Machine Learning

1 C

2 B

3 C

4 B

5 B

6 A,D

7 B

8 D

9 B

10 The adjusted R-squared takes into account the number of predictors in a model. It lowers the value of R-squared value based on the number of predictors, it penalizes those predictors which doesn't contribute. R-squared value increases with increase in number of predictors, which draws conclusion that model is accurately fitted but in actual case it is not. This is taken into account in adjusted R-squared.

Adjusted R-squared = 1 - [(1 - R-squared) * (n - 1) / (n - k - 1)]

Where:

n = number of observations

k = number of predictors (excluding the intercept)

R-squared = coefficient of determination, a measure of the goodness of fit of the model.

11 Ridge Regression and Lasso Regression are two common regularization techniques used to tackle overfitting in linear regression models.

Ridge Regression: It is a type of regularization that adds a penalty term to the least square loss function to reduce the magnitude of the coefficients. The penalty term is proportional to the **square** of the magnitude of the coefficients and is controlled by a hyperparameter alpha. The ridge regression encourages the coefficients to be small but does **not** encourage them to be **exactly zero**, which means it does not perform feature selection.

Lasso Regression: It is another type of regularization that adds a penalty term to the least square loss function, but it is proportional to the **absolute value** of the coefficients. Unlike Ridge Regression, Lasso Regression can shrink some **coefficients to zero**, which **performs feature selection**. Lasso Regression is useful in cases where we have a large number of features and only a few of them are really important.

12 VIF stands for Variance Inflation Factor, which is a measure of the extent of multicollinearity in a multiple regression model. VIF is calculated for each predictor in a regression model by dividing the variance of the predictor by the variance of the residuals from a model with all other predictors. The suitable value of VIF for a feature to be included in a regression modeling is typically less than 10. A high VIF value indicates that the feature is highly correlated with other features in the model, which can cause issues with interpreting the model and overfitting. In such cases, the feature with the highest VIF value should be removed from the model and the VIF values of the remaining features should be re-calculated.

$VIF = 1/(1-R^2)$

Where:

R^2 is the coefficient of determination (R-squared) of the regression of the feature with all the other features in the model.

13 Scaling the data is needed before training the model because it helps to bring all features to the same level of magnitude. This helps the model to weigh all the features equally and not be influenced by a particular feature. Without scaling the model has high tendency to be biased towards a feature with high magnitude as it will miss the important relationship between features of smaller magnitude. Scaling the data ensure smooth movement towards minima, as it will help in convergence faster.

14 The different metrics which are used to check the goodness of fit in linear regression are:

- R-squared (Coefficient of Determination): It measures the proportion of variance in the dependent variable that is explained by the independent variables.
- Adjusted R-squared: It takes into account the number of independent variables in the model, thus penalizing the presence of unnecessary predictors.
- Mean Squared Error (MSE): It measures the average of the squared differences between the predicted and actual values of the dependent variable.
- Root Mean Squared Error (RMSE): It is the square root of MSE and it measures the average difference between the predicted and actual values.
- Mean Absolute Error (MAE): It measures the average of the absolute differences between the predicted and actual values.
- Mean Absolute Percentage Error (MAPE): It measures the average of the absolute differences between the predicted and actual values as a percentage of the actual values.

15

- Sensitivity (True Positive Rate or Recall) = True Positives / (True Positives + False Negatives)= 1000 / (1000 + 250) = 0.8

- Specificity (True Negative Rate) = True Negatives / (True Negatives + False Positives) 1200 / (1200 + 50) = 0.96

- Precision (Positive Predictive Value) = True Positives / (True Positives + False Positives) =1000 / (1000 + 50) = 0.95

- Recall (True Positive Rate or Sensitivity) = True Positives / (True Positives + False Negatives) =1000 / (1000 + 250) = 0.8

- Accuracy = (True Positives + True Negatives) / Total = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.9

Statistics

1 D

2 A

3 A

4 C

5 D

6 D

7 C

8 B

9 A

10 A boxplot and a histogram are two different types of visualizations used to represent and summarize a dataset.

A boxplot, also known as a box-and-whisker plot, is a graph that displays the summary of a set of data values having properties of the minimum, first quartile, median, third quartile, and maximum. It is useful in identifying the presence of outliers and skewness in the data.

A histogram, on the other hand, is a bar graph that displays the frequency distribution of a set of continuous or discrete data values. It groups the data into bins, and the height of each bar

represents the frequency of the data that falls into that bin. Histograms provide a visual representation of the distribution of the data, including the shape, central tendency, and spread.

11 The selection of metrics for evaluating a machine learning model depends on various factors such as the type of problem being solved (classification, regression, etc.), the nature of the data, and the desired outcome.

Commonly used metrics for classification problems are accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) curve.

Commonly used metrics for regression problems are mean absolute error (MAE), mean squared error (MSE), and R-squared.

12 To assess the statistical significance of an insight, one commonly uses statistical hypothesis testing.

In hypothesis testing, one formulates **a null hypothesis**, which represents the **baseline assumption**, and an **alternate hypothesis**, which represents the insight under investigation. The null hypothesis is usually that there is no difference or no relationship between the variables under investigation, while the alternate hypothesis is the opposite of the null hypothesis.

Next, one collects data and calculates a test statistic, which summarizes the relationship between the variables in the sample. Based on the test statistic and the sample size, one can calculate the p-value, which is the probability of observing the test statistic under the null hypothesis.

If the p-value is smaller than a predetermined significance level (e.g., 0.05), one can reject the null hypothesis and conclude that the insight is statistically significant, meaning that it is unlikely to have occurred by chance. On the other hand, if the p-value is larger than the significance level, one fails to reject the null hypothesis and concludes that there is insufficient evidence to support the insight.

13 Data that doesn't follow a Gaussian or log-normal distribution can come from many sources, including:

1. Categorical data: For example, the number of people who choose a certain political party or brand of toothpaste.
2. Count data: For example, the number of times a website is visited in a day, or the number of emails sent by a user.
3. Binary data: For example, whether a customer made a purchase or not.
4. Time series data: For example, stock prices, traffic volume, or weather patterns.
5. Non-parametric data: For example, data that follows a Poisson, exponential, or Weibull distribution.

14 The median is often a better measure than the mean **when there are outliers** or extreme values in a data set. For example, consider a data set that represents the income of a group of people, where some people have very high incomes compared to the majority. In this case, the **mean would be heavily influenced by the high incomes**, giving an overestimate of the typical income in the group. On the other hand, the median represents the value that separates the lower and upper halves of the data, and would give a better representation of the typical income in the group.

15 The likelihood is a function that measures the probability of obtaining the observed data given a set of parameters. In statistical modeling, the likelihood is used to estimate the parameters of a **model that best fit the observed data.** The likelihood function is **calculated by multiplying the individual probabilities of observing the data for each data point**, given the set of parameters. The maximum likelihood estimates (MLE) of the parameters are the values that maximize the likelihood function, and are considered as the best estimates of the parameters based on the data.

SQL

1 B, C, D

2 A, C, D

3 B

4 C

5 C

6 B

7 A

8 C

9 D

10 A

11 Denormalization is the process of intentionally adding redundant data to a database to improve performance and simplify queries. This involves grouping data into larger tables, reducing the number of joins needed and duplicating data across tables. By doing so, the database is optimized, the tradeoff is that it may become slower for writes and more difficult to maintain.

12 A database cursor is a pointer to a specific row within a result set of a SQL query. It is a mechanism for iterating over the records of a result set and is used to retrieve data one row at a time. Cursors provide a way to perform operations on a row-by-row basis, which is useful when dealing with large data sets or complex data operations. Cursors can be used in stored procedures, triggers, and other database programming constructs to enable iterative processing of query results.

13 Some of the most common types of SQL queries are:

1. SELECT: This query is used to retrieve data from one or more tables in a database. It allows you to specify the columns you want to retrieve, the tables you want to query, and any filtering or sorting conditions.
2. INSERT: This query is used to add new data to a table in a database. It requires you to specify the table you want to insert data into, as well as the values you want to add.
3. UPDATE: This query is used to modify existing data in a table. It allows you to specify the table you want to update, the columns you want to modify, and the new values you want to set.
4. DELETE: This query is used to remove data from a table in a database. It requires you to specify the table you want to delete data from, as well as any filtering conditions.
5. JOIN: This query is used to combine data from two or more tables in a database. It allows you to specify the columns you want to retrieve from each table, as well as the join conditions that specify how the tables should be combined.
6. GROUP BY: This query is used to group data from a table based on one or more columns. It allows you to perform aggregate functions such as SUM, COUNT, AVG, and MAX on the grouped data.
7. HAVING: This query is used to filter the results of a GROUP BY query based on aggregate functions. It allows you to specify filtering conditions that apply to the grouped data.

14 A constraint is a rule that is enforced on the data in a table. It helps to ensure the accuracy, validity, and reliability of the data, and prevents incorrect or incomplete data from being entered into the database. Constraints can be defined on a single column or a combination of columns in a table, the different type of constraints are:

1. NOT NULL: This constraint ensures that a column cannot contain a null value.
2. UNIQUE: This constraint ensures that the values in a column or a group of columns are unique across all the rows in the table.
3. PRIMARY KEY: This constraint is a combination of NOT NULL and UNIQUE constraints and is used to uniquely identify each row in a table.
4. FOREIGN KEY: This constraint is used to establish a relationship between two tables by referencing the primary key of one table in another table.
5. CHECK: This constraint is used to ensure that the values in a column or a group of columns satisfy a specific condition.
6. DEFAULT: This constraint is used to provide a default value for a column when no value is specified for the column during an insert operation.

15 Auto increment is a feature in SQL that allows a unique number to be generated automatically when a new record is inserted into a table. In MySQL, the AUTO_INCREMENT attribute is used to specify a column as an auto-increment column.

Code:

CREATE TABLE users (id INT AUTO_INCREMENT PRIMARY KEY, name VARCHAR(50) NOT NULL, email VARCHAR(50) NOT NULL);

In this example, the **id** column is an auto-increment column and is also the primary key for the table. When a new record is inserted into the **users** table, a unique number will be automatically generated for the **id** column.