MACHINE LEARNING

ASSIGNMENT - 4

1 In Q1 to Q7, only one option is correct, Choose the correct option:

1. The value of correlation coefficient will always be:

A) between 0 and 1 B) greater than -1

C) between -1 and 1 D) between 0 and -1

Answer: c

2. Which of the following cannot be used for dimensionality reduction?

A) Lasso Regularisation B) PCA

C) Recursive feature elimination D) Ridge Regularisation

Answer: c

3. Which of the following is not a kernel in Support Vector Machines?

A) linear B) Radial Basis Function

C) hyperplane D) polynomial

Answer: c

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

A) Logistic Regression B) Naïve Bayes Classifier

C) Decision Tree Classifier D) Support Vector Classifier

Answer: c

5. In a Linear Regression problem, 'X' is independent variable and 'Y' is dependent variable, where 'X' represents weight in pounds. If you convert the unit of 'X' to kilograms, then new coefficient of 'X' will be?

(1 kilogram = 2.205 pounds)

A) 2.205 × old coefficient of 'X' B) same as old coefficient of 'X'

C) old coefficient of 'X' ÷ 2.205 D) Cannot be determined

Answer: b

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

A) remains same B) increases

C) decreases D) none of the above

Answer: b

7. Which of the following is not an advantage of using random forest instead of decision trees?

A) Random Forests reduce overfitting

B) Random Forests explains more variance in data then decision trees

C) Random Forests are easy to interpret

D) Random Forests provide a reliable feature importance estimate

Answer: c

In Q8 to Q10, more than one options are correct, Choose all the correct options:

8. Which of the following are correct about Principal Components?

A) Principal Components are calculated using supervised learning techniques

B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

D) All of the above

Answer: d

9. Which of the following are applications of clustering?

A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.

C) Identifying spam or ham emails

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

Answer:b

10. Which of the following is(are) hyper parameters of a decision tree?

A) max_depth B) max_features

C) n_estimators D) min_samples_leaf

Answer:c

Q10 to Q15 are subjective answer type questions, Answer them briefly.

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Answer: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

Q1 represents the 25th percentile of the data.

Q2 represents the 50th percentile of the data.

Q3 represents the 75th percentile of the data.

If a dataset has 2n / 2n+1 data points, then

Q1 = median of the dataset.

Q2 = median of n smallest data points.

Q3 = median of n highest data points.

IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

12. What is the primary difference between bagging and boosting algorithms?

Answer:

Sl.No.	Bagging Algorithms	Boosting Algorithms
1	The simplest way of combining predictions that belong to the same group	A way of combining the predictions that belong to different applications
2	It decreases variance not bias	It decreases bias not variance
3	Each model receives equal weight	The model receives weights baed on the performance
4	It decreases overfitting	It decreases the bias
5	Model is built independently	New models are influenced by the performance of the previous models.
6	The base of the classifiers are trained parallelly	The base of the classifiers are trained parallelly are trained sequencially
7	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
8	Ex: Random forest	Ex: Adaboost

13. What is adjusted R2 in linear regression. How is it calculated?

Answer: Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.R2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of effects are included in the model. Adjusted R2 attempts to correct for this overestimation. Adjusted R2 might decrease if a specific effect does not improve the model. Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R2 is always less than or equal to R2. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0 indicates a model that has no predictive value. In the real world, adjusted R2 lies between these values.

The *coefficient of determination*, or R2R2, is a measure that provides information about the goodness of fit of a model. In the context of regression it is a statistical measure of how well the regression line approximates the actual data. It is therefore important when a statistical model is used either to predict future outcomes or in the testing of hypotheses. There are a number of variants (see comment below); the one presented here is widely used

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

The *sum squared regression* is the sum of the residuals squared, and the *total sum of squares* is the sum of the distance the data is away from the mean all squared. As it is a percentage it will take values between 0 and 1.

14. What is the difference between standardisation and normalisation?

Answer:

Sl.No.	Standardisation	Normalisation
1	Maximum and minimum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

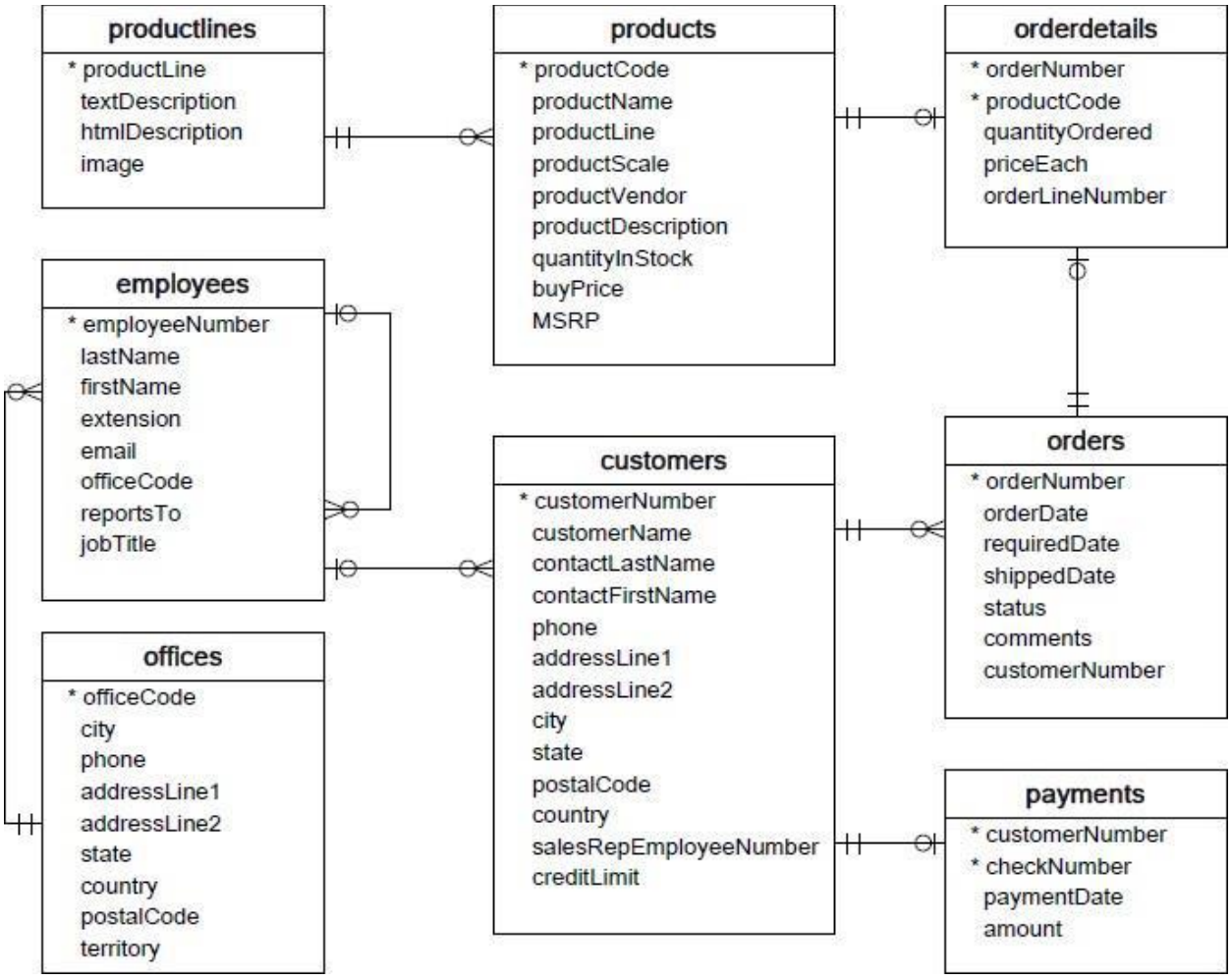
Answer: Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

The advantages are reduces Overfitting: In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

The disadvantages are Increases Training Time and computing efficiency: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.

WORKSHEET 4 SQL

Refer the following ERD and answer all the questions in this worksheet. You have to write the queries using MySQL for the required Operation.



Customers: stores customer's data.
 Products: stores a list of scale model cars.
 Product Lines: stores a list of product line categories.
 Orders: stores sales orders placed by customers.
 Order Details: stores sales order line items for each sales order.
 Payments: stores payments made by customers based on their accounts.
 Employees: stores all employee information as well as the organization structure such as who reports to whom.
 Offices: stores sales office data.

QUESTIONS:

1. Write a SQL query to show average number of orders shipped in a day (use Orders table).

Answer: SELECT AVG (quantityOrdered) FROM orderdetails

2. Write a SQL query to show average number of orders placed in a day.

Answer: SELECT orderDate,AVG (quantityOrdered) FROM orderdetails,orders.

3. Write a SQL query to show the product name with minimum MSRP (use Products table).

Answer: SELECT product name FROM products WHERE MIN(MSRP)

4. Write a SQL query to show the product name with maximum value of stock Quantity.

Answer: SELECT product name FROM products WHERE MAX(stock Quantity)

5. Write a query to show the most ordered product Name (the product with maximum number of orders).

Answer: SELECT product name FROM products,orderdetails WHERE MAX(quantityordered)

6. Write a SQL query to show the highest paying customer Name.

Answer: SELECT customerName FROM customer,payments WHERE MAX(amount)

7. Write a SQL query to show cutomerNumber, customerName of all the customers who are from Melbourne city.

Answer: SELECT cutomerNumber, customerName FROM customer WHERE city==' Melbourne city'

8. Write a SQL query to show name of all the customers whose name start with "N".

Answer: SELECT customerName FROM customer WHERE customerName LIKE "N%"

9. Write a SQL query to show name of all the customers whose phone start with '7' and are from city 'Las Vegas'.

Answer: SELECT customerName FROM customer WHERE phone LIKE "7%" and city== 'Las Vegas'

10. Write a SQL query to show name of all the customers whose creditLimit < 1000 and city is either "Las Vegas" or "Nantes" or "Stavern".

Answer: SELECT customername FROM customers WHERE creditLimit < 1000 and city=="Las Vegas" or "Nantes" or "Stavern"

11. Write a SQL query to show all the orderNumber in which quantity ordered <10.

Answer: SELECT orderNumber FROM orderdetails WHERE quantity ordered <10

12. Write a SQL query to show all the orderNumber whose customer Name start with letter 'N'.

Answer: SELECT customername, orderNumber FROM customers, orders WHERE customer Name LIKE 'N%'

13. Write a SQL query to show all the customerName whose orders are "Disputed" in status.

Answer: SELECT customerName FROM customers,orders WHERE status=="Disputed"

14. Write a SQL query to show the customerName who made payment through cheque with checkNumber starting with H and made payment on "2004-10-19".

Answer: SELECT customerName, checkNumber FROM customers, payments WHERE checkNumber LIKE 'H%' and paymentDate=="2004-10-19"

15. Write a SQL query to show all the checkNumber whose amount > 1000.

Answer: SELECT checkNumber FROM payments WHERE amount>1000

STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

1. What is central limit theorem and why is it important?

Answer: The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.
The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.
Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.
A sufficiently large sample size can predict the characteristics of a population more accurately.
CLT is useful in finance when analyzing a large collection of securities to estimate portfolio distributions and traits for returns, risk, and correlation.
2. What is sampling? How many sampling methods do you know?
Answer: A sample is a subset of individuals from a larger population. Sampling means selecting the group that you will actually collect data from in your research. Sampling methods are:

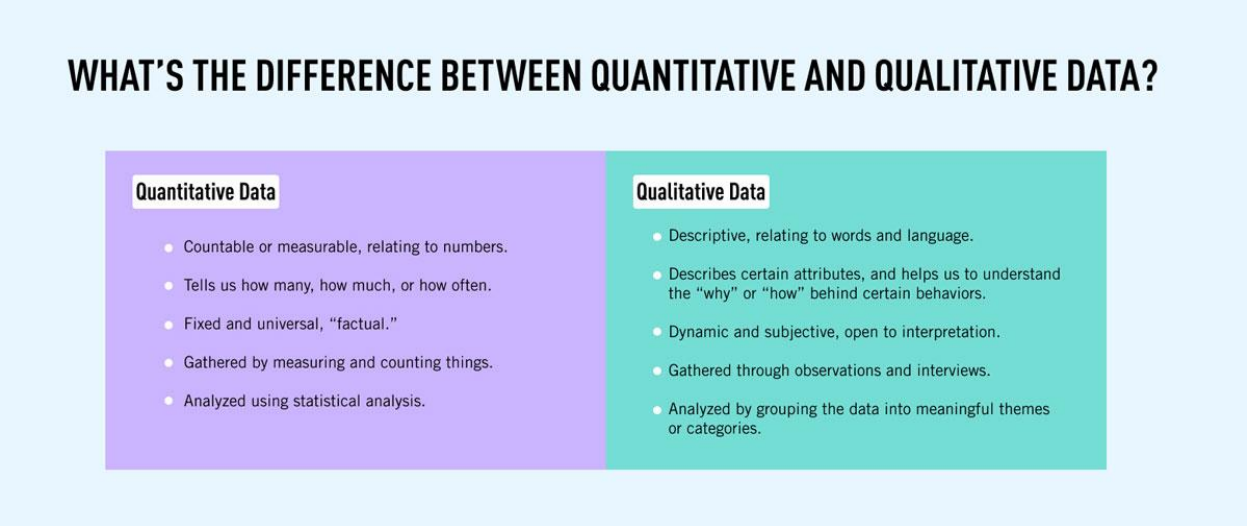
1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling
5. Accidental sampling

3. What is the difference between type1 and type II error?
Answer: A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.
4. What do you understand by the term Normal distribution?
Answer: In statistics, a normal distribution (also known as Gaussian, Gauss, or Laplace–Gauss distribution) is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The parameter μ is the mean or expectation of the distribution (and also its median and mode), while the parameter σ is its standard deviation. The variance of the distribution is σ^2. A random variable with a Gaussian distribution is said to be normally distributed, and is called a normal deviate.
5. What is correlation and covariance in statistics?
Answer: Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.
Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related.

6. Differentiate between univariate, Biavariate,and multivariate analysis.
Answer: Explorative data analysis has different analysis like:
Univariate statistics summarize only one variable at a time.
Bivariate statistics compare two variables.
Multivariate statistics compare more than two variables.
7. What do you understand by sensitivity and how would you calculate it?
Answer: Sensitivity is a measure of how well a machine learning model can detect positive instances. It is also known as the true positive rate (TPR) or recall. Sensitivity is used to evaluate model performance because it allows us to see how many positive instances the model was able to correctly identify. A model with high sensitivity will have few false negatives, which means that it is missing a few of the positive instances. In other words, sensitivity measures the ability of a model to correctly identify positive examples. This is important because we want our models to be able to find all of the positive instances in order to make accurate predictions. The sum of sensitivity (true positive rate) and false negative rate would be 1. The higher the true positive rate, the better the model is in identifying the positive cases in the correct manner.
Mathematically, sensitivity or true positive rate can be calculated as the following:
Sensitivity = (True Positive)/(True Positive + False Negative)
A high sensitivity means that the model is correctly identifying most of the positive results, while a low sensitivity means that the model is missing a lot of positive results.
The following are the details in relation to True Positive and False Negative used in the above equation.
True Positive: Persons predicted as suffering from the disease (or unhealthy) are actually suffering from the disease (unhealthy); In other words, the true positive represents the number of persons who are unhealthy and are predicted as unhealthy.
False Negative: Persons who are actually suffering from the disease (or unhealthy) are actually predicted to be not suffering from the disease (healthy). In other words, the false-negative represents the number of persons who are unhealthy and got predicted as healthy. Ideally, we would seek the model to have low false negatives as it might prove to be life-threatening or business threatening.
8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
Answer:
9. What is quantitative data and qualitative data?
Answer: The main differences between quantitative and qualitative data lie in what they tell us, how they are collected, and how they are analyzed. Let's summarize the key differences before exploring each aspect in more detail:
Quantitative data is countable or measurable, relating to numbers. Qualitative data is descriptive, relating to language.
Quantitative data is gathered by measuring and counting. Qualitative data is collected by interviewing and observing.
Quantitative data is analyzed using statistical analysis, while qualitative data is analyzed by grouping it in terms of meaningful categories or themes.



10. How to calculate range and interquartile range?

Answer: The Interquartile range formula helps in finding the difference between the third quartile and the first quartile. The Interquartile range formula measures the variability, based on dividing an ordered set of data into quartiles. Quartiles are three values or cuts that divide each respective part as the first, second, and third quartiles, denoted by Q1Q1, Q2Q2, and Q3Q3, respectively.

Q1Q1 is the cut in the first half of the rank-ordered data set

Q2Q2 is the median value of the set

Q3Q3 is the cut in the second half of the rank-ordered data set.

The Interquartile Range (IQR) formula is a measure of the middle 50% of a data set. The smallest of all the measures of dispersion in statistics is called the Interquartile Range. The difference between the upper and lower quartile is known as the interquartile range.

Interquartile range = Upper Quartile – Lower Quartile

Q2=Q3–Q1Q2=Q3–Q1

where,

IQR = Interquartile range (IQR = Q2Q2)

Q1Q1 = (1/4)[(n + 1)]th term)

Q3Q3 = (3/4)[(n + 1)]th term)

n = number of data points

The following steps help us to find the IQR:

The simple trick is to arrange the data points in ascending order.

Q2Q2 is the median of the data. If the number of data points is odd, the middle term is (n+1)/2 and if the number of data points is even, the median is the mean of the two middle points.

Q1Q1 is the median of the data points to the left of the median found in step 2.

Q3Q3 is the median of the data points to the right of the median found in step 2.

IQR = Q2=Q3–Q1

11. What do you understand by bell curve distribution?

Answer: A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell. The top of the curve shows the mean, mode, and median of the data collected. Its standard deviation depicts the bell curve's relative width around the mean.

12. Mention one method to find outliers.
Answer: Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your statistical analyses and skew the results of any hypothesis tests.
It's important to carefully identify potential outliers in your dataset and deal with them in an appropriate manner for accurate results.
There are four ways to identify outliers:
Sorting method
Data visualization method
Statistical tests (z scores)
Interquartile range method
13. What is p-value in hypothesis testing?
Answer: The p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P-values are used in hypothesis testing to help decide whether to reject the null hypothesis.
14. What is the Binomial Probability Formula?
Answer:

Formula

$$P_x = \binom{n}{x} p^x q^{n-x}$$

P = binomial probability
x = number of times for a specific outcome within n trials
$\binom{n}{x}$ = number of combinations
p = probability of success on a single trial
q = probability of failure on a single trial
n = number of trials

15. Explain ANOVA and its applications.
Answer: Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. An ANOVA test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance. The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples. Application of ANOVA in Quality and cost comparison, Product safety tests, Optimize production.