Microsoft

DP-100

Designing and Implementing a Data Science Solution on Azure

http://killexams.com/pass4sure/exam-detail/DP-100
Question: 114

You use the Two-Class Neural Network module in Azure Machine Learning Studio to build a binary classification model. You use the Tune Model Hyperparameters module to tune accuracy for the model. You need to select the hyperparameters that should be tuned using the Tune Model Hyperparameters module. Which two hyperparameters should you use? Each correct answer presents part of the solution. Each correct selection is worth one point. NOTE:

A. Number of hidden nodes
B. Learning Rate
C. The type of the normalizer
D. Number of learning iterations
E. Hidden layer specification

Answer: D,E.

Question: 115

HOTSPOT - You are evaluating a Python NumPy array that contains six data points defined as follows: data = [10, 20, 30, 40, 50, 60] You must generate the following output by using the k-fold algorithm implementation in the Python Scikit-learn machine learning library: train: [10 40 50 60], test: [20 30] train: [20 30 40 60], test: [10 50] train: [10 20 30 50], test: [40 60] You need to implement a cross-validation to generate the output. How should you complete the code segment? To answer, select the appropriate code segment in the dialog box in the answer area. Each correct selection is worth one point. NOTE: Hot Area:

Answer: See Explanation

Explanation:

Question: 116

You create a binary classification model by using Azure Machine Learning Studio. You must tune hyperparameters by performing a parameter sweep of the model. The parameter sweep must meet the following requirements: ? iterate all possible combinations of hyperparameters ? minimize computing resources required to perform the sweep

You need to perform a parameter sweep of the model. Which parameter sweep mode should you use?

A. Random sweep
B. Sweep clustering
C. Entire grid
D. Random grid
E. Random seed D

Answer: See Explanation

Explanation:
You are building a recurrent neural network to perform a binary classification. The training loss, validation loss, training accuracy, and validation accuracy of each training epoch has been provided. You need to identify whether the classification model is overfitted. Which of the following is correct?

A. The training loss stays constant and the validation loss stays on a constant value and close to the training loss value when training the model.
B. The training loss decreases while the validation loss increases when training the model.
C. The training loss stays constant and the validation loss decreases when training the model.
D. The training loss increases while the validation loss decreases when training the model.

Answer: B

Question: 118

Question HOTSPOT - You need to use the Python language to build a sampling strategy for the global penalty detection models. How should you complete the code segment? To answer, select the appropriate options in the answer area. Each correct selection is worth one point. NOTE: Hot Area:

Answer: See Explanation

Explanation:
https://github.com/pytorch/pytorch/blob/master/torch/utils/data/distributed.py

Question: 119

Introductory Info Case study - Overview - You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules. Datasets - There are two datasets in CSV format that contain property details for two cities, London and Paris, with the following columns:

<table>
<thead>
<tr>
<th>Column heading</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CapitaCrimeRate</td>
<td>per capita crime rate by town</td>
</tr>
<tr>
<td>Zoned</td>
<td>proportion of residential land zoned for lots over 25,000 square feet</td>
</tr>
<tr>
<td>NonRetailAcres</td>
<td>proportion of retail business acres per town</td>
</tr>
<tr>
<td>NextToRiver</td>
<td>proximity of the property to the river</td>
</tr>
<tr>
<td>NitrogenOxideConcentration</td>
<td>nitric oxides concentration (parts per 10 million)</td>
</tr>
<tr>
<td>AvgRoomsPerHouse</td>
<td>average number of rooms per dwelling</td>
</tr>
<tr>
<td>Age</td>
<td>proportion of owner-occupied units built prior to 1940</td>
</tr>
<tr>
<td>DistanceToEmploymentCenter</td>
<td>weighted distances to employment centers</td>
</tr>
<tr>
<td>AccessibilityToHighway</td>
<td>index of accessibility to radial highways to a value of two decimal places</td>
</tr>
<tr>
<td>Tax</td>
<td>full value property tax rate per $10,000</td>
</tr>
<tr>
<td>PupilTeacherRatio</td>
<td>pupil to teacher ratio by town</td>
</tr>
<tr>
<td>ProfessionalClass</td>
<td>professional class percentage</td>
</tr>
<tr>
<td>LowerStatus</td>
<td>percentage lower status of the population</td>
</tr>
<tr>
<td>MedianValue</td>
<td>median value of owner-occupied homes in $1000s</td>
</tr>
</tbody>
</table>

The two datasets have been added to Azure Machine Learning Studio as separate datasets and included as the starting point of the experiment. Dataset issues - The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values. Columns in each dataset contain missing and null values. The dataset also contains many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column. The MedianValue and AvgRoomsPerHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail. Model fit - The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting. Experiment requirements - You must set up the experiment to cross-validate the
Linear Regression and Bayesian Linear Regression modules to evaluate performance. In each case, the predictor of the dataset is the column named MedianValue. An initial investigation showed that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in numerical format, whereas the larger London dataset contains the MedianValue in numerical format. You must ensure that the datatype of the MedianValue column of the Paris dataset matches the structure of the London dataset. You must prioritize the columns of data for predicting the outcome. You must use non-parameters statistics to measure the relationships.

You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns. Model training - Given a trained model and a test dataset, you need to compute the permutation feature importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings. You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful. You are concerned that the model might not efficiently use compute resources in hyperparameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, you need to implement an early stopping criterion on models that provides savings without terminating promising jobs. Testing - You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio. You must create three equal partitions for cross-validation. You must also configure the cross-validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. The data that identifies that a property is near a river is held in the column named NextToRiver. You want to complete this task before the data goes through the sampling process. When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent. Data visualization - You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results. You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another. HOTSPOT - You need to replace the missing data in the AccessibilityToHighway columns. How should you configure the Clean Missing Data module? To answer, select the appropriate options in the answer area. Each correct selection is worth one point. NOTE: Hot Area:

Answer: See Explanation

Explanation:

Question: 120

Introductory Info Case study - Overview - You are a data scientist in a company that provides data science for professional sporting events. Models will use global and local market data to meet the following business goals: Understand sentiment of mobile device users at sporting events based on audio from crowd reactions. Assess a user's tendency to respond to an advertisement. Customize styles of ads served on mobile devices. Use video to detect penalty events Current environment - Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared using social media. The images and videos will have varying sizes and formats. The data available for model building comprises of seven years of sporting event media. The sporting event media includes; recorded video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events. Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats. Penalty detection and sentiment - Data scientists must build an intelligent solution by using multiple machine learning models for penalty event detection. Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines. Notebooks must be deployed to retrain by using Spark instances with dynamic worker allocation. Notebooks must execute with the same code on new Spark instances to recode only the source of the data. Global penalty detection models must be trained by using dynamic runtime graph computation during training. Local penalty detection models must be written by using BrainScript. Experiments for local crowd sentiment models must combine local penalty detection data. Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds. All shared features for local models are continuous variables. Shared features must use double precision. Subsequent layers must have aggregate running mean and standard deviation metrics available. Advertisements - During the initial weeks in production, the following was observed: Ad response rate declined. Drops were not consistent across ad styles. The distribution of features across training and production data are not consistent Analysis shows that, of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrelated features. Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models. All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow. Audio samples show that the length of a catch phrase varies between 25%-47% depending on region The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases. Ad response models must be trained at the beginning of each event and applied during the sporting event. Market segmentation models must optimize for similar ad response history. Sampling must guarantee mutual and collective exclusively between local and global segmentation models that share the same features. Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement. Ad response models must support non-linear boundaries of features. The ad propensity model uses a cut threshold is 0.45 and retraining occur if weighted Kappa deviated from 0.1 +/- 5%. The ad propensity model uses cost factors shown in the following diagram:
DRAG DROP - You need to define an evaluation strategy for the crowd sentiment models. Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order. Select and Place:

Answer: See Explanation

Explanation:

Question: 121

Introductory Info Case study - Overview - You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules. Datasets - There are two datasets in CSV format that contain property details for two cities, London and Paris, with the following columns:
The two datasets have been added to Azure Machine Learning Studio as separate datasets and included as the starting point of the experiment. Dataset issues - The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values. Columns in each dataset contain missing and null values. The dataset also contains many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column. The MedianValue and AvgRoomsinHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail. Model fit - The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting. Experiment requirements - You must set up the experiment to cross-validate the Linear Regression and Bayesian Linear Regression modules to evaluate performance. In each case, the predictor of the dataset is the column named MedianValue. An initial investigation showed that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the MedianValue in numerical format. You must ensure that the datatype of the MedianValue column of the Paris dataset matches the structure of the London dataset. You must prioritize the columns of data for predicting the outcome. You must use non-parameters statistics to measure the relationships. You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsinHouse columns. Model training - Given a trained model and a test dataset, you need to compute the permutation feature importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings. You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful. You are concerned that the model might not efficiently use compute resources in hyperparameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, you need to implement an early stopping criterion on models that provides savings without terminating promising jobs. Testing - You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio. You must create three equal partitions for cross-validation. You must also configure the cross-validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. The data that identifies that a property is near a river is held in the column named NextToRiver. You want to complete this task before the data goes through the sampling process. When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent. Data visualization - You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results. You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another. HOTSPOT - You need to set up the Permutation Feature Importance module according to the model training requirements. Which properties should you select? To answer, select the appropriate options in the answer area. Each correct selection is worth one point. NOTE: Hot Area:

Answer: See Explanation

Introductory Info Case study - Overview - You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules. Datasets - There are two datasets in CSV format that contain property details for two cities, London and Paris, with the following columns: The two datasets have been added to Azure Machine Learning Studio as separate datasets and included as the starting point of the experiment. Dataset issues - The
AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values. Columns in each dataset contain missing and null values. The dataset also contains many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column. The MedianValue and AvgRoomsInHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail. Model fit - The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting. Experiment requirements - You must set up the experiment to cross-validate the Linear Regression and Bayesian Linear Regression modules to evaluate performance. In each case, the predictor of the dataset is the column named MedianValue. An initial investigation showed that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the MedianValue in numerical format. You must ensure that the datatype of the MedianValue column of the Paris dataset matches the structure of the London dataset. You must prioritize the columns of data for predicting the outcome. You must use non-paramaters statistics to measure the relationships. You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns. Model training - Given a trained model and a test dataset, you need to compute the permutation feature importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings. You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful. You are concerned that the model might not efficiently use compute resources in hyperparameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, you need to implement an early stopping criterion on models that provides savings without terminating promising jobs. Testing - You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio. You must create three equal partitions for cross-validation. You must also configure the cross-validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. The data that identifies that a property is near a river is held in the column named NextToRiver. You want to complete this task before the data goes through the sampling process. When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent. Data visualization - You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results. You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another. Question HOTSPOT - You need to configure the Feature Based Feature Selection module based on the experiment requirements and datasets. How should you configure the module properties? To answer, select the appropriate options in the dialog box in the answer area. Each correct selection is worth one point. NOTE: Hot Area:

Answer: See Explanation

Explanation:

Question: 123

Introductory Info Case study - Overview - You are a data scientist in a company that provides data science for professional sporting events. Models will use global and local market data to meet the following business goals: Understand sentiment of mobile device users at sporting events based on audio from crowd reactions. Assess a user's tendency to respond to an advertisement. Customize styles of ads served on mobile devices. Use video to detect penalty events Current environment - Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared using social media. The images and videos will have varying sizes and formats. The data available for model building comprises of seven years of sporting event media. The sporting event media includes; recorded video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events. Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats. Penalty detection and sentiment - Data scientists must build an intelligent solution by using multiple machine learning models for penalty event detection. Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines. Notebooks must be deployed to retrain by using Spark instances with dynamic worker allocation. Notebooks must execute with the same code on new Spark instances to recode only the source of the data. Global penalty detection models must be trained by using dynamic runtime graph computation during training. Local penalty detection models must be written by using BrainScript. Experiments for local crowd sentiment models must combine local penalty detection data. Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds. All shared features for local models are continuous variables. Shared features must use double precision. Subsequent layers must have aggregate running mean and standard deviation metrics available. Advertisements - During the initial weeks in production, the following was observed: Ad response rating declined. Drops were not consistent across ad styles. The distribution of features across training and production data are not consistent Analysis shows that, of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrelated features. Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models. All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow. Audio samples show that the length of a catch phrase varies between 25%-47% depending on region The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases. Ad response models must be trained at the beginning of each event and applied during the sporting event. Market segmentation models must optimize for similar ad response history. Sampling must guarantee mutual and collective exclusively between
local and global segmentation models that share the same features. Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement. Ad response models must support non-linear boundaries of features. The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated from 0.1 +/- 5%. The ad propensity model uses cost factors shown in the following diagram:

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

The ad propensity model uses proposed cost factors shown in the following diagram:

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Performance curves of current and proposed cost factor scenarios are shown in the following diagram:

![Performance curves diagram]

Question DRAG DROP - You need to define a modeling strategy for ad response. Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order. Select and Place:

Answer: See Explanation

Explanation:

Question: 124

You are creating a binary classification by using a two-class logistic regression model. You need to evaluate the model results for imbalance. Which evaluation metric should you use?
A. Relative Absolute Error
B. AUC Curve
C. Mean Absolute Error
D. Relative Squared Error

Answer: B

Question: 125
HOTSPOT - You are using a decision tree algorithm. You have trained a model that generalizes well at a tree depth equal to 10. You need to select the bias and variance properties of the model with varying tree depth values. Which properties should you select for each tree depth? To answer, select the appropriate options in the answer area. Hot Area:
Answer: See Explanation

Explanation:

Question: 126
DRAG DROP - You have a model with a large difference between the training and validation error values. You must create a new model and perform cross-validation. You need to identify a parameter set for the new model using Azure Machine Learning Studio. Which module you should use for each step? To answer, drag the appropriate modules to the correct steps. Each module may be used once or more than once, or not at all. You may need to drag the split bar between panes or scroll to view content. Each correct selection is worth one point. NOTE: Select and Place:
Answer: See Explanation

Explanation:

Question: 127
OTSPOT - You are evaluating a Python NumPy array that contains six data points defined as follows: data = [10, 20, 30, 40, 50, 60] You must generate the following output by using the k-fold algorithm implantation in the Python Scikit-learn machine learning library: train: [10 40 50 60], test: [20 30] train: [20 30 40 60], test: [10 50] train: [10 20 30 50], test: [40 60] You need to implement a cross-validation to generate the output. How should you complete the code segment? To answer, select the appropriate code segment in the dialog box in the answer area. Each correct selection is worth one point. NOTE: Hot Area:
Answer: See Explanation

Explanation:
For More exams visit https://killexams.com/vendors-exam-list

Kill your exam at First Attempt....Guaranteed!