MACHINE LEARNING

Q1 to Q11

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

D) Both A and B

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

A) Linear regression is sensitive to outliers

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

B) Negative

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

B) Correlation

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

C) Low bias and high variance

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

B) Predictive modal

7. Lasso and Ridge regression techniques belong to _________?

D) Regularization

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

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

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

B) False

11. Pick the feature extraction from below:

A) Construction bag of words from a email

Q12

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.

D) It does not make use of dependent variable

Q13 and Q15

13. Explain the term regularization?

Regularization is a technique used in machine learning to prevent overfitting and improve the generalization performance of a model.

In the process of building a model, the goal is to fit the training data as closely as possible. However, sometimes the model may be too complex and try to fit the noise in the data, which can lead to overfitting. Overfitting occurs when the model captures the noise in the training data instead of the underlying pattern, causing the model to perform poorly on new, unseen data.

Regularization helps to address this problem by adding a penalty term to the model's loss function. The penalty term discourages the model from assigning too much weight to any single feature or parameter, which can help to reduce the complexity of the model and prevent overfitting.

There are several types of regularization techniques, including L1 regularization (also known as Lasso), L2 regularization (also known as Ridge), and dropout regularization. L1 regularization adds a penalty proportional to the absolute value of the parameters, while L2 regularization adds a penalty proportional to the square of the parameters. Dropout regularization randomly drops out a percentage of neurons in a neural network during training to prevent over-reliance on any single neuron.

Overall, regularization is an important tool for building machine learning models that can generalize well to new data and perform reliably in real-world applications.

14. Which particular algorithms are used for regularization?

There are several algorithms used for regularization, including:

1. L1 regularization (also known as Lasso regularization): This algorithm adds a penalty term to the cost function of a linear model, which encourages the model to select only the most important features and set the weights of less important features to zero.
2. L2 regularization (also known as Ridge regularization): This algorithm adds a penalty term to the cost function of a linear model, which encourages the model to keep all the features, but to keep their weights as small as possible.
3. Dropout regularization: This algorithm randomly drops out some neurons during training, which helps prevent overfitting and encourages the network to learn more robust representations.
4. Early stopping: This algorithm stops training when the performance on a validation set stops improving, which helps prevent overfitting.
5. Data augmentation: This algorithm artificially increases the size of the training set by applying transformations to the existing data, which helps prevent overfitting and improves generalization.
6. Batch normalization: This algorithm normalizes the inputs to each layer, which helps prevent overfitting and improves generalization.

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

In linear regression, the "error" or "residual" refers to the difference between the predicted value and the actual value of the dependent variable.

The linear regression equation can be written as:

y = mx + b + e

where y is the dependent variable (or response variable), x is the independent variable (or predictor variable), m is the slope of the regression line, b is the intercept, and e is the error term.

The error term (e) represents the variability or randomness in the dependent variable that is not explained by the independent variable. In other words, it is the difference between the actual value of the dependent variable and the value predicted by the regression line.

The goal of linear regression is to minimize the sum of the squared errors, which is known as the residual sum of squares (RSS). This is done by finding the values of m and b that minimize the distance between the predicted values and the actual values of the dependent variable. The regression line that results from this process is the "best-fit" line that represents the relationship between the independent and dependent variables.

PYTHON

Q1 to Q8

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

C) %

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

B) 0

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

C) 24

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

A) 2

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

D) 6

6. What does the finally keyword denotes in python?

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

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

A) It is used to raise an exception.

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

C) in defining a generator

Q9 and Q10

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

A) _abc and C) abc2

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

A) yield and B) raise

STATISTICS

Q1 to Q9

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

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

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

c) Modeling contingency tables

4. Point out the correct statement.

d) All of the mentioned

5. _______ random variables are used to model rates.

c) Poisson

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

b) False

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

b) Hypothesis

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

a) 0

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

c) Outliers cannot conform to the regression relationship

Q10and Q15

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

Normal distribution, also known as Gaussian distribution, is a continuous probability distribution that is commonly used to model real-world variables that tend to cluster around a mean value, with the majority of the observations falling close to the mean, and fewer observations further away from the mean. In a normal distribution, the probability density function is symmetric and bell-shaped, with the mean, median, and mode all equal to each other. The spread of the distribution is controlled by the standard deviation, which determines the width of the curve. Many natural phenomena, such as heights, weights, and IQ scores, follow a normal distribution, and it is widely used in statistics, hypothesis testing, and inferential analysis.

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

Handling missing data is an important task in data analysis as missing data can negatively affect the accuracy and effectiveness of the analysis. There are various techniques that can be used to handle missing data, some of which are:

1. Listwise Deletion: In this method, the entire observation with missing data is removed from the dataset. This method is easy to apply but can result in a loss of information.
2. Pairwise Deletion: In this method, only the missing values are removed from the analysis, and the rest of the data is used. This method is less biased than the listwise deletion but can result in an increased variability.
3. Mean/Median/Mode Imputation: In this method, the missing values are replaced by the mean, median or mode value of the non-missing values in the same column. This method is simple to apply but can result in biased estimates.
4. Regression Imputation: In this method, a regression model is used to predict the missing values based on the non-missing values in the same column. This method can provide more accurate estimates than mean/median/mode imputation.
5. Multiple Imputation: In this method, missing values are imputed multiple times using regression imputation, and the results are combined to provide a more accurate estimate.

The choice of imputation technique depends on the nature and extent of the missing data, and the specific requirements of the analysis. No single imputation technique is universally suitable for all situations. However, multiple imputation is generally recommended as it provides the most accurate estimates and accounts for the uncertainty in the imputed values.

12. What is A/B testing?

A/B testing is a method of comparing two versions of a product or service to determine which one performs better. It involves dividing a group of users into two subgroups, where one subgroup is shown version A and the other subgroup is shown version B. The performance of each version is then measured based on a predetermined metric, such as click-through rates, conversion rates, or revenue.

A/B testing can be used for a variety of purposes, such as testing new features, improving user experience, or optimizing marketing campaigns. It is commonly used in web design, software development, and digital marketing. The results of A/B testing can help businesses make data-driven decisions and improve the effectiveness of their products or services.

13. Is mean imputation of missing data acceptable practice?

Mean imputation of missing data is a commonly used technique for handling missing data. However, it has some drawbacks and limitations. One major limitation of mean imputation is that it assumes that the missing values are missing completely at random (MCAR), which means that the probability of a value being missing is unrelated to its actual value. In many cases, this assumption may not hold, and missing values may be related to other variables in the dataset.

Moreover, mean imputation can lead to biased estimates of the mean and standard deviation of the variable. This is because mean imputation reduces the variance of the variable by artificially inflating the sample size. As a result, the standard errors of the estimates are underestimated, and the significance levels of the statistical tests are overestimated.

In summary, mean imputation can be a quick and easy solution for handling missing data, but it should be used with caution and only when the assumption of MCAR is justifiable. In cases where the assumption is not met, more sophisticated imputation techniques, such as multiple imputation or maximum likelihood estimation, should be used.

14. What is linear regression in statistics?

Linear regression is a statistical method used to model the linear relationship between a dependent variable and one or more independent variables. It involves finding the best-fit line that represents the relationship between the variables. The goal of linear regression is to estimate the parameters of the linear equation, such as the slope and intercept, which can then be used to predict the value of the dependent variable based on the values of the independent variables. Linear regression is widely used in various fields, including economics, social sciences, engineering, and business, to analyze and model the relationships between variables.

15. What are the various branches of statistics?

Statistics is a broad field that encompasses a wide range of branches and sub-disciplines. Some of the major branches of statistics include:

1. Descriptive statistics: This branch deals with summarizing and describing data using measures such as mean, median, mode, standard deviation, etc.
2. Inferential statistics: This branch deals with making inferences and predictions about populations based on sample data.
3. Probability theory: This branch deals with the study of random events and the likelihood of their occurrence.

4. Biostatistics: This branch applies statistical methods to analyze and interpret data in the field of biology and medicine.
5. Econometrics: This branch applies statistical methods to analyze economic data and test economic theories.
6. Psychometrics: This branch applies statistical methods to measure and analyze human behavior and psychology.
7. Time series analysis: This branch deals with analyzing and forecasting time-dependent data.
8. Spatial statistics: This branch deals with analyzing and modeling spatial data, such as geographical data.
9. Bayesian statistics: This branch deals with the use of Bayes' theorem to update beliefs and probabilities based on new evidence or data.
10. Nonparametric statistics: This branch deals with statistical methods that do not assume a particular distribution of the data.

15. What are the various branches of statistics?

Statistics can be broadly classified into two main branches:

1. Descriptive Statistics: This branch deals with the collection, organization, analysis, interpretation, and presentation of data. It includes techniques like measures of central tendency, measures of dispersion, frequency distributions, graphical representations, etc.
2. Inferential Statistics: This branch deals with making predictions, inferences, and decisions about a population based on a sample of data. It includes techniques like hypothesis testing, confidence intervals, analysis of variance (ANOVA), regression analysis, etc.

Apart from these two main branches, there are various other sub-branches of statistics, such as:

1. Biostatistics: This branch applies statistical methods to analyze biological and medical data.
2. Econometrics: This branch applies statistical methods to analyze economic data.
3. Social statistics: This branch deals with the analysis of social phenomena and data.
4. Business statistics: This branch deals with the application of statistical methods in the field of business and commerce.
5. Environmental statistics: This branch deals with the analysis of environmental data.
6. Psychometrics: This branch applies statistical methods to analyze psychological data.
7. Educational statistics: This branch deals with the application of statistical methods in the field of education.
8. Statistical genetics: This branch deals with the analysis of genetic data.
9. Quality control: This branch deals with the application of statistical methods to improve the quality of products and services.
10. Time series analysis: This branch deals with the analysis of time-based data, such as stock prices, weather data, etc.

Link to Jupyter Notebook

http://localhost:8888/notebooks/Python%20worksheet.ipynb