Assignment- 4

MACHINE LEARNING

Batch DS2311

Name: Alvin V Regi

1. **Which of the following methods do we use to find the best fit line for data in Linear Regression?**

Answer-

Least Square Error.

This method minimizes the sum of the squared differences between the actual and predicted values of the dependent variable. It is also known as the ordinary least squares (OLS) method or the method of least squares.

2. **Which of the following statement is true about outliers in linear regression?**

Answer-

Linear regression is sensitive to outliers.

This means that the presence of outliers can affect the slope and intercept of the regression line, and thus the accuracy of the predictions. Outliers are usually caused by errors in data collection, measurement, or entry, or by extreme values that do not represent the typical behavior of the data.

3. **A line falls from left to right if a slope is ______?**

Answer-

Negative.

A line falls from left to right if it has a negative slope, which means that the rise is negative, and the run is positive. A negative slope indicates that the line is decreasing when viewed from left to right.

4. **Which of the following will have symmetric relation between dependent variable and independent variable?**

Answer-

Correlation.

Correlation is a measure of the strength and direction of the linear relationship between two variables. It is symmetric, meaning that the correlation between x and y is the same as the correlation between y and x.

5. **Which of the following is the reason for over fitting condition?**

Answer-

Low bias and high variance.

This means that the model is too complex and fits the training data too well but fails to generalize to new data. Its also known as overfitting.

Overfitting is a common problem in machine learning, and it can be caused by various factors, such as:

- Using too many features or parameters in the model, which increases the model's flexibility and complexity.
- Using too few training examples, which makes the model learn the noise and outliers in the data, rather than the underlying pattern.

6. **If output involves label then that model is called as:**

Answer-

Predictive model.

A predictive model is a type of machine learning model that outputs a target variable based on the input variables. For example, a predictive model can classify an email as spam or not spam, or predict the price of a house based on its features.

7. **Lasso and Ridge regression techniques belong to _________?**

Answer-

Regularization.

Lasso and Ridge regression are two types of regularization techniques that are used to prevent overfitting and improve the performance of linear regression models. They both add a penalty term to the cost function of the model, which reduces the magnitude of the coefficients and makes the model simpler and more generalizable.

8. **To overcome with imbalance dataset which technique can be used?**

Answer-

SMOTE.

SMOTE stands for Synthetic Minority Over-sampling Technique, and it is a method of generating new samples from the minority class by using a k-nearest neighbors algorithm. SMOTE helps to balance the dataset and improve the performance of the classification models.

9. **The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?**

Answer-

TPR and FPR.

TPR, also known as sensitivity or recall, measures the proportion of actual positive instances that are correctly identified by the model. FPR, also known as the fall-out, measures the proportion of actual negative instances that are incorrectly identified by the model.

10. **In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.**

Answer-

False.

The AUC-ROC curve, or Area Under the Receiver Operating Characteristic curve, is a graphical representation of the performance of a binary classification model at various classification thresholds. It is commonly used in machine learning to assess the ability of a model to distinguish between two classes, typically the positive class (e.g. sick) and the negative class (e.g. healthy)

11. **Pick the feature extraction from below:**
    A) Construction bag of words from a email
    B) Apply PCA to project high dimensional data
    C) Removing stop words
    D) Forward selection

Answer-

Apply PCA to project high dimensional data.

PCA, or Principal Component Analysis, is a technique for reducing the dimensionality of a data set by finding a set of orthogonal axes that capture the most variance in the data. PCA can be used for feature extraction, as it transforms the original features into a smaller set of new features that are linear combinations of the original ones.

Feature extraction is the process of transforming the raw data into a more suitable and informative representation for machine learning tasks. Feature extraction can help to reduce the complexity, noise, and redundancy of the data, and improve the performance and interpretability of the models

12. **Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?**

Answer-

It becomes slow when number of features is very large.

The normal equation is a closed-form solution for finding the optimal coefficients of a linear regression model. It does not require choosing a learning rate or iterating, and it does use the dependent variable. However, it involves computing the inverse of a matrix that has the size of the number of features, which can be very costly and inefficient when the number of features is large.

13. **& 14 Explain the term regularization? Which particular algorithms are used for regularization?**

Answer-

Regularization is a technique used in machine learning to prevent overfitting and improve the generalization of the models. Overfitting occurs when the model fits the training data too well, but fails to perform well on new or unseen data. Regularization helps to reduce the complexity of the model by adding a penalty term to the cost function, which shrinks the coefficients of the features or parameters. This way, the model becomes simpler and less prone to learn the noise or outliers in the data.

There are different types of regularization techniques, such as:

- L1 regularization, also known as Lasso, which adds the absolute value of the coefficients to the cost function. This technique tends to produce sparse solutions, meaning that some of the coefficients become zero and are eliminated from the model.

- L2 regularization, also known as Ridge, which adds the square of the coefficients to the cost function. This technique tends to produce small but non-zero coefficients, meaning that all the features are kept in the model, but with reduced importance.

- Elastic net, which combines both L1 and L2 regularization, and allows to control the balance between them. This technique can be useful when there are many correlated features in the data.

Regularization is an important concept in machine learning, as it can help to improve the performance, interpretability, and robustness of the models.

15. **Explain the term error present in linear regression equation?**

Answer-

The error term in a linear regression equation is the difference between the observed value of the dependent variable and the predicted value of the dependent variable at a given value of the independent variable. It represents the random variation or noise in the data that is not explained by the linear relationship between the variables

The error term is also known as the residual, and it can be calculated by subtracting the predicted value from the observed value for each data point. The error term has some important properties, such as:

- The mean of the error term is zero.
- The variance of the error term is constant and equal to the residual variance.
- The error term is independent of the independent variable and the predicted value.
- The error term follows a normal distribution.

The error term is used to measure the goodness-of-fit of the linear regression model, and to perform statistical tests and confidence intervals for the model parameters

PYTHON

1. **Which of the following operators is used to calculate remainder in a division?**

Answer-

%.

It is also called the modulus or modulo operator.

2. **In python 2//3 is equal to?**

Answer-

0.

This operator // is used to perform integer division, which means it returns the quotient without the fractional part.

3. **In python, 6<<2 is equal to?**

Answer-

24.

This operator << is used to perform a bitwise left shift, which means it shifts the bits of the first operand to the left by the number of positions specified by the second operand.

4. **In python, 6&2 will give which of the following as output?**

Answer-

2

In Python, the operator & is used to perform a bitwise AND, which means it returns a bit-wise logical AND of the two operands.

5. **In python, 6|2 will give which of the following as output?**

Answer-

6

In Python, the operator | is used to perform a bitwise OR, which means it returns a bit-wise inclusive OR of the two operands.

6. **What does the finally keyword denotes in python?**

Answer-

The finally block will be executed no matter if the try block raises an error or not.

The finally keyword is used to define a block of code that will always run after the try and except blocks, regardless of whether an exception was raised or not. It is useful for releasing resources or performing cleanup tasks. For example, you can use the finally block to close a file that was opened in the try block.

7. **What does raise keyword is used for in python?**

Answer-

It is used to raise an exception.

The raise keyword is used to explicitly trigger an exception in Python. It can be used to create custom exceptions or re-raise existing ones. For example, you can use the raise keyword to raise a ValueError if the input is invalid.

8. **Which of the following is a common use case of yield keyword in python?**

Answer-

In defining a generator.

The yield keyword is used to return a value from a function without exiting the function. It can be used to create a generator, which is a special type of iterator that produces values on demand.

9. **Which of the following are the valid variable names?**

Answer-

_abc and abc2.

In Python, a variable name can start with an underscore or a letter, but not a digit. It can contain any combination of letters, digits, and underscores.

10. **Which of the following are the keywords in python?**

Answer-

Yield & raise.

Keywords are reserved words that have a special meaning and syntax in the language. They cannot be used as identifiers for variables, functions, classes, etc.

Statistics

1. **Bernoulli random variables take (only) the values 1 and 0.**

Answer-

True.

A Bernoulli random variable is a discrete random variable that has only two possible outcomes: success or failure, which are usually coded as 1 or 0.

2. **Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?**

Answer-

Central Limit Theorem.

The Central Limit Theorem is one of the most important results in probability and statistics. It states that if X1, X2,...,Xn are independent and identically distributed random variables with mean. and variance , then the sample mean is approximately normally distributed with mean and variance as n becomes large. Therefore, the standardized sample mean is approximately standard normally distributed with mean 0 and variance 1 as n becomes large. This theorem allows us to use the normal distribution to approximate the sampling distribution of many statistics, such as the mean, the proportion, the difference of means, etc.

3. **Which of the following is incorrect with respect to use of Poisson distribution?**

Answer-

Modeling bounded count data.

The Poisson distribution is used to model the number of events that occur in a fixed interval of time or space, assuming that the events are independent and occur at a constant rate. However, the Poisson distribution is unbounded, meaning that it can take any non-negative integer value. Therefore, it is not suitable for modeling count data that has a natural upper limit, such as the number of heads in a coin toss, or the number of students in a class.

4. **Point out the correct statement.**

a) The exponent of a normally distributed random variables follows what is called the log-normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

Answer-

The square of a standard normal random variable follows what is called chi-squared distribution.

This is a special case of the more general result that the sum of squares of independent standard normal random variables follows a chi-squared distribution with degrees of freedom equal to the number of variables.

5. **______ random variables are used to model rates.**

Answer-

Poisson.

Poisson random variables are used to model the number of events that occur in a fixed interval of time or space, assuming that the events are independent and occur at a constant rate.

6. **Usually replacing the standard error by its estimated value does change the CLT.**

Answer-

False.

The Central Limit Theorem (CLT) states that the standardized sample mean of independent and identically distributed (iid) random variables is approximately standard normally distributed as the sample size becomes large. The standard error is the standard deviation of the sampling distribution of the sample mean, which is usually unknown and estimated by the sample standard deviation. Replacing the standard error by its estimated value does not change the CLT, because the estimated standard error converges to the true standard error as the sample size becomes large.

7. **Which of the following testing is concerned with making decisions using data?**

Answer-

Hypothesis.

Hypothesis testing is a statistical method that is used to make decisions or draw conclusions using data. It involves testing a claim or a hypothesis about a population parameter, such as the mean, the proportion, the difference, etc., based on a sample of data.

8. **Normalized data are centered at______and have units equal to standard deviations of the original data.**

Answer-

0.

Normalized data are data that have been transformed to have a mean of 0 and a standard deviation of 1. This process is also called standardization or z-scoring. Normalized data are useful for comparing data that have different scales or units, or for applying statistical methods that assume normality. To normalize a data point, you can subtract the mean and divide by the standard deviation of the original data.

9. **Which of the following statement is incorrect with respect to outliers?**

Answer-

Outliers cannot conform to the regression relationship.

Outliers are data points that deviate significantly from the rest of the data, or from the expected pattern or trend. Outliers can sometimes conform to the regression relationship, meaning that they lie on or close to the fitted line or curve, but they are far away from the other data points. These outliers are called vertical outliers, and they do not have much influence on the slope or the intercept of the regression model.

10. **What do you understand by the term Normal Distribution?**

Answer-

The normal distribution is a type of probability distribution that describes how likely it is to get different values of a continuous random variable. It is also called the Gaussian distribution or the bell curve, because it has a symmetrical shape that looks like a bell. The normal distribution is characterized by two parameters: the mean and the standard deviation. The mean is the average value of the random variable, and the standard deviation is a measure of how much the values vary around the mean. The normal distribution has some important properties, such as:

- About 68% of the values are within one standard deviation of the mean.
- About 95% of the values are within two standard deviations of the mean.
- About 99.7% of the values are within three standard deviations of the mean.

The normal distribution is useful for modeling many natural phenomena, such as heights, weights, IQ scores, blood pressure, etc. It is also widely used in statistics and inference because many statistical methods and tests are based on the assumption of normality.

11. **How do you handle missing data? What imputation techniques do you recommend?**

Answer-

Missing data is a common problem in data analysis, and it can affect the quality and validity of the results. There are different ways to handle missing data, depending on the type, pattern, and mechanism of the missingness. Some of the common methods are:

- Deleting the missing data: This method involves removing the rows or columns that contain missing values from the data set. This is the simplest and fastest way to deal with missing data, but it can reduce the sample size and introduce bias if the missing data is not random.
- Imputing the missing data: This method involves replacing the missing values with some estimated values based on the available data. There are different techniques for imputation, such as mean, median, mode, regression, interpolation, k-nearest neighbours, etc. This method can preserve the sample size and reduce bias, but it can also introduce uncertainty and error in the imputed values.
- Ignoring the missing data: This method involves using statistical methods or models that can handle missing data without deleting or imputing them. For example, some methods can use the likelihood function or the expectation-maximization algorithm to estimate the parameters and the missing values simultaneously. This method can avoid the drawbacks of deleting or imputing, but it can also be complex and computationally intensive.

12. **What is A/B testing?**

Answer-

A/B testing is a method of comparing two versions of something to see which one performs better. For example, you can use A/B testing to compare two versions of a website, an app, an email, an ad, etc., and measure how they affect the behaviour of the users, such as clicks, conversions, purchases, etc. A/B testing can help you make data-driven decisions and improve your design, marketing, or product.

13. **Is mean imputation of missing data acceptable practice?**

Answer-

Mean imputation is a technique that replaces missing values with the mean of the non-missing values of the same variable. It is a simple and fast way to deal with missing data, but it has many drawbacks that can affect the quality and validity of the analysis. Some of the drawbacks are:

- It reduces the variability and standard deviation of the imputed variable, making the data less representative of the true population.
- It introduces bias in the estimates of the correlations and regression coefficients involving the imputed variable, making them closer to zero than they should be.
- It does not account for the uncertainty and error associated with the imputed values, making the confidence intervals and hypothesis tests too narrow or too optimistic.
- It does not handle the cases where the missing data is not missing completely at random (MCAR), meaning that the probability of missingness depends on the observed or unobserved data.

Therefore, mean imputation is not a recommended practice for handling missing data, unless the proportion of missing data is very small, the variable is not of interest for the analysis, and the missing data mechanism is MCAR. Otherwise, more advanced, and robust methods, such as multiple imputation, should be used instead.

14. **What is linear regression in statistics?**

Answer-

Linear regression is a statistical method that estimates the relationship between one or more independent variables and a dependent variable. It can be used to test hypotheses, make predictions, or explore trends based on data. There are different types of linear regression, such as simple, multiple, and logistic, depending on the number and nature of the variables involved.

15. **What are the various branches of statistics?**

Answer-

Statistics is the science of collecting, organizing, analyzing, and interpreting data to make decisions or draw conclusions. There are two main branches of statistics: descriptive and inferential. Descriptive statistics summarizes and displays the properties of a data set, such as the mean, median, mode, range, standard deviation, frequency, distribution, etc. Inferential statistics uses the data to test hypotheses, make predictions, or estimate parameters, such as the confidence interval, p-value, correlation, regression, ANOVA, etc.