Question Set

Brij, Kirti, Abhishek, Nitesh, Ananda

1. *MSQ* We can remove the missing values from a dataset if

   (A) MCAR
   (B) MAR
   (C) MNAR
   (D) less than 5% data are missing

2. *MSQ* Imputing the missing values by the mean of the data

   (A) increases variance
   (B) decreases variance
   (C) increases peakedness in the density curve of the data
   (D) decreases peakedness in the density curve of the data

3. *MCQ* X_1 and X_2 are two normally distributed random variates with Pearson's correlation coefficient (a measure of linear relationship between two continuous variables) 0. Choose the most appropriate statement.

   (A) X_1 and X_2 are only linearly independent but non-linear or functional dependence might be there.
   (B) Even though Pearson's correlation coefficient is a measure of linear relationship, here X_1 and X_2 are completely independent.
   (C) Value of the correlation coefficient is too little information to conclude anything about independence.
   (D) The joint distribution of X_1 and X_2 can never be bivariate normal.

4. *MSQ* In PCA, we desire the principal components to have

   (A) low variance
   (B) high variance
   (C) low covariance
   (D) high covariance

5. *MCQ* An illegal site's servers were seized in a recent operation. Which of the following queries will submit all users' details sorted by access times in descending order ?

   (A) SELECT * FROM users;
   (B) SELECT * FROM users ORDER BY AccessTime;
   (C) SELECT * FROM users GROUP BY AccessTime;
   (D) SELECT * FROM users ORDER BY AccessTime DESC;

6. *MCQ Movie* table has 3 columns : *Movie_name* (primary key), *release_year*, *genre*. Select the query to find the oldest released movie of each genre.

   (A) SELECT genre, MIN(release_year) FROM movie ORDER BY genre;

   (B) SELECT genre, MIN(release_year) FROM movie GROUP BY genre;

   (C) SELECT genre, MIN(release_year) FROM movie;

   (D) SELECT genre, MIN(release_year) FROM movie COUNT BY genre;

7. *MCQ Employees* table has 2 columns : *id* (primary key) and *name*. Each row of this table contains the id and the name of an employee in a company. *EmployeeUNI* table also has two columns *id* and *unique_id*; (*id*, *unique_id*) is the primary key (combination of columns with unique values) for this table. Each row of this table contains the id and the corresponding unique id of an employee in the company. Select the correct query to show the unique ID of each user, if a user does not have a unique ID replace just show 'null'.

   (A) SELECT unique_id, name FROM Employees AS a INNER JOIN EmployeeUNI AS b ON a.id = b.id;

   (B) SELECT unique_id, name FROM Employees AS a LEFT JOIN EmployeeUNI AS b ON a.id = b.id;

   (C) SELECT unique_id, name FROM Employees AS a RIGHT JOIN EmployeeUNI AS b ON a.id = b.id;

   (D) SELECT id, name FROM Employees AS a LEFT JOIN EmployeeUNI AS b ON a.id = b.id;

8. *MCQ* Why is feature selection important?

   (A) to increase training time

   (B) to make models more complex

   (C) to reduce overfitting and improve accuracy

   (D) to increase the number of predictors

9. *MCQ* Which of the following techniques is used for encoding categorical variables?

   (A) PCA

   (B) Min-Max Scaling

   (C) One-hot Encoding

   (D) Dropout

10. *MCQ* Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and found a relationship between them. Which of the following conclusions do you make about this situation?

    (A) Since there is a relationship means our model is not good

    (B) Since there is a relationship means our model is good

    (C) can't say

    (D) None of these