of the final sale price for houses they are considering in various cities. To do this, the company wants to use a fully connected neural network trained on data from the previous ten years of home sales, as well as other features. What kind of machine learning problem does this situation most likely represent? **[Correct Answer: A]**

A. Regression

D. Binary classification

C. Recommender system

D. Multiclass classification

4 - In which phase of the ML pipeline does the machine learn from the data? **[Correct Answer: B]**

|  |
| --- |
| 7 |

A. Data preprocessing

B. Model training

C. Feature engineering

D. Deployment

5 - Which feature of Amazon SageMaker can you use for preprocessing the data? **[Correct Answer: A]**

A. SageMaker notebook instances

B. SageMaker training jobs

C. SageMaker hyperaparameter tuning

D. SageMaker endpoints

6 - What setting, when creating an Amazon SageMaker notebook instance, can you use to install libraries and import data? **[Correct Answer: A]**

A. Lifecycle configuration

B. Default Git repositories

C. VPC

D. IAM role

7 - You work for the largest coffee chain in the world. You've recently decided to source beans from a new market to create new blends and flavors. These beans come from 30 different growers, in 3 different countries. In order to keep a consistent flavor, you have each grower send samples of their beans to your tasting baristas who rate the beans on 20 different dimensions. You now need to group the beans together so the supply can be diversified yet the flavor of the final product kept as consistent as possible. What is one way you could convert this business situation into a machine learning problem? **[Correct Answer: C]**

A. Predict a numerical output for different types of beans based on past ratings of similar beans

B. Categorize different types of beans into multiple categories indicating different types of flavors

C. Group similar beans together based on ratings by baristas

D. Categorize beans into one of two categories indicating the two main flavors represented

Quiz 2

1 - A data scientist wants to visualize the correlation between features in their dataset. What tool(s) can they use to visualize this in a correlation matrix? (Select TWO) **[Correct Answer: B, D]**

A. Pandas

B. Matplotlib

C. Scikitlearn

D. Seaborn

2 - You are preprocessing a dataset that includes categorical features. You want to determine which categories of

|  |
| --- |
| 8 |

particular features are most common in your dataset. Which basic descriptive statistic could you use? **[Correct Answer: A]**

A. Mode

B. Median

C. Standard Deviation

D. Mean

3 - You've plotted the correlation matrix of your dataset's features and realized that two of the features present a high negative correlation (-0.95). What should you do? **[Correct Answer: D]**

A. Do nothing because correlated features won't affect the model performance

B. Do nothing as the correlation is not positive

C. Remove both features

D. Remove one of the features

4 - You are in charge of preprocessing the data your publishing company wants to use for a new ML model they’re building, which aims to predict the influence an academic journal will have in its field. The preprocessing step is necessary to prepare the data for model training.

What type of issue with the data might you encounter during this preprocessing phase. (Select TWO.) **[Correct Answer: A, C]**

A. Outliers

B. Overfit data

C. Missing values

D. Residuals

E. Insufficient batch size

5 - A Machine Learning Engineer is creating and preparing data for a linear regression model. However, while preparing the data, the Engineer notices that about 20% of the numerical data contains missing values in the same two columns. The shape of the data is 500 rows by 4 columns, including the target column. **[Correct Answer: C, D]**

How can the Engineer handle the missing values in the data? (Select TWO.)

A. Remove the rows containing the missing values

B. Remove the columns containing the missing values

C. Fill the missing values with mean of the column

D. Impute the missing values using regression

E. Add regularization to the model

6 - A Data Scientist created a correlation matrix between nine variables and the target variable. The correlation coefficient between two of the numerical variables, variable 1 and variable 5, is -0.95. How should the Data Scientist interpret the correlation coefficient? **[Correct Answer: B]**

|  |
| --- |
| 9 |

A. As variable 1 increases, variable 5 increases

B. As variable 1 increases, variable 5 decreases

C. Variable 1 does not have any influence on variable 5

D. The data is not sufficient to make a well-informed interpretation

7 - An advertising and analytics company uses machine learning to predict user response to online advertisements using a custom XGBoost model. The company wants to improve its ML pipeline by porting its training and inference code, written in R, to Amazon SageMaker, and do so with minimal changes to the existing code.

How should the company set up this new pipeline?**[Correct Answer: C]**

A. Use the Amazon pre-built R container option and port the existing code over to the container. Register the container in Amazon Elastic Container Registry (Amazon ECR). Finally, run the training and inference jobs using this container.

B. Use Amazon in-built algorithms to run their training and inference jobs.

C. Use the Build Your Own Container (BYOC) Amazon SageMaker option. Create a new Docker container with the existing code. Register the container in Amazon Elastic Container Registry (ECR). Finally, run the training and inference jobs using this container.

D. Create a new Amazon SageMaker notebook instance. Copy the existing code into an Amazon SageMaker notebook. Then, run the pipeline from this notebook.

8 - An ML engineer at a text analytics startup wants to develop a text classification model. The engineer collected large amounts of data to develop a supervised text classification model. The engineer is getting 99% accuracy on the dataset but when the model is deployed to production, it performs significantly worse.

What is the most likely cause of this? **[Correct Answer: B]**

A. The engineer did not correctly deploy the model to production

B. The engineer did not split the data to validate the model on unseen data

C. The engineer did not use the right machine learning algorithm for this use case

D. The engineer did not scale the data before training

9 - For a classification problem, what does the loss function measure? **[Correct Answer: B]**

A. A loss function measures the steps you need to take to reach the optimal point

B. A loss function measures how accurate your prediction is with respect to the true values

C. A loss function calculates the iterations that you will take to reach your goal

D. A loss function calculates the gradients of the weights for a particular iteration

10 - Gradient Descent is an important optimization method. Which one of the following is TRUE about the gradient descent method? (Select THREE) **[Correct Answer: A, B, E]**

A. It can involve multiple iterations

|  |
| --- |
| 10 |

B. It uses learning rate to multiply the effect of gradients

C. It uses a batch of training data points to calculate its next step

D. It is more computationally efficient to use for large data sets

E. It tries to find the minimum of a loss function

Quiz 3

1 - A text analytics company is developing a text classification model to detect whether a document involves offensive content or not. The training dataset included ten non-offensive documents for every one offensive document. Their model resulted in an accuracy score of 94%.

 What can we conclude from this result? **[Correct Answer: C]**

A. Accuracy is the right metric of choice here as it resulted in a good model performance

B. Accuracy is the wrong metric of choice here because the model's performance is clearly not as high as it should be for this situation

C. Accuracy is the wrong metric of choice here because it can be heavily influenced by the large class (non-offensive documents)

D. Accuracy is the right metric of choice when dealing with unbalanced data like this

2 - A Machine Learning Engineer is creating a regression model for forecasting company revenue based on an internal dataset made up of past sales and other related data?

What metric should the Engineer use to evaluate the ML model? **[Correct Answer: C]**

A. Cross-entropy log loss

B. Sigmoid

C. Root Mean squared error (RMSE)

D. Precision

3 - An ML scientist has built a decision tree model using scikit-learn with 1,000 trees. The training accuracy for the model was 99.2% and the test accuracy was 70.3%. Should the Scientist use this model in production? **[Correct Answer: C]**

A. Yes, because it is generalizing well on the training set

B. No, because it is generalizing well on the training set

C. No, because it is not generalizing well on the test set

D. Yes, because it is not generalizing well on the test set

4 - The curse of dimensionality relates to which of the following?

**[Correct Answer: A]**

A. A high number of features in a dataset

B. A low number of features in a dataset

|  |
| --- |
| 11 |

6 - What is the primary reason that one might want to pick either random search or Bayesian optimization over grid search when performing hyperparameter optimization? **[Correct Answer: A]**

A. Random search and Bayesian methods leave smaller unexplored regions than grid searches

B. Random search can be parallelized, whereas grid search cannot

C. Deterministic policies are unable to handle uncertainty in the data

D. Random search can explore more parameter space than grid search

7 - A Data Scientist trained an XGBoost model to classify internal documents for further inquiry, and now wants to evaluate the model's performance by looking at the results visually.

What technique should the Data Scientist use in this situation? **[Correct Answer: C]**

A. Scatterplot to visualize the predicted labels versus the true label

B. Correlation matrix to visualize the predicted labels versus the true label

C. Confusion matrix to visualize the predicted labels

D. Box plot to visualize the predicted labels versus true labels

C. A high number of datapoints

D. A low number of datapoints

E. It tries to find the minimum of a loss function

5 - A Data Scientist wants to include “month” as a categorical column in a training dataset for an ML model that is being built. However, the ML algorithm gives an error when the column is added to the training data.

What should the Data Scientist do to add this column? **[Correct Answer: A]**

A. Convert the “month” column to 12 different columns, one for each month, by using one-hot encoding

B. Map the “month" column data to the numbers 1 to 12 and use this new numerical mapped column

C. Scale the months using StandardScaler

D. Use pandas fillna() to convert the column to numerical data