Worksheet Solution -3

Machine Learning Solutions

1. d) All of the above
2. d) None
3. c) Reinforcement Learning and Unsupervised Learning
4. b) The tree representing how close the data points are to each other
5. d) None
6. c) . k-nearest neighbour is same as k-means
7. d) 1,2 and 3
8. a) 1 only
9. a) 2
10. a) .Given sales data from a large number of products in a supermarket, estimate future sales for each of these products
11. a)
12. b)
13. Clustering is useful for exploring data. If there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings. Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build supervised models.
14. K-means clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm.When the data has overlapping clusters, k-means can improve the results of the initialization technique. When the data has well separated clusters, the performance of k-means depends completely on the goodness of the initialization.

SQL Solutions :

1. CREATE TABLE customers (customerNumber int, customerName varchar(100), contactLastName varchar(100), contactFirstName varchar(100), phone int, addressLine1 varchar(100), addressLine2 varchar(100), city varchar(50), state varchar(50), postalCode int, country varchar(50), salesRepEmployeeNumber int, creditLimit int, FOREIGN KEY(salesRepEmployeeNumber) REFERENCES Employees(employeeNumber), PRIMARY KEY(customerNumber));
2. CREATE TABLE orders(orderNumber int, orderDate Date, requiredDate Date, shippedDate Date, status varchar(100), comments varchar(100), customerNumber int, FOREIGN KEY(customerNumber) REFERENCES Customers(customerNumber), PRIMARY KEY(orderNumber));
3. Select * from orders;
4. Select comments from orders;
5. Select orderDate, count(*) from orders group by orderDate;
6. Select employeeNumber, lastName, firstName from Employees;
7. Select O.orderNumber, C.customerName from Orders O Join Customers C ON C.customerNumber = O.customerNumber;
8. Select C.customerName, E.firstName from Customers C Join Employees E on C.salesRepEmployeeNumber = E.employeeNumber;
9. Select paymentDate, amount from Payments.
10. Select productName, mSRP, productDescription from Products;
11. Select P.productName, P.productDescription, sum(O.quantityOrdered) as Quantity from Products p Join orderDetails O on P.productCode = O.orderNumber group by P.productName order by sum(O.quantityOrdered) desc;
12. Select C.city from Customers C Join Orders O on O.customerNumber = C.customerNumber Join orderDetails D on D.orderNumber = O.orderNumber group by C.city order bu D.quantityOrdered Desc limit 1;

13. Select State, count(*) from Customers Group by State, order by count(*) desc limit 1;
14. Select employeeNumber, concat(firstName,' ', lastName) from employees;
15. Select O.orderNumber, C.customerName, D.quantityOrdered*D.priceEach as Total_Amount_Paid from Orders O Join Customer C on C.customerNumber = O.customerNumber Join orderDetails D on D.orderNumber = O.orderNumber;

Statistics Worksheet Solution:

1. b) Total Variation = Residual Variation + Regression Variation
2. c) binomial
3. a) 2
4. a) Type-I error
5. d) Confidence coeeficient
6. d) None
7. b) Hypothesis
8. d) All of the mentioned
9. a) 0
10. Bayes' theorem is used for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence.
11. A z-score measures exactly how many standard deviations above or below the mean a data point is. It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution
12. A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.
13. a percentile is a number where a certain percentage of scores fall below that number.
14. Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among means.
15. ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.