DATABRICKS-MACHINE-LEARNING-ASSOCIATE$^{Q\&As}$

Databricks Certified Machine Learning Associate Exam

Pass Databricks DATABRICKS-MACHINE-LEARNING-ASSOCIATE Exam with 100% Guarantee

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

A data scientist has produced two models for a single machine learning problem. One of the models performs well when one of the features has a value of less than 5, and the other model performs well when the value of that feature is greater than or equal to 5. The data scientist decides to combine the two models into a single machine learning solution.

Which of the following terms is used to describe this combination of models?

A. Bootstrap aggregation

B. Support vector machines

C. Bucketing

D. Ensemble learning

E. Stacking

Correct Answer: D

Ensemble learning is a machine learning technique that involves combining several models to solve a particular problem. The scenario described fits the concept of ensemble learning, where two models, each performing well under different

conditions, are combined to create a more robust model. This approach often leads to better performance as it combines the strengths of multiple models.

References:

Introduction to Ensemble Learning:

https://machinelearningmastery.com/ensemble-machine-learning-algorithms-python-scikit-learn/

QUESTION 2

The implementation of linear regression in Spark ML first attempts to solve the linear regression problem using matrix decomposition, but this method does not scale well to large datasets with a large number of variables.

Which of the following approaches does Spark ML use to distribute the training of a linear regression model for large data?

A. Logistic regression

B. Singular value decomposition

C. Iterative optimization

D. Least-squares method

Correct Answer: C

For large datasets, Spark ML uses iterative optimization methods to distribute the training of a linear regression model.

Specifically, Spark MLlib employs techniques like Stochastic Gradient Descent (SGD) and Limited-memory Broyden

QUESTION 3

A data scientist is developing a single-node machine learning model. They have a large number of model configurations to test as a part of their experiment. As a result, the model tuning process takes too long to complete. Which of the following approaches can be used to speed up the model tuning process?

A. Implement MLflow Experiment Tracking

B. Scale up with Spark ML

C. Enable autoscaling clusters

D. Parallelize with Hyperopt

Correct Answer: D

To speed up the model tuning process when dealing with a large number of model configurations, parallelizing the hyperparameter search using Hyperopt is an effective approach. Hyperopt provides tools likeSparkTrialswhich can run

hyperparameter optimization in parallel across a Spark cluster.

Example:

```
fromhyperoptimportfmin, tpe, hp, SparkTrials search_space = {\\'x\\': hp.uniform(\\'x\\',0,1),\\'y\\':

hp.uniform(\\'y\\',0,1) }defobjective(params):returnparams[\\'x\\'] **2+ params[\\'y\\'] **2spark_trials =
SparkTrials(parallelism=4) best = fmin(fn=objective, space=search_space, algo=tpe.suggest, max_evals=100,
trials=spark_trials) References:
```

Hyperopt Documentation

QUESTION 4

A data scientist has been given an incomplete notebook from the data engineering team. The notebook uses a Spark DataFrame spark_df on which the data scientist needs to perform further feature engineering. Unfortunately, the data scientist has not yet learned the PySpark DataFrame API.

Which of the following blocks of code can the data scientist run to be able to use the pandas API on Spark?

A. import pyspark.pandas as ps df = ps.DataFrame(spark_df)

B. import pyspark.pandas as ps df = ps.to_pandas(spark_df)

C. spark_df.to_pandas()

D. import pandas as pd df = pd.DataFrame(spark_df)

Correct Answer: A

To use the pandas API on Spark, the data scientist can run the following code block:

importpyspark.pandasasps df = ps.DataFrame(spark_df) This code imports the pandas API on Spark and converts the Spark DataFramespark_df into a pandas-on-Spark DataFrame, allowing the data scientist to use familiar pandas functions

for further feature engineering.

References:

Databricks documentation on pandas API on Spark: pandas API on Spark

QUESTION 5

An organization is developing a feature repository and is electing to one-hot encode all categorical feature variables. A data scientist suggests that the categorical feature variables should not be one-hot encoded within the feature repository.

Which of the following explanations justifies this suggestion?

A. One-hot encoding is a potentially problematic categorical variable strategy for some machine learning algorithms.

B. One-hot encoding is dependent on the target variable\\'s values which differ for each apaplication.

C. One-hot encoding is computationally intensive and should only be performed on small samples of training sets for individual machine learning problems.

D. One-hot encoding is not a common strategy for representing categorical feature variables numerically.

Correct Answer: A

The suggestion not to one-hot encode categorical feature variables within the feature repository is justified because one-hot encoding can be problematic for some machine learning algorithms. Specifically, one-hot encoding increases the dimensionality of the data, which can be computationally expensive and may lead to issues such as multicollinearity and overfitting. Additionally, some algorithms, such as tree-based methods, can handle categorical variables directly without requiring one-hot encoding. References: Databricks documentation on feature engineering: Feature Engineering