MACHINE LEARNING

1. Which of the following methods do we use to find the best fit line for data in Linear Regression? A) Least Square Error B) Maximum Likelihood C) Logarithmic Loss D) Both A and B

Answer - A) Least Square Error

2. Which of the following statement is true about outliers in linear regression? A) Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers C) Can't say D) none of these

Answer - A) Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is _____? A) Positive B) Negative C) Zero D) Undefined

Answer - B) Negative

4. Which of the following will have symmetric relation between dependent variable and independent variable? A) Regression B) Correlation C) Both of them D) None of these

Answer - B) Correlation

5. Which of the following is the reason for over fitting condition? A) High bias and high variance B) Low bias and low variance C) Low bias and high variance D) none of these

Answer- C) Low bias and high variance

6. If output involves label then that model is called as: A) Descriptive model B) Predictive modal C) Reinforcement learning D) All of the above

Answer - B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____? A) Cross validation B) Removing outliers C) SMOTE D) Regularization

Answer - D) Regularization

8. To overcome with imbalance dataset which technique can be used? A) Cross validation B) Regularization C) Kernel D) SMOTE

Answer - D) SMOTE

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph? A) TPR and FPR B) Sensitivity and precision C) Sensitivity and Specificity D) Recall and precision

Answer- A) TPR and FPR

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

Answer - B) False

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 - A) Construction bag of words from a email

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression? A) We don't have to choose the learning rate. B) It becomes slow when number of features is very large. C) We need to iterate. D) It does not make use of dependent variable.

Answer - A) We don't have to choose the learning rate.

B) It becomes slow when number of features is very large.

13. Explain the term regularization?

Answer - In machine learning and statistics, regularisation is a technique used to improve a model's generalisation and prevent overfitting. A model is said to be over fitted if it learns noise and random fluctuations in addition to the underlying pattern from the training set. This can result in the model performing poorly on fresh, untested data.
The primary goal of regularisation is to deter the model from fitting the training data too closely by including a penalty term in the objective function. The complexity of the model, which is frequently determined by the size of the coefficients or parameters, usually determines this penalty.

Regularisation approaches can be broadly classified into two types:

1. L2 Regularisation, also known as Ridge Regularisation:
o The penalty in L2 regularisation, commonly referred to as Ridge regularisation, is squared by the magnitudes of the model coefficients. The objective function of the form λ∑j=1pθj2\lambda \sum_{j=1}^{p} \theta_j^2λ∑j=1pθj2 is enhanced by a term, where the model coefficients are denoted by θj\theta_jθj, and the regularisation parameter λ\lambdaλ regulates the penalty strength. By using this method, the coefficients are encouraged to be small, which effectively shrinks them towards zero.

.

2 . L1 Regularisation (Lasso Regularisation): o As a first step, a penalty term is added that is based on the absolute values of the model coefficients: Assuming \sum_{j=1}^{p} |\theta_j|, λ∑j=1p|θj∗\lambda \sum_{j=1}^{p}. The regularisation parameter in this case is λ\lambdaλ, just like in L2 regularisation; nevertheless, it tends to produce sparse coefficients by driving some coefficients exactly to zero. When the model contains a large number of irrelevant features, L1 regularisation becomes helpful due to this feature selection aspect.

The particulars of the dataset and the situation at hand determine whether to use L2 or L1 regularisation, or a combination known as Elastic Net regularisation. By exchanging decreased variance (better generalisation to unknown data) for increased bias (caused by the penalty on the coefficients), regularisation aids in the control of overfitting.
In conclusion, regularisation is an essential machine learning technique that, by penalising overly complex models, helps to produce models that perform better and generalise effectively to fresh data.

14. Which particular algorithms are used for regularization?

Answer - Regularisation methods are frequently used to enhance model generalisation and avoid overfitting in a range of machine learning algorithms. The particular algorithms that frequently use regularisation are as follows:

1. Linear Regression: o Ridge Regression (L2 Regularisation): Encourages smaller coefficient values by adding a penalty term to the least squares objective.
   o Lasso Regression (L1 Regularisation): This technique is comparable to Ridge Regression but makes use of the coefficients' absolute values to encourage feature selection and sparsity.

   2. Logistic Regression: o Ridge (L2) or Lasso (L1) regularisation can be used to penalise high coefficients in logistic regression models in a manner akin to that of linear regression models.

   3. Support Vector Machines (SVM): o By modifying the C parameter, which regulates the trade-off between maximising the margin and minimising the classification error, SVMs can be made more regular. Regularisation decreases with a higher C value and increases with a lower C value.

   4. Neural Networks: o Regularisation of neural networks can be achieved using methods like: L2 Regularisation: This involves appending a penalty term to the loss function, which is determined by the squared magnitudes of the weights.

      Dropout: During training, neurons are randomly removed to avoid co-adaptation and minimise overfitting.
      Early Stopping: To avoid the model overfitting the training data, training is stopped when performance on a validation set begins to deteriorate.

5. Discriminant analysis techniques, linear (LDA) and quadratic (QDA):
   a. By modifying the parameters that govern covariance estimation, regularisation can be applied to LDA and QDA models to enhance performance with sparse data.

   These algorithms and techniques show how different machine learning approaches can incorporate regularisation techniques to improve model performance and generalisation capabilities. The particulars of the data and the learning algorithm being utilised determine which regularisation strategy is best.

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

Answer - The difference between the values predicted by the linear regression model and the actual observed values of the dependent variable (or target variable) is referred to as "error" in the context of linear regression. Another name for these faults is residuals.

This is a thorough explanation:

1. Observed Values (Actual Values): The dependent variable yyy's genuine values as observed or measured in the dataset are these. A standard linear regression setup involves a dataset made up of independent variable pairs (XXX) and dependent variable pairs (yyy), where yyy is the response variable or actual outcome that you are attempting to predict.

2. Predicted Values: Based on the input values XXX, the linear regression model predicts the values of yyy. By combining the input variables and the corresponding coefficients in a linear fashion, the linear regression model calculates these anticipated values
.

3. Error (Residuals): In linear regression, the error is the difference between the predicted values y^\hat{y}y^ and the observed values yyy. (where the expected value of yyy is shown by y^\hat{y}y^). In terms of math, for every data point iii:

4. Regression's Objective: The line (or hyperplane, in higher dimensions) that minimises these errors across all data points is the target of linear regression. Usually, to do this, the sum of squared errors (SSE), or the total of the squared disparities between the actual and predicted values, is minimised:

5.Interpreting mistakes: If the model is not correctly predicting the dependent variable from the independent variables, it is likely showing large mistakes (residuals). A good fit between the model and the observed data is suggested by small errors. A further indicator of whether the linear regression model is suitable for the data is the distribution of errors, such as whether they are regularly distributed around zero.

STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0. a) True b) False

Answer - A) True

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? a) Central Limit Theorem b) Central Mean Theorem c) Centroid Limit Theorem d) All of the mentioned

Answer - a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution? a) Modeling event/time data b) Modeling bounded count data c) Modeling contingency tables d) All of the mentioned

Answer - b) Modeling bounded count data

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

5. _____ random variables are used to model rates. a) Empirical b) Binomial c) Poisson d) All of the mentioned

Answer - c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT. a) True b) False

Answer - b) False

7. 1. Which of the following testing is concerned with making decisions using data? a) Probability b) Hypothesis c) Causal d) None of the mentioned

Answer - b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data. a) 0 b) 5 c) 1 d) 10

Answer - a) 0

9. . Which of the following statement is incorrect with respect to outliers? a) Outliers can have varying degrees of influence b) Outliers can be the result of spurious or real processes c) Outliers cannot conform to the regression relationship d) None of the mentioned

Answer - c) Outliers cannot conform to the regression relationship

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

Answer - A key idea in probability theory and statistics is the "Normal Distribution," commonly referred to as the Gaussian distribution. It depicts a probability distribution that is bell-shaped and symmetric, with two characteristic variables: variance (σ^2) and mean (μ). The following are the main attributes and traits of the normal distribution:

Shape: Around its mean, the Normal Distribution is symmetrical and bell-shaped. It features a single peak at the mean, which indicates that it is unimodal.

1. Central Tendency: A Normal Distribution's centre is determined by its mean (μ). It is an illustration of the distribution's average value.

2. Dispersion: The variance (σ^2) or standard deviation (σ) define the distribution's spread. Greater data variability is indicated by a bigger standard deviation.

3. Probability Density Function: The Normal Distribution's probability density function (pdf) is provided by: f(x)=12πσ2exp⁡(−(x−μ)22σ2)f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)f(x)=2πσ21exp(−2σ2(x−μ)2), where xxx is a random variable, μ\muμ is the mean, and σ2\sigma^2σ2 is the variance.

4.. Standard Normal Distribution: The Standard Normal Distribution is a special case of the Normal Distribution with a mean (μ) of 0 and a variance (σ^2) of 1. Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ is the formula for standardising random variables from a normal distribution to this distribution: ZZZ is a standard normal random variable, and XXX is a normal random variable with mean μ\muμ and standard deviation σ\sigmaσ.

5. Empirical Rule: The Normal Distribution complies with the 68-95-99.7 rule, which is an

empirical rule.
o The data is distributed within one standard deviation of the mean in about 68% of cases.

A little over 95% of data are contained within two standard deviations.
o 99.7% of the data are contained within three standard deviations.

6 . Applications: The Normal Distribution finds extensive use in a multi
tude of sectors, including economics, engineering, statistics, social sciences, and natural sciences. In addition to being crucial for inferential statistics, hypothesis testing, and confidence interval construction, it functions as a basic model for a variety of natural phenomena.

In conclusion, the Normal Distribution is a fundamental idea in statistics that is widely applied because of its frequent occurrence in natural processes and bell-shaped curve, which is defined by mean and variance. It is also mathematically tractable.

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

Answer - One of the most important steps in the data preparation process is addressing missing data before utilising machine learning algorithms or statistical analysis. Some common techniques for handling missing data and recommended imputation algorithms are as follows:

1. Locate Missing Data: To begin with, locate the missing data in your dataset and comprehend its pattern and mechanism. Ascertain whether the missing values are related to the missing variable itself (missing not at random, MNAR), to certain observable variables (missing at random, MAR), or to random (missing completely at random, MCAR).

2. Understanding Data: Examine the extent to which your analysis is affected by missing data. Take into account each variable's percentage of missing values and how they can affect the model's performance.

3.Managing Missing Data: • Removal: Eliminate any rows or columns that contain missing data. Although this method is simple, if the data are not MCAR, it could result in the loss of important information.
o Suggested when: There are few missing values and the data are MAR or MCAR.

• Imputation: Approximate values based on other available data are used to fill in missing values.
o Suggested methods: Mean/Median imputation: Use the mean or median of the observed values for that variable to fill in any missing values. Although this approach is easy to use and efficient, links between variables might not be maintained.

Hot-deck imputation involves substituting absent values with equivalent records' values according to specific standards (such nearest neighbour).
Regression imputation involves utilising a regression model to predict missing values by

considering other factors.
To capture uncertainty, provide multiple plausible values for each missing value by multiple imputation. This approach works well for intricate missing data patterns.

o Suggested when: A significant portion of the data is missing, or the data are MAR.

4.Advanced Methods (if relevant):
• Model-based imputation: Predict missing values based on other variables by using machine learning methods (e.g., k-nearest neighbours, decision trees).
• Deep learning techniques: Neural networks and deep learning models function particularly well for complicated patterns when used to impute missing data.

5. Assess and Validate: • Determine how imputation affects your analysis. To evaluate the validity and robustness of imputation techniques, take into consideration comparing outcomes with and without imputation.

6.Documentation: • Keep a record of the imputation technique used and any assumptions you made on missing data. For outcomes to be interpretable and reproducible, transparency is essential.

In brief, there are various techniques for imputation such as mean/median, mode, hot-deck, regression, and multiple imputation.
•Take into account: Select imputation methods according to the objectives of your research, the missing data mechanism, and the properties of the data. Analyse how imputation affects the validity and performance of the model.

12. What is A/B testing?

Answer - A controlled experiment called A/B testing, sometimes referred to as split testing, compares two or more versions of a website, email, advertisement, or other digital material. Finding the variant that performs better in reaching a particular goal, such raising click-through rates, conversions, or user engagement, is its main objective.

Essential Ideas in A/B Testing:

1. Variants: In A/B testing, a web page, email, or other piece of content is created in numerous versions (variants), each of which differs in one specific area (e.g., colour, layout, phrasing, call-to-action).

2. Randomization: Different versions are assigned to users or visitors at random. By using randomization, it is made more likely that variations in performance measures will only be caused by the effect of the variant and not by other outside influences.

3. Objective: Prior to carrying out the test, a precise objective or hypothesis is developed. This could be increasing sales, decreasing bounce rates, or enhancing conversion rates, among other things.

4. measures: To gauge each variant's degree of success, key performance measures (KPIs)

are established. Common measures include time on page, income produced, click-through rates (CTR), conversion rates, and so forth.

5.Statistical Analysis: To ascertain whether there is a statistically significant difference between the variants, statistical methods are employed to analyse the data. This makes it less likely that discrepancies you see are the result of random variation.

The A/B testing procedure:

1.Identify Your Hypothesis: Come up with a theory regarding how a modification to a particular element (such as the colour of the button) will affect user behaviour or performance metrics.
2. Produce Variants: Provide several iterations of the material, each with a solitary variation. Variant B, for instance, might have a green button and Variant A, a red button.

3. Randomise and Implement: Assign users to various variations at random. Apply the modifications to your email campaign, website, or other platforms.
4. Gather Data: Over a predetermined time frame, track and gather data on each variant's performance indicators (such as conversion rates and CTR).
4. analytic: Compare each variant's performance using statistical analytic tools. Find out if there is a difference between them that is statistically significant.
6. Conclusion: Determine which variant outperformed the others based on the analysis. Use the victorious option to accomplish the intended goal.

Advantages of A/B Testing
• Data-Driven Decisions: Empirical data from A/B testing is used to guide choices regarding marketing tactics, design, and content.
• Optimisation: By recognising successful design or content adjustments, it helps optimise conversion rates and user experience.

• Cost-Effectiveness: A/B testing lowers the risk of introducing significant changes that might not be well-received by users by testing minor adjustments.
• Culture of Continuous Improvement: Through iterative testing and optimisation of digital assets, A/B testing promotes a culture of continuous improvement.

In conclusion, A/B testing is a potent technique that is used in web development and digital marketing to analyse data and conduct controlled experiments to objectively assess and enhance the efficacy of design, content, and user experience.

13. Is mean imputation of missing data acceptable practice?

Answer - Mean imputation, where missing values are replaced with the mean of the observed values for that variable, is a straightforward and commonly used technique for handling missing data. However, its acceptability and appropriateness depend on several factors and should be considered carefully:

Benefits of Mean Imputation

1. Simple and Easy: Mean imputation provides a rapid fix for missing data since it is simple to use and comprehend.
2. Preserves Sample Size: By keeping all of the dataset's observations, it can help to preserve statistical power and lessen bias in later studies.

3. Helpful for Linear Models: Mean imputation is particularly useful in linear regression models when data are missing at random (MAR) or entirely missing at random (MCAR).

Drawbacks and Things to Think About:

1. Distorts Data Distribution: If missing data are not MCAR, mean imputation may artificially reduce variance and correlations between variables, which may distort the variable's distribution. Biassed parameter estimations and underestimating of standard errors may result from this.
2. Unsuitable for Categorical Data: Mean imputation is not suited for variables that have non-normal distributions or for categorical data. It might add irrational values to the data that weren't there in the beginning.

3. Does Not Take Into Account Uncertainty: This could lead to an underestimation of variability and compromise the reliability of statistical conclusions since it does not take into consideration uncertainty in the imputed values.
4. Effect on Relationships: In more intricate studies or when significant interactions exist between variables, mean imputation may have an impact on the relationships between the variables.

Alternatives to Mean Imputation:

• Median Imputation: This technique uses the median, which is less susceptible to outliers and distributional skewness, to fill in missing values.
• Multiple Imputation: Using the observed data distribution as a guide, produce several believable values for each missing value. With this method, uncertainty is captured and more precise estimations are produced.

• Model-Based Imputation: By utilising correlations between variables, forecast missing values using statistical models like regression, nearest neighbours, or Bayesian techniques.

14. What is linear regression in statistics?

Answer - A basic and popular method in statistics for simulating the connection between a dependent variable (yyy) and one or more independent variables (XXX) is linear regression. It makes the assumption that the dependent variable and the predictors have a linear relationship.

.

Essential Ideas in Linear Regression:

1. Formulation of the Model: The following linear relationship is assumed by linear regression: y = β0+β1x1+β2x2+…+βpxp+ϵy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \The dependent variable (response variable) that we wish to forecast is epsilony=β0+β1x1+β2x2+…+βpxp+ϵo~yyy.

o~x1,x2,…,xpx_1, x_2, \ldots, x_px1,x2,…,xp: Independent variables (predictors) that account for differences in yyy.
The coefficients, also known as parameters, o~β0,β1,β2,…,βp\beta_0, \beta_1, \beta_2, \ldots, \beta_pβ0,β1,β2,…,βp: These indicate the slope of the connection between each predictor and the dependent variable.
oϵ\epsilonϵ: Error term in the model that denotes noise or unexplained variance.

2 Using linear regression, the objective is to find the coefficients β0,β1,…,βp\beta_0, \beta_1, \ldots, \beta_pβ0,β1,…,βp that minimise the difference between the observed values yyy and the anticipated values y^\hat{y}y^ (obtained from the model).
2. Presumptions
o Linearity: There is a linear relationship between the predictors xix_ixi and yyy.

o Independence: The disparities between the observed and predicted values, or residuals, are unrelated to one another.
o Homoscedasticity: At all predictor levels, the residuals' variance remains constant.
o Normality: A normal distribution is followed by the residuals.
.

3. Types of Linear Regression: o One independent variable, xxx, is used in Simple Linear Regression.
ldots, x_px1,x2,…,xp, and other independent variables are included in multiple linear regression.

o Polynomial Regression: Polynomial terms of predictors (such as x2,x3x^2, x^3x2,x3) are added to linear regression to extend its capabilities.
o Generalised Linear Models (GLM): GLMs extend linear regression to address error distributions other than normal and non-normal response variables.

4. Model Assessment
o The coefficient of determination (R2R^2R2) quantifies the percentage of the dependent variable's variance that can be predicted based on the independent variables.
o Standard Error of the Estimate: Shows how far apart from expected values the observed values are on average.

Uses for Nonlinear Regression

• Prediction: Using predictor variables, forecast continuous outcomes (like sales or temperature).
Analysing the direction and intensity of correlations between variables is known as

relationship analysis.

Understanding how changes in predictor variables impact the dependent variable is known as causal inference.

In conclusion, the core statistical method of linear regression is used to describe the relationship between variables in a variety of domains, including business, engineering, social sciences, and economics. When used appropriately and with the premise that it is met, its simplicity and interpretability make it a useful tool for both inference and prediction tasks.

15. What are the various branches of statistics ?

Answer - Each branch of the large discipline of statistics focuses on a distinct facet of data analysis, inference, and application. The following are a few of the major areas of statistics:

1. Methods for condensing and characterising data sets are included in descriptive statistics. measurements of dispersion (variance, standard deviation), measurements of central tendency (mean, median, mode), and graphical depiction (histograms, box plots, etc.) are all included in this branch.

2. Statistics that Infer:

o Based on sample data, inferential statistics entails drawing conclusions or forecasts about the population. Regression analysis, confidence intervals, and hypothesis testing are all included.

3. Theory of Probabilities: oThe basis of statistics is probability theory, which addresses the possibility of events in ambiguous circumstances. It covers ideas like random variables, stochastic processes, and probability distributions (such the binomial and normal distributions).

4.Biostatistics: Biostatistics is the study of applying statistical techniques to data pertaining to biology, medicine, and health. It is essential to genetics, public health, epidemiology, and clinical trials.

5. Measurement of economic activity

o Ecometrics analyses economic relationships, tests hypotheses, and projects future trends by using statistical techniques to economic data. It blends statistics, mathematics, and economics.

6. Psychometrics: The area of statistics that deals with measuring psychological characteristics and aptitudes is known as psychometrics. It entails creating and approving psychological scales and exams.

7. Social Statistics: o Sociology, political science, anthropology, and demography

research are all subjected to statistical methodologies in social statistics. It looks at trends in human behaviour, public opinion, and social phenomena.

8.Statistical Computing: o The creation and use of computational techniques for statistical analysis is the main emphasis of statistical computing. It comprises software tools for data processing, modelling, and visualisation as well as programming languages (like R and Python).

9. Quality Management and Trustworthiness:

o In manufacturing and industrial operations, quality control and reliability statistics are used to guarantee consistency, track and enhance product quality, and forecast failure rates.

10. Environmental Statistics: o Environmental statistics analyses and studies environmental data using statistical techniques. Natural resource management, environmental impact assessments, and studies on climate change are all included.

11. Spatial Statistics: o Addressing patterns, connections, and processes that differ across space, spatial statistics deals with the examination of spatial or geographical data. It consists of interpolation, spatial

The interdisciplinary character of statistical methods and their many applications across sectors and fields of study are reflected in these branches of statistics. Every discipline offers distinct methods and resources to tackle certain problems with data interpretation, analysis, and decision-making.

.

PYTHON – WORKSHEET 1

1. Which of the following operators is used to calculate remainder in a division? A) # B) & C) % D) $

Answer - C) %

2. In python 2//3 is equal to? A) 0.666 B) 0 C) 1 D) 0.67

Answer - B) 0

3. In python, 6<<2 is equal to? A) 36 B) 10 C) 24 D) 45

Answer - C) 24

4. In python, 6&2 will give which of the following as output? A) 2 B) True C) False D) 0

Answer- A) 2

1. In python, 6|2 will give which of the following as output? A) 2 B) 4 C) 0 D) 6

Answer – D) 6

2. What does the finally keyword denotes in python? A) It is used to mark the end of the code B) It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block. C) the finally block will be executed no matter if the try block raises an error or not. D) None of the above

Answer - C) the finally block will be executed no matter if the try block raises an error or not

3. What does raise keyword is used for in python? A) It is used to raise an exception. B) It is used to define lambda function C) it's not a keyword in python. D) None of the above

Answer - A) It is used to raise an exception

4. Which of the following is a common use case of yield keyword in python? A) in defining an iterator B) while defining a lambda function C) in defining a generator D) in for loop

Answer - C) in defining a generator

5. Which of the following are the valid variable names? A) _abc B) 1abc C) abc2 D) None of the above

Answer - A) _abc

C) abc2

6. Which of the following are the keywords in python? A) yield B) raise C) look-in D) all of the above

Answer - A) yield

B) raise