4.0 Pandas Groupby, along with their answers and justifications:

1. Question:

What is the purpose of the `groupby` method in pandas?

Options:

A. To perform numerical calculations on the entire DataFrame.

B. To filter out specific rows based on a condition.

C. To group rows of data together based on a categorical column.

D. To rename columns in the DataFrame.

Answer: C

Explanation: The `groupby` method in pandas is used to group rows of data together based on a categorical column. It allows us to perform aggregate functions on the grouped data.

2. Question:

Which of the following is an example of an aggregate function that can be used with the `groupby` method?

Options:

A. add

B. multiply

C. mean

D. subtract

Answer: C

Explanation: The `mean` function is an example of an aggregate function that can be used with the `groupby` method. It calculates the average value for each group of data.

3. Question:

What does the following code do?

```python
df.groupby('Year').mean().sort_index(ascending=False)
```

Options:

A. Groups the DataFrame by the 'Year' column and calculates the mean for each group.

B. Sorts the DataFrame in descending order based on the 'Year' column.

C. Calculates the mean of the 'Year' column and sorts it in descending order.

D. Calculates the mean of the 'Year' column and sorts the result in ascending order.

Answer: A

Explanation: The code groups the DataFrame by the 'Year' column and calculates the mean value for each group. The `sort_index` function is used to sort the results based on the 'Year' column in descending order.

4. Question:

Which of the following is NOT an aggregate function available in pandas?

Options:

A. count

B. mode

C. median

D. normalize

Answer: D

Explanation: The `normalize` function is not an aggregate function available in pandas. The other options (`count`, `mode`, `median`) are valid aggregate functions.

5. Question:

How can you group rows based on multiple columns simultaneously?

Options:

A. Using the `group` method with a single column name.

B. Using the `groupby` method and passing a list of column names.

C. Using the `groupby` method with multiple arguments.

D. Using the `group` method and providing a tuple of column names.

Answer: B

Explanation: You can group rows based on multiple columns simultaneously by using the `groupby` method and passing a list of column names as an argument.

6. Question:

What is the purpose of the `describe` method when used with the `groupby` operation?

Options:

A. To generate a summary statistics of each group.

B. To calculate the total number of groups in the DataFrame.

C. To display descriptive statistics of individual columns in the DataFrame.

D. To visualize the distribution of data in each group.

Answer: A

Explanation: The `describe` method, when used with the `groupby` operation, generates a summary statistics of each group. It provides information like count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values for each group.

7. Question:

Which method is used to sort the output of the `groupby` operation based on specific columns?

Options:

A. `sort`

B. `order`

C. `sort_values`

D. `sortby`

Explanation: The `sort_values` method is used to sort the output of the `groupby` operation based on specific columns. It allows you to specify the column by which you want to sort the data.

8. Question:

If you want to find the maximum value of the 'Completions' column for each year, which aggregate function should you use with `groupby`?

Options:

A. `max`

B. `mean`

C. `sum`

D. `std`

Answer: A

Explanation: To find the maximum value of the 'Completions' column for each year, you should use the `max` aggregate function with `groupby`.

9. Question:

What is the result of the following code?

```python
df.groupby(['Year', 'Sector']).mean()
```

Options:

A. Calculates the mean of the 'Year' column.

B. Groups the DataFrame by both 'Year' and 'Sector' columns and calculates the mean for each group.

C. Calculates the mean of the 'Sector' column.

D. Groups the DataFrame by the 'Year' column only.

Answer: B

Explanation: The code groups the DataFrame by both 'Year' and 'Sector' columns and calculates the mean value for each group. It provides a hierarchical index with 'Year' and 'Sector' as the levels.

10. Question:

Which aggregate function can be used to find the total number of observations in each 'Year' group?

Options:

A. `count`

B. `mean`

C. `sum`

D. `median`

Answer: A

Explanation: The `count` aggregate function can be used to find the total number of observations in each 'Year' group. It returns the count of non-null values for each group.

5.0 Pandas Operations, along with their answers and justifications:

1. Question:

What is the purpose of the `unique` method in pandas?

Options:

A. To remove duplicate rows from a DataFrame.

B. To count the occurrences of unique values in a column.

C. To create new columns with custom functions.

D. To retrieve the unique values from a column.

Answer: D

Explanation: The `unique` method in pandas is used to retrieve the unique values from a column in a DataFrame. It returns an array containing the distinct values present in that column.

2. Question:

How can you determine the number of unique values in a column of a DataFrame?

Options:

A. Use the `count` method.

B. Use the `nunique` method.

C. Use the `unique` method.

D. Use the `value_counts` method.

Answer: B

Explanation: The `nunique` method is used to determine the number of unique values in a column of a DataFrame. It returns the count of distinct values present in that column.

3. Question:

Which method is used to count the occurrences of each unique value in a column?

Options:

A. `value_counts`

B. `unique`

C. `nunique`

D. `count`

Answer: A

Explanation: The `value_counts` method is used to count the occurrences of each unique value in a column. It returns a Series containing the counts of each distinct value.

4. Question:

Suppose you have a DataFrame named `df_one` with columns 'k1', 'col1', and 'col2.' You want to drop the duplicate rows based on all columns. Which method should you use?

Options:

A. `drop_duplicates`

B. `unique`

C. `drop`

D. `remove_duplicates`

Answer: A

Explanation: The `drop_duplicates` method is used to drop duplicate rows from a DataFrame. By default, it considers all columns to identify duplicates, and only the first occurrence of each duplicate row is kept.

5. Question:

What is the purpose of the `map` method in pandas?

Options:

A. To create new columns with custom functions.

B. To transform numerical data into categorical data.

C. To retrieve the unique values from a column.

D. To replace specific values in a column with new values based on a dictionary.

Answer: D

Explanation: The `map` method in pandas is used to replace specific values in a column with new values based on a dictionary or a Series. It allows you to perform custom value replacement.

6. Question:

How can you find the index position of the minimum value in a specific column of a DataFrame?

Options:

A. `min`

B. `idxmin`

C. `max`

D. `idxmax`

Answer: B

Explanation: The `idxmin` method is used to find the index position of the minimum value in a specific column of a DataFrame.

7. Question:

What is the result of applying the `sort_values` method on the DataFrame `df_one` with respect to the 'C3' column?

Options:

A. DataFrame sorted in ascending order based on the 'C3' column.

B. DataFrame sorted in descending order based on the 'C3' column.

C. DataFrame sorted in ascending order based on the 'C2' column.

D. DataFrame sorted in descending order based on the 'C2' column.

Answer: A

Explanation: The `sort_values` method, when applied to the DataFrame `df_one` with respect to the 'C3' column, will sort the DataFrame in ascending order based on the values in the 'C3' column.

8. Question:

What is the purpose of the `concat` function in pandas?

Options:

A. To join two DataFrames based on common columns.

B. To concatenate the rows of two DataFrames.

C. To perform arithmetic operations on two DataFrames.

D. To create dummy variables from a categorical column.

Answer: B

Explanation: The `concat` function in pandas is used to concatenate the rows of two or more DataFrames. It allows you to combine DataFrames along a particular axis (rows or columns).

9. Question:

Which method is used to create dummy variables from a categorical column in a DataFrame?

Options:

A. `map`

B. `concat`

C. `get_dummies`

D. `unique`

Answer: C

Explanation: The `get_dummies` method is used to create dummy variables from a categorical column in a DataFrame. It converts the categorical column into a set of binary columns for each unique category.

10. Question:

What is the purpose of the `apply` method in pandas?

Options:

A. To filter rows based on specific conditions.

B. To transform numerical data into categorical data.

C. To create new columns with custom functions.

D. To drop duplicate rows from a DataFrame.

Answer: C

Explanation: The `apply` method in pandas is used to create new columns in a DataFrame with custom functions. It allows you to apply a function to each element or row of a column and create a new column based on the function's result.

6.0 Pandas Data Input Output, along with their answers and justifications:

1. Question:

Which method is used to read data from a CSV file into a pandas DataFrame?

Options:

A. `read_csv`

B. `read_excel`

C. `read_json`

D. `read_html`

Answer: A

Explanation: The `read_csv` method in pandas is used to read data from a CSV file and store it into a DataFrame.

2. Question:

If your .csv file is located in a different directory, how should you pass the file path to the `read_csv` method?

Options:

A. Pass the entire file path as a string to the `read_csv` method.

B. Use the `pwd` command to set the file path.

C. Place the .csv file in the same directory as your Python script or Jupyter notebook.

D. Use the `ls` command to list the files in the current directory.

Answer: A

Explanation: If the .csv file is located in a different directory, you should pass the entire file path as a string to the `read_csv` method to correctly read the file.

3. Question:

Which method is used to write a pandas DataFrame to a CSV file?

Options:

A. `write_csv`

B. `to_csv`

C. `save_csv`

D. `export_csv`

Answer: B

Explanation: The `to_csv` method in pandas is used to write a pandas DataFrame to a CSV file.

4. Question:

What should be set to the `index` parameter in the `to_csv` method to exclude the index column from the CSV file?

Options:

A. `index=True`

B. `index=False`

C. `index=0`

D. `index=None`

Answer: B

Explanation: The `index` parameter in the `to_csv` method should be set to `False` to exclude the index column from the CSV file.

5. Question:

Which file format is used for comma-separated values?

Options:

A. JSON

B. HTML

C. Excel

D. CSV

Answer: D

Explanation: CSV stands for Comma-Separated Values, which is a file format for storing tabular data separated by commas.

6. Question:

If your firewall is blocking pandas from accessing the internet, will the `read_html` method still work?

Options:

A. Yes, it will work.

B. No, it will not work.

C. It depends on the website being accessed.

D. `read_html` method does not access the internet.

Answer: B

Explanation: If the firewall is blocking pandas from accessing the internet, the `read_html` method will not work as it requires internet access to read tables from a webpage.

7. Question:

What should you install to enable the `read_html` method to work properly?

Options:

A. xlrd and openpyxl

B. lxml and html5lib

C. htmllib5 and BeautifulSoup4

D. numpy and pandas

Answer: B

Explanation: To enable the `read_html` method to work properly, you should install the `lxml`, `html5lib`, and `beautifulsoup4` libraries.

8. Question:

Which method is used to read tables from a webpage and return a list of DataFrame objects?

Options:

A. `read_html`

B. `read_csv`

C. `read_excel`

D. `read_json`

Answer: A

Explanation: The `read_html` method in pandas is used to read tables from a webpage and return a list of DataFrame objects.

9. Question:

How can you exclude the index column when writing a DataFrame to a CSV file using the `to_csv` method?

Options:

A. `to_csv(index=True)`

B. `to_csv(include_index=False)`
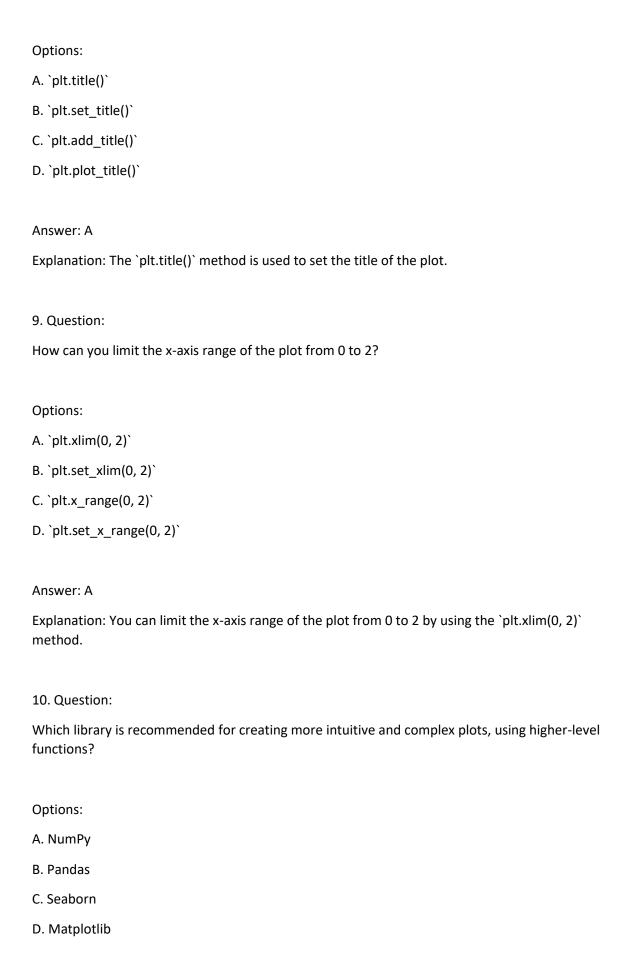
C. `to_csv(with_index=False)`

D. `to_csv(index=False)`

Answer: D

Explanation: To exclude the index column when writing a DataFrame to a CSV file, you should use the `to_csv(index=False)` parameter.

10. Question:

Which online resource provides the best reference for how to read/write to a variety of data sources in pandas?

Options:

A. pandas.org

B. scikit-learn.org

C. pytorch.org

D. https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html

Answer: D

Explanation: The best online resource for how to read/write to a variety of data sources in pandas is available at https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html.

1.0 Matplotlib Basics, along with their answers and justifications:

1. Question:

Which library is used for visualizing plots and is the foundation for Pandas plotting and Seaborn?

Options:

A. NumPy

B. Pandas

C. Matplotlib

D. Seaborn

Answer: C

Explanation: Matplotlib is the library used for visualizing plots and serves as the foundation for Pandas plotting and Seaborn.

2. Question:

What is the purpose of `%matplotlib inline` in Jupyter Notebook?

Options:

A. It enables inline plotting in the Jupyter Notebook.

B. It imports the Matplotlib library.

C. It creates a new figure for plotting.

D. It sets the style of the plot.

Answer: A

Explanation: `%matplotlib inline` is used in Jupyter Notebook to enable inline plotting, meaning that plots will be displayed directly below the code cells.

3. Question:

Which method is used to create a line plot in Matplotlib?

Options:

A. `plt.scatter()`

B. `plt.plot()`

C. `plt.bar()`

D. `plt.hist()`

Answer: B

Explanation: The `plt.plot()` method is used to create a line plot in Matplotlib.

4. Question:

When running a `.py` file, what should you add at the end of your Matplotlib commands to display the plot?

Options:

A. `plt.display()`

B. `plt.show()`

C. `plt.plot()`

D. `plt.save()`

Answer: B

Explanation: When running a `.py` file, you should add `plt.show()` at the end of your Matplotlib commands to display the plot.

5. Question:

What method is used to create a scatter plot in Matplotlib?

Options:

A. `plt.bar()`

B. `plt.scatter()`

C. `plt.plot()`

D. `plt.hist()`

Answer: B

Explanation: The `plt.scatter()` method is used to create a scatter plot in Matplotlib.

6. Question:

What is the purpose of the `marker` parameter in the `plt.plot()` method?

Options:

A. It sets the color of the plot.

B. It adds markers to the data points on the plot.

C. It specifies the linestyle of the plot.

D. It controls the size of the plot.

Answer: B

Explanation: The `marker` parameter in the `plt.plot()` method is used to add markers (symbols) to the data points on the plot.

7. Question:

How can you set the color of the plot to red?

Options:

A. `plt.color('red')`

B. `plt.linecolor('red')`

C. `plt.plot(color='red')`

D. `plt.plot(color='r')`

Answer: D

Explanation: You can set the color of the plot to red by using the `color='r'` parameter in the `plt.plot()` method.

8. Question:

Which method is used to set the title of the plot?

Options:

A. `plt.title()`

B. `plt.set_title()`

C. `plt.add_title()`

D. `plt.plot_title()`

Answer: A

Explanation: The `plt.title()` method is used to set the title of the plot.

9. Question:

How can you limit the x-axis range of the plot from 0 to 2?

Options:

A. `plt.xlim(0, 2)`

B. `plt.set_xlim(0, 2)`

C. `plt.x_range(0, 2)`

D. `plt.set_x_range(0, 2)`

Answer: A

Explanation: You can limit the x-axis range of the plot from 0 to 2 by using the `plt.xlim(0, 2)` method.

10. Question:

Which library is recommended for creating more intuitive and complex plots, using higher-level functions?

Options:

A. NumPy

B. Pandas

C. Seaborn

D. Matplotlib

Answer: C

Explanation: Seaborn is recommended for creating more intuitive and complex plots, using higher-level functions, built on top of Matplotlib.

2.0 Seaborn Basics, along with their answers and justifications:

1. Question:

Which library is used for creating distribution plots, count plots, box plots, scatter plots, and pair plots?

Options:

A. Pandas

B. NumPy

C. Seaborn

D. Matplotlib

Answer: C

Explanation: Seaborn is the library used for creating distribution plots, count plots, box plots, scatter plots, and pair plots.

2. Question:

What type of plot displays the distribution of a single continuous variable and shows the density estimate?

Options:

A. Scatter Plot

B. Pair Plot

C. Histogram

D. Box Plot

Answer: C

Explanation: A Histogram displays the distribution of a single continuous variable and shows the density estimate using bars.

3. Question:

How can you resize a Seaborn plot to have dimensions 12x8?

Options:

A. `sns.resize_plot(12, 8)`

B. `plt.resize(12, 8)`

C. `sns.figsize(12, 8)`

D. `plt.figure(figsize=(12, 8))`

Answer: D

Explanation: You can resize a Seaborn plot to have dimensions 12x8 using `plt.figure(figsize=(12, 8))`.

4. Question:

What type of plot shows the distribution across different categories and uses boxes to represent the quartiles?

Options:

A. Scatter Plot

B. Pair Plot

C. Box Plot

D. Histogram

Answer: C

Explanation: A Box Plot shows the distribution across different categories and uses boxes to represent the quartiles.

5. Question:

In a count plot, which axis represents the category to be counted?

Options:

A. x-axis

B. y-axis

Answer: A

Explanation: In a count plot, the x-axis represents the category to be counted.

6. Question:

What type of plot displays the relationship between two continuous features and can be used to add a third dimension using colors or sizes?

Options:

A. Scatter Plot

B. Pair Plot

C. Count Plot

D. Box Plot

Answer: A

Explanation: A Scatter Plot displays the relationship between two continuous features and can be used to add a third dimension using colors or sizes.

7. Question:

Which function is used to create a pair plot in Seaborn?

Options:

A. `sns.pairplot()`

B. `sns.scatterplot()`

C. `sns.histplot()`

D. `sns.boxplot()`

Answer: A

Explanation: The function `sns.pairplot()` is used to create a pair plot in Seaborn.

8. Question:

What does the `hue` parameter in the `sns.scatterplot()` function do?

Options:

A. It sets the size of the scatter plot markers.

B. It specifies the color palette for the plot.

C. It adds a third dimension to the plot using colors.

D. It adjusts the transparency of the plot markers.

Answer: C

Explanation: The `hue` parameter in the `sns.scatterplot()` function adds a third dimension to the plot using colors, allowing you to differentiate data points based on a categorical variable.

9. Question:

What does the pair plot display along the diagonal when the `hue` parameter is specified?

Options:

A. Scatter plots

B. Histograms

C. Box plots

D. KDE plots

Answer: D

Explanation: When the `hue` parameter is specified in the pair plot, it displays KDE (Kernel Density Estimation) plots along the diagonal instead of histograms.

10. Question:

What is the main drawback of using pair plots for large datasets?

Options:

A. Pair plots do not show any relationship between variables.

B. Pair plots take a long time to create.

C. Pair plots can only be used for categorical data.

D. Pair plots are not supported in Seaborn.

Answer: B

Explanation: The main drawback of using pair plots for large datasets is that they can take a long time to create and the figures may become too small to interpret properly.

1.0 ML Linear Regression, along with their answers and justifications:

Certainly! Here are 10 multiple-choice questions based on the provided information about the Linear Regression Model and the evaluation metrics:

1. Question:

What type of plot is used to visualize the correlation between different numerical features in the dataset?

Options:

A. Histogram

B. Pair Plot

C. Heatmap

D. Scatter Plot

Answer: C

Explanation: A heatmap is used to visualize the correlation between different numerical features in the dataset.

2. Question:

What is the purpose of splitting the dataset into a training set and a testing set in machine learning?

Options:

A. To create new features

B. To preprocess the data

C. To avoid overfitting

D. To increase the number of data points

Answer: C

Explanation: The purpose of splitting the dataset into a training set and a testing set is to avoid overfitting, i.e., to ensure that the model performs well on unseen data.

3. Question:

Which library is used to perform linear regression and train the model?

Options:

A. Matplotlib

B. Seaborn

C. Scikit-learn

D. Pandas

Answer: C

Explanation: Scikit-learn is the library used to perform linear regression and train the model.

4. Question:

What is the interpretation of the coefficient associated with the 'Avg. Area House Age' feature in the linear regression model?

Options:

A. For each unit increase in 'Avg. Area House Age', the price increases by $164883.28.

B. For each unit increase in 'Avg. Area House Age', the price decreases by $164883.28.

C. There is no significant relationship between 'Avg. Area House Age' and the price.

D. The coefficient value cannot be interpreted for 'Avg. Area House Age'.

Answer: A

Explanation: The interpretation of the coefficient associated with the 'Avg. Area House Age' feature is that for each unit increase in 'Avg. Area House Age', the price increases by $164883.28, holding all other features constant.

5. Question:

What evaluation metric is used to measure the average error in a regression model?

Options:

A. Mean Absolute Error (MAE)

B. Root Mean Squared Error (RMSE)

C. Mean Squared Error (MSE)

D. R-squared

Answer: A

Explanation: Mean Absolute Error (MAE) is used to measure the average error in a regression model.

6. Question:

Which evaluation metric is interpretable in the "y" units and is more popular than Mean Squared Error (MSE)?

Options:

A. R-squared

B. Mean Absolute Error (MAE)

C. Root Mean Squared Error (RMSE)

D. Mean Squared Error (MSE)

Answer: C

Explanation: Root Mean Squared Error (RMSE) is interpretable in the "y" units and is more popular than Mean Squared Error (MSE) due to its interpretability.

7. Question:

What is the purpose of the scatter plot comparing predictions with actual values?

Options:

A. To visualize the model's coefficients

B. To check the linearity of the data

C. To assess the residual errors of the model

D. To evaluate the accuracy of the model

Answer: D

Explanation: The purpose of the scatter plot comparing predictions with actual values is to evaluate the accuracy of the model by visualizing how well the predictions align with the actual data points.

8. Question:

Which evaluation metric for regression problems "punishes" larger errors, making it more useful in the real world?

Options:

A. R-squared

B. Mean Absolute Error (MAE)

C. Mean Squared Error (MSE)

D. Root Mean Squared Error (RMSE)

Answer: C

Explanation: Mean Squared Error (MSE) "punishes" larger errors, making it more useful in the real world as it penalizes larger deviations from the actual values.

9. Question:

What is the purpose of using a heatmap in exploratory data analysis?

Options:

A. To visualize the correlation between numerical features

B. To visualize the distribution of a single numerical feature

C. To compare different regression models

D. To identify outliers in the dataset

Answer: A

Explanation: The purpose of using a heatmap in exploratory data analysis is to visualize the correlation between numerical features, allowing us to understand the relationships between different variables.

10. Question:

What evaluation metric is calculated as the square root of the mean of the squared errors in a regression model?

Options:

A. R-squared

B. Mean Absolute Error (MAE)

C. Mean Squared Error (MSE)

D. Root Mean Squared Error (RMSE)

Answer: D

Explanation: Root Mean Squared Error (RMSE) is calculated as the square root of the mean of the squared errors in a regression model.

2.0 ML KNN, along with their answers and justifications:

Sure! Here are 10 multiple-choice questions based on the provided information about K Nearest Neighbors (KNN) with Python:

1. Question:

What is the primary objective of the K Nearest Neighbors (KNN) algorithm?

Options:

A. To classify data points into different clusters

B. To predict a continuous target variable

C. To predict the class of a new data point based on its features

D. To find the optimal value of K for a given dataset

Answer: C

Explanation: The primary objective of the K Nearest Neighbors (KNN) algorithm is to predict the class of a new data point based on its features by identifying the K nearest data points in the training set.

2. Question:

What is the purpose of standardizing the variables in KNN?

Options:

A. To reduce the dimensionality of the dataset

B. To improve the accuracy of the model

C. To make the data follow a normal distribution

D. To ensure all variables have the same scale and influence on the KNN classifier

Answer: D

Explanation: The purpose of standardizing the variables in KNN is to ensure all variables have the same scale and influence on the KNN classifier. Variables on a large scale can have a much larger effect on the distance between observations, affecting the KNN classifier more than variables on a small scale.

3. Question:

What does the elbow method help determine in the KNN algorithm?

Options:

A. The optimal number of features for the model

B. The optimal value of K for the model

C. The accuracy of the model

D. The number of training and testing data points

Answer: B

Explanation: The elbow method helps determine the optimal value of K for the KNN algorithm. It involves plotting the error rate against different values of K and selecting the K value at the "elbow" point where the error rate starts to stabilize.

4. Question:

What is the significance of the confusion matrix in evaluating a KNN model?

Options:

A. It helps visualize the clusters formed by the KNN algorithm

B. It helps compare the accuracy of the KNN model with other models

C. It provides the count of correctly and incorrectly predicted classes by the model

D. It determines the distance between data points in the feature space

Answer: C

Explanation: The confusion matrix provides the count of correctly and incorrectly predicted classes by the KNN model. It helps evaluate the model's performance by showing the true positive, true negative, false positive, and false negative predictions.

5. Question:

Which parameter specifies the number of nearest neighbors to be considered in the KNN algorithm?

Options:

A. K

B. X_train

C. y_train

D. scaled_features

Answer: A

Explanation: The parameter "K" specifies the number of nearest neighbors to be considered in the KNN algorithm.

6. Question:

What is the main advantage of using the KNN algorithm?

Options:

A. It works well with large datasets

B. It is easy to implement and understand

C. It provides accurate predictions for complex models

D. It requires minimal preprocessing of the data

Answer: B

Explanation: The main advantage of using the KNN algorithm is that it is easy to implement and understand, making it suitable for simple classification tasks.

7. Question:

In the elbow method, what does the "elbow" point represent on the plot?

Options:

A. The optimal number of features for the model

B. The point where the error rate is maximum

C. The point where the error rate starts to decrease rapidly

D. The point where the error rate starts to stabilize or flatten

Answer: D

Explanation: In the elbow method, the "elbow" point on the plot represents the point where the error rate starts to stabilize or flatten. This point is used to determine the optimal value of K for the KNN model.

8. Question:

Which data split is used to evaluate the KNN model's performance?

Options:

A. Train-Test Split

B. K-Fold Cross-Validation

C. Validation Set Split

D. Random Split

Answer: A

Explanation: The Train-Test Split is used to evaluate the KNN model's performance. The dataset is split into a training set and a testing set, and the model is trained on the training set and evaluated on the testing set.

9. Question:

What is the primary goal of the KNN algorithm when selecting the value of K?

Options:

A. To minimize the error rate on the training set

B. To minimize the error rate on the testing set

C. To maximize the accuracy of the model

D. To avoid overfitting

Answer: B

Explanation: The primary goal of the KNN algorithm when selecting the value of K is to minimize the error rate on the testing set, ensuring that the model performs well on unseen data.

10. Question:

What is the maximum error rate observed in the error rate vs. K value plot after K>23?

Options:

A. 0.05

B. 0.06

C. 0.23

D. 1

Answer: B

Explanation: The maximum error rate observed in the error rate vs. K value plot after K>23 is 0.06. It hovers around 0.05-0.06.

1.0 ANN Keras Basics, along with their answers and justifications:

1. Question:

What is the purpose of splitting the dataset into training and testing sets?

Options:

A. To increase the number of training samples

B. To evaluate the model's performance on unseen data

C. To improve the training speed of the model

D. To avoid overfitting the model

Answer: B

Explanation: The purpose of splitting the dataset into training and testing sets is to evaluate the model's performance on unseen data, which helps assess how well the model generalizes to new data.

2. Question:

What is the primary objective of the Keras Syntax Basics example project?

Options:

A. To predict the sale price of gem stones

B. To create a deep learning model with multiple layers

C. To explore the real dataset of gem stones

D. To demonstrate the syntax of TensorFlow 2.0 with Keras

Answer: D

Explanation: The primary objective of the Keras Syntax Basics example project is to demonstrate the syntax of TensorFlow 2.0 with Keras, focusing on creating a simple regression model.

3. Question:

Which function is used to normalize or scale the feature data?

Options:

A. MinMaxScaler()

B. StandardScaler()

C. normalize()

D. scale()

Answer: A

Explanation: The function used to normalize or scale the feature data is MinMaxScaler().

4. Question:

What type of problem does the model.compile() method indicate with loss='mse'?

Options:

A. Multi-class classification problem

B. Binary classification problem

C. Regression problem

D. Clustering problem

Answer: C

Explanation: The model.compile() method indicates a regression problem with loss='mse', as mean squared error (MSE) is a common loss function used in regression tasks.

5. Question:

What is the purpose of evaluating the model on both the training set and the test set?

Options:

A. To ensure the training set is well-distributed

B. To compare the performance of different models

C. To check for overfitting or underfitting

D. To balance the data between training and testing sets

Answer: C

Explanation: The purpose of evaluating the model on both the training set and the test set is to check for overfitting or underfitting. If the model performs significantly better on the training set compared to the test set, it may indicate overfitting.

6. Question:

How many epochs were used for training the model?

Options:

A. 100

B. 150

C. 250

D. 300

Answer: C

Explanation: 250 epochs were used for training the model, as seen in the line of code: `model.fit(X_train,y_train,epochs=250)`.

7. Question:

Which activation function was used for the output node of the model?

Options:

A. sigmoid

B. tanh

C. softmax

D. relu

Answer: No options provided in the given code snippet.

8. Question:

What is the significance of the "Test Y" and "Model Predictions" columns in the pred_df DataFrame?

Options:

A. "Test Y" contains the true labels, and "Model Predictions" contains the predicted labels

B. "Test Y" contains the features, and "Model Predictions" contains the target variable

C. "Test Y" contains the training set, and "Model Predictions" contains the test set

D. "Test Y" contains the true labels, and "Model Predictions" contains the scaled feature data

Answer: A

Explanation: "Test Y" contains the true labels (target variable) from the test set, and "Model Predictions" contains the predicted labels generated by the trained model for the test set.

9. Question:

What evaluation metric is used to measure the average absolute difference between the true labels and the predicted labels?

Options:

A. Mean Squared Error (MSE)

B. Mean Absolute Error (MAE)

C. Root Mean Squared Error (RMSE)

D. Accuracy

Answer: B

Explanation: The evaluation metric used to measure the average absolute difference between the true labels and the predicted labels is Mean Absolute Error (MAE).

10. Question:

Which method is used to save the trained Keras model?

Options:

A. save_model()

B. save()

C. save_model_file()

D. save_model_hdf5()

Answer: B

Explanation: The method used to save the trained Keras model is `model.save('my_model.h5')`.

2.0 ANN Keras Regression, along with their answers and justifications:

1. Question:

What is the primary objective of the Keras Regression Code Along Project?

Options:

A. To predict the sale price of gem stones

B. To explore geographical properties of houses

C. To work on feature engineering and clean the data

D. To build a deep learning model with multiple layers

Answer: C

Explanation: The primary objective of the Keras Regression Code Along Project is to work on feature engineering and clean the data to prepare it for building a regression model using Keras.

2. Question:

What is the purpose of using the MinMaxScaler()?

Options:

A. To convert categorical features to numerical

B. To normalize the target variable 'price'

C. To scale the feature data to a specific range

D. To handle missing values in the dataset

Answer: C

Explanation: The purpose of using the MinMaxScaler() is to scale the feature data to a specific range (usually between 0 and 1) to ensure that all features have similar scales and do not dominate the learning process.

3. Question:

How many features are used as input for training the regression model?

Options:

A. 17

B. 18

C. 19

D. 20

Answer: C

Explanation: 19 features are used as input for training the regression model as seen in the code: `model.add(Dense(19,activation='relu'))`.

4. Question:

What is the purpose of the `model.compile()` method in the Keras regression model?

Options:

A. To add layers to the model

B. To define the optimizer and loss function for training

C. To split the data into training and testing sets

D. To visualize the training progress

Answer: B

Explanation: The `model.compile()` method is used to define the optimizer and loss function for training the Keras regression model.

5. Question:

What evaluation metric is used to measure the average absolute difference between the true labels and the predicted labels?

Options:

A. Mean Squared Error (MSE)

B. Mean Absolute Error (MAE)

C. Explained Variance Score (EVS)

D. Median Absolute Error (MedAE)

Answer: B

Explanation: The evaluation metric used to measure the average absolute difference between the true labels and the predicted labels is Mean Absolute Error (MAE).

6. Question:

How many epochs were used for training the regression model?

Options:

A. 250

B. 300

C. 400

D. 500

Answer: C

Explanation: 400 epochs were used for training the regression model, as seen in the code: `model.fit(x=X_train, y=y_train.values, validation_data=(X_test, y_test.values), batch_size=128, epochs=400)`.

7. Question:

Which layer activation function was used for the output node of the regression model?

Options:

A. sigmoid

B. tanh

C. softmax

D. linear

Answer: D

Explanation: The layer activation function used for the output node of the regression model is linear, which is the default activation function for the regression task.

8. Question:

What is the significance of the "losses" DataFrame generated during model training?

Options:

A. It represents the training loss at each epoch

B. It represents the test loss at each epoch

C. It is used for feature engineering

D. It contains the predicted values of the model

Answer: A

Explanation: The "losses" DataFrame represents the training loss at each epoch, providing insights into the model's learning progress during training.

9. Question:

What is the value of the mean absolute error on the test data?

Options:

A. 200,000

B. 100,000

C. 50,000

D. 25,000

Answer: C

Explanation: The value of the mean absolute error on the test data is 50,000 as computed by the code: `mean_absolute_error(y_test, predictions)`.

10. Question:

What type of data visualization is used to compare actual prices with predicted prices?

Options:

A. Line plot

B. Scatter plot

C. Box plot

D. Histogram

Answer: B

Explanation: A scatter plot is used to compare actual prices with predicted prices, as seen in the code: `plt.scatter(y_test, predictions)`.

3.0 ANN Keras Classification, along with their answers and justifications:

1. Question:

What is the total number of instances in the breast cancer Wisconsin (diagnostic) dataset?

Options:

A. 569

B. 30

C. 212

D. 357

Answer: A

Explanation: The total number of instances in the breast cancer Wisconsin (diagnostic) dataset is 569.

2. Question:

Which attribute has the highest correlation with the target variable 'benign_0__mal_1'?

Options:

A. smoothness (mean)

B. fractal dimension (worst)

C. concave points (mean)

D. compactness (worst)

Answer: C

Explanation: The attribute with the highest correlation with the target variable 'benign_0__mal_1' is 'concave points (mean)' as per the correlation values computed.

3. Question:

What type of classification task is being performed in this project?

Options:

A. Multiclass classification

B. Binary classification

C. Regression

D. Clustering

Answer: B

Explanation: The classification task being performed in this project is binary classification, as the target variable 'benign_0__mal_1' has two classes - WDBC-Malignant and WDBC-Benign.

4. Question:

Which data preprocessing technique is applied to scale the feature data?

Options:

A. StandardScaler

B. MinMaxScaler

C. RobustScaler

D. Normalizer

Answer: B

Explanation: The MinMaxScaler is applied to scale the feature data between 0 and 1.

5. Question:

What activation function is used for the output layer in the classification model?

Options:

A. tanh

B. sigmoid

C. ReLU

D. softmax

Answer: B

Explanation: The activation function used for the output layer in the classification model is sigmoid, as it is commonly used for binary classification tasks.

6. Question:

What is the purpose of using the EarlyStopping callback during model training?

Options:

A. To add dropout layers to the model

B. To prevent overfitting by stopping training when validation loss increases

C. To shuffle the training data at each epoch

D. To save the best model during training

Answer: B

Explanation: The purpose of using the EarlyStopping callback during model training is to prevent overfitting by stopping training when the validation loss starts increasing.

7. Question:

How many units are there in the first hidden layer of the classification model?

Options:

A. 15

B. 30

C. 45

D. 60

Answer: B

Explanation: There are 30 units in the first hidden layer of the classification model, as seen in the code: `model.add(Dense(units=30,activation='relu'))`.

8. Question:

What is the number of epochs used during model training with dropout layers?

Options:

A. 200

B. 400

C. 600

D. 800

Answer: C

Explanation: The number of epochs used during model training with dropout layers is 600.

9. Question:

What evaluation metric is used to measure the performance of the classification model?

Options:

A. F1-score

B. R-squared

C. Mean Absolute Error (MAE)

D. Mean Squared Error (MSE)

Answer: A

Explanation: The classification model's performance is measured using the F1-score, as shown in the classification report.

10. Question:

What is the primary purpose of using dropout layers in the model?

Options:

A. To reduce the number of units in the hidden layers

B. To increase the number of units in the hidden layers

C. To introduce random noise to prevent overfitting

D. To remove specific features from the input data

Answer: C

Explanation: The primary purpose of using dropout layers in the model is to introduce random noise during training, which helps prevent overfitting by reducing the reliance on specific neurons and improving generalization.

1.0 CNN Keras, along with their answers and justifications:

1. Question:

What is the shape of a single image in the MNIST dataset before preprocessing?

Options:

A. (28, 28)

B. (28, 28, 1)

C. (1, 28, 28)

D. (60000, 28, 28)

Answer: A

Explanation: The shape of a single image in the MNIST dataset before preprocessing is (28, 28).

2. Question:

What is the purpose of using "one hot encoding" for the labels in this dataset?

Options:

A. To convert numerical labels into categorical labels

B. To scale the labels between 0 and 1

C. To reduce the number of categories in the labels

D. To convert categorical labels into numerical labels

Answer: A

Explanation: The purpose of using "one hot encoding" for the labels is to convert numerical labels into categorical labels, making them suitable for the CNN model.

3. Question:

What is the range of pixel values in the images after normalization?

Options:

A. (0, 255)

B. (0, 1)

C. (-1, 1)

D. (-255, 255)

Answer: B

Explanation: After normalization, the pixel values in the images are scaled between 0 and 1.

4. Question:

What is the total number of neurons in the dense hidden layer of the CNN model?

Options:

A. 64

B. 128

C. 256

D. 512

Answer: B

Explanation: The dense hidden layer of the CNN model has 128 neurons, as specified in the code.

5. Question:

What activation function is used in the convolutional layer of the CNN model?

Options:

A. ReLU

B. Sigmoid

C. Tanh

D. Leaky ReLU

Answer: A

Explanation: The activation function used in the convolutional layer of the CNN model is ReLU (Rectified Linear Unit).

6. Question:

What is the loss function used during model compilation?

Options:

A. Mean Squared Error (MSE)

B. Binary Crossentropy

C. Categorical Crossentropy

D. Mean Absolute Error (MAE)

Answer: C

Explanation: The loss function used during model compilation is Categorical Crossentropy, suitable for multiclass classification tasks.

7. Question:

What is the purpose of using early stopping during model training?

Options:

A. To add dropout layers to the model

B. To increase the number of epochs for better performance

C. To stop training when the model starts overfitting

D. To increase the number of convolutional layers

Answer: C

Explanation: The purpose of using early stopping during model training is to stop training when the model starts overfitting, preventing it from learning noise in the training data.

8. Question:

How many epochs are used for training the CNN model?

Options:

A. 5

B. 10

C. 15

D. 20

Answer: B

Explanation: The CNN model is trained for 10 epochs, as specified in the code.

9. Question:

What is the evaluation metric used to measure the performance of the CNN model?

Options:

A. Mean Squared Error (MSE)

B. Accuracy

C. R-squared

D. Mean Absolute Error (MAE)

Answer: B

Explanation: The evaluation metric used to measure the performance of the CNN model is Accuracy.

10. Question:

What is the total number of classes in the MNIST dataset?

Options:

A. 5

B. 8

C. 10

D. 15

Answer: C

Explanation: The MNIST dataset contains a total of 10 classes representing digits from 0 to 9.

1.0 RNN Sine Wave, along with their answers and justifications:

1. Question:

What is the purpose of the "MinMaxScaler" used in this code?

Options:

A. To normalize the sine wave data between 0 and 1

B. To convert categorical labels into numerical labels

C. To scale the training and testing data separately

D. To perform feature engineering on the dataset

Answer: A

Explanation: The "MinMaxScaler" is used to normalize the sine wave data between 0 and 1, which is a common preprocessing step to ensure that the data is on a similar scale for training the RNN model.

2. Question:

What is the length of the output sequences used in the "TimeseriesGenerator"?

Options:

A. 2

B. 10

C. 49

D. 50

Answer: D

Explanation: The length of the output sequences used in the "TimeseriesGenerator" is 50, as specified in the code.

3. Question:

What is the number of features used in the LSTM layer of the model?

Options:

A. 1

B. 10

C. 49

D. 50

Answer: A

Explanation: The number of features used in the LSTM layer of the model is 1, as there is only one feature in the time series data (sine wave).

4. Question:

What is the loss function used during model compilation?

Options:

A. Mean Squared Error (MSE)

B. Binary Crossentropy

C. Categorical Crossentropy

D. Mean Absolute Error (MAE)

Answer: A

Explanation: The loss function used during model compilation is Mean Squared Error (MSE), which is commonly used for regression tasks.

5. Question:

How many epochs are used for training the LSTM model?

Options:

A. 5

B. 6

C. 20

D. 50

Answer: C

Explanation: The LSTM model is trained for 20 epochs, as specified in the code.

6. Question:

What is the purpose of using "EarlyStopping" during model training?

Options:

A. To add dropout layers to the model

B. To increase the number of epochs for better performance

C. To stop training when the model starts overfitting

D. To increase the number of LSTM layers

Answer: C

Explanation: The purpose of using "EarlyStopping" during model training is to stop training when the model starts overfitting, which helps prevent unnecessary training and reduces the risk of overfitting.

7. Question:

What is the shape of the input data to the LSTM layer?

Options:

A. (batch_size, length, n_features)

B. (length, n_features, batch_size)

C. (n_features, length, batch_size)

D. (batch_size, n_features, length)

Answer: A

Explanation: The shape of the input data to the LSTM layer is (batch_size, length, n_features), where batch_size is the number of samples in each batch, length is the number of timesteps in each input sequence, and n_features is the number of features in the input data.

8. Question:

What is the activation function used in the LSTM layer?

Options:

A. ReLU

B. Sigmoid

C. Tanh

D. Leaky ReLU

Answer: C

Explanation: The activation function used in the LSTM layer is Tanh (hyperbolic tangent), which is a commonly used activation function in LSTM networks.

9. Question:

What is the final output of the model?

Options:

A. A sequence of time series data

B. A single value prediction

C. A classification label

D. A binary output

Answer: B

Explanation: The final output of the model is a single value prediction, as the problem is framed as a regression task to predict the next value in the time series.

10. Question:

What is the forecasting range used in the code?

Options:

A. 5.0 to 5.1

B. 50.0 to 50.1

C. 50.1 to 55.1

D. 55.0 to 55.1

Answer: C

Explanation: The forecasting range used in the code is 50.1 to 55.1, as specified in the code.

1.0 RL, along with their answers and justifications:

1. Question:

What is the purpose of the OpenAI Gym library mentioned in the note?

Options:

A. To provide a simple interface for control tasks and Markov decision processes

B. To offer a collection of classic control tasks and video games

C. To support continuous control tasks and robotic arm manipulation

D. All of the above

Answer: D

Explanation: The OpenAI Gym library serves the purpose of providing a simple interface for control tasks and Markov decision processes. It also offers a collection of classic control tasks, video games, continuous control tasks, and robotic arm manipulation.

2. Question:

Which statement is true about the environment created using the Maze() class?

Options:

A. It has a large number of states (more than 100)

B. Transitions between states are probabilistic (uncertain)

C. The reward obtained in this environment is always -1

D. It is an image-based environment

Answer: C

Explanation: In the Maze environment created using the Maze() class, all rewards are the same (-1) until the episode concludes. This simplifies the study of the value and action-value functions.

3. Question:

What does the env.step() method return when an action is applied in the environment?

Options:

A. The next state

B. The reward obtained

C. A boolean indicating if the task has been completed

D. All of the above

Answer: D

Explanation: The env.step() method returns a tuple of four objects: the next state, the reward obtained, a boolean indicating if the task has been completed, and any other relevant information in a Python dictionary.

4. Question:

How many actions are there in the Maze environment?

Options:

A. 2

B. 3

C. 4

D. 5

Answer: C

Explanation: In the Maze environment, there are four different actions represented by integers: 0 (move up), 1 (move right), 2 (move down), and 3 (move left).

5. Question:

What is the state space of the Maze environment?

Options:

A. 5x5 elements

B. 10 elements

C. 25 elements

D. 50 elements

Answer: C

Explanation: The state space of the Maze environment consists of 25 elements, representing all possible combinations of rows and columns in the 5x5 maze.

6. Question:

What is a trajectory in the context of reinforcement learning?

Options:

A. The sequence generated by moving from one state to another

B. The return to the beginning of the episode

C. The set of all possible states in the task

D. The probability of an action given the current state

Answer: A

Explanation: A trajectory is the sequence generated by moving from one state to another in reinforcement learning. It consists of states, actions, rewards, and next states.

7. Question:

What is the return associated with a moment in time *t*?

Options:

A. The reward obtained at time *t*

B. The sum of all rewards up to time *t*

C. The discounted sum of all rewards up to time *t*

D. The next state at time *t*

Answer: C

Explanation: The return associated with a moment in time *t* is the discounted sum of all rewards obtained from that moment onward. It is represented as: $G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + ... + \gamma^{T-1} R_T$.

8. Question:

What is the purpose of a policy in reinforcement learning?

Options:

A. To define the environment's transition probabilities

B. To compute the probabilities of all actions given the current state

C. To maximize the return in an episode

D. To minimize the number of states in the environment

Answer: B

Explanation: The purpose of a policy in reinforcement learning is to compute the probabilities of all actions given the current state. It defines the agent's behavior in the environment.

9. Question:

What is the policy function used in the note for the Maze environment?

Options:

A. A deterministic policy that always selects the same action

B. A random policy that selects actions uniformly at random

C. A greedy policy that selects the action with the highest probability

D. A fixed policy that follows a predefined sequence of actions

Answer: B

Explanation: The policy function used in the note for the Maze environment is a random policy that selects actions uniformly at random. It assigns an equal probability to all four actions.

10. Question:

What is the type of action space in the Maze environment?

Options:

A. MultiDiscrete([5 5])

B. MultiDiscrete([0, 1, 2, 3])

C. Discrete(4)

D. Discrete(5)

Answer: C

Explanation: The type of action space in the Maze environment is Discrete(4), which means it consists of four different actions represented by integers 0, 1, 2, and 3.

4.0 Extra – Theory ANN, along with their answers and justifications:

1. Question:

In an Artificial Neural Network (ANN), what is the purpose of the activation function?

Options:

A. To adjust the weights and biases of the neurons during training.

B. To control the learning rate of the optimization algorithm.

C. To compute the gradient of the loss function during backpropagation.

D. To introduce non-linearity and enable the network to learn complex patterns.

Answer: D

Explanation: The activation function introduces non-linearity to the output of each neuron, allowing the network to learn complex patterns and make it capable of approximating any arbitrary function.

2. Question:

Which type of ANN layer connects every neuron from the previous layer to every neuron in the current layer?

Options:

A. Fully Connected Layer (Dense Layer)

B. Convolutional Layer

C. Recurrent Layer

D. Pooling Layer

Answer: A

Explanation: A Fully Connected Layer, also known as a Dense Layer, connects every neuron from the previous layer to every neuron in the current layer, making it the most common type of layer in ANNs.

3. Question:

What is the purpose of the Loss Function in ANN training?

Options:

A. To regularize the weights and prevent overfitting.

B. To compute the accuracy of the model on the training data.

C. To measure the dissimilarity between the predicted output and the true target.

D. To update the learning rate during the optimization process.

Answer: C

Explanation: The Loss Function measures the dissimilarity between the predicted output and the true target, quantifying how well the model is performing on the training data. It serves as a guide for the optimization process.

4. Question:

In the context of ANN training, what is the role of the Backpropagation algorithm?

Options:

A. To update the learning rate during the training process.

B. To initialize the weights and biases of the network.

C. To compute the gradient of the loss function with respect to the network parameters.

D. To decide the number of layers and neurons in the network.

Answer: C

Explanation: The Backpropagation algorithm is responsible for computing the gradient of the loss function with respect to the network parameters (weights and biases). It allows the network to adjust its parameters to minimize the loss during training.

5. Question:

What is the vanishing gradient problem in ANN training?

Options:

A. When the learning rate is too high, causing the model to diverge during training.

B. When the gradients of the activation functions become zero, leading to slow or stalled learning.

C. When the model is unable to generalize well to new, unseen data.

D. When the model suffers from overfitting and fails to perform on the test set.

Answer: B

Explanation: The vanishing gradient problem occurs when the gradients of the activation functions become very close to zero during backpropagation. This leads to slow or stalled learning in deep networks and hinders their ability to learn complex patterns.

6. Question:

Which type of ANN layer is commonly used for handling sequential data, such as time series or natural language?

Options:

A. Fully Connected Layer (Dense Layer)

B. Convolutional Layer

C. Recurrent Layer

D. Pooling Layer

Answer: C

Explanation: Recurrent Layers are commonly used for handling sequential data, such as time series or natural language. They allow information to be passed between time steps, making them suitable for tasks with temporal dependencies.

7. Question:

What is the purpose of regularization techniques in ANN?

Options:

A. To increase the number of layers and neurons in the network.

B. To improve the computational efficiency during training.

C. To prevent overfitting and improve the generalization of the model.

D. To adjust the learning rate during optimization.

Answer: C

Explanation: Regularization techniques, such as L1 and L2 regularization, are used to prevent overfitting and improve the generalization of the model. They add penalty terms to the loss function, encouraging the network to avoid overly complex solutions.

8. Question:

What is the role of the Learning Rate in the training of an ANN?

Options:

A. To determine the number of training iterations.

B. To control the magnitude of the weight updates during backpropagation.

C. To decide the number of neurons in each layer of the network.

D. To compute the gradients of the activation functions.

Answer: B

Explanation: The Learning Rate determines the magnitude of the weight updates during backpropagation. It controls the step size in the direction of the negative gradient, influencing the rate of convergence during training.

9. Question:

Which ANN architecture is designed to process data in a grid-like structure, such as images?

Options:

A. Feedforward Neural Network

B. Convolutional Neural Network (CNN)

C. Recurrent Neural Network (RNN)

D. Autoencoder

Answer: B

Explanation: Convolutional Neural Networks (CNNs) are specifically designed to process data in a grid-like structure, such as images. They utilize convolutional layers to detect local patterns and hierarchical representations.

10. Question:

What is the role of the Activation Function in the output layer of a binary classification ANN?

Options:

A. To introduce non-linearity and enable complex pattern learning.

B. To compute the probability of each class in a multi-class classification task.

C. To squash the output into a range between 0 and 1, representing the probability of the positive class.

D. To prevent the gradients from vanishing during backpropagation.

Answer: C

Explanation: In the output layer of a binary classification ANN, the Activation Function typically used is the Sigmoid function. It squashes the output into a range between 0 and 1, representing the probability of the positive class (class 1).