Rajesh kakumanu -1843

MACHINE LEARNING – ASSIGNMENT 39

1. The method do we use to find the best fit line for data in Linear Regression is [**Least square method] – (A)**
2. The statement is true about outliers in linear regression – **[Linear regression is sensitive to outliers] – (A)**
3. A line falls from left to right if a slope is **[Negative] – (B)**
4. Which of the following will have symmetric relation between dependent variable and independent variable is **[None of these] – (D)**
5. The following is the reason for over fitting condition is **[Low bias and high variance] – (C)**
6. If output involves label then that model is called as**: [Predictive model] – (B)**
7. Lasso and Ridge regression techniques belong to **[Regularization] – (D)**
8. To overcome with imbalance dataset which technique can be used is **[Cross validation] – (A)**
9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph **[TPR & FPR] – (A)**
10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less is **[True] – (A)**
11. Pick the feature extraction from below: **[Apply PCA to project high dimensional data] – (B)**
12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression – [**We need to iterate**.] –**(C)**
13. Explain the term regularization?

Regularization means- a set of techniques that regularizes learning from particular features for traditional algorithms. It normalizes and moderates weights attached to a feature or a neuron so that algorithms do not rely on just a few features. It is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.I**t is deployed for reducing overfitting by putting network weights small and also enhances the performance of models for new inputs.** ***Regularization assumes that least weights may produce simpler models and hence assist in avoiding overfitting. Examples of regularization are K-means, Neural networks and Random forests etc..There are various regularization techniques to solve the overfitting issues.***

14. Which particular algorithms are used for regularization?

When any dataset not have any noise in it, it will face overfitting problem and parameters will not generalize well on unseen data so, and to avoid these you we need to regularize the weights for better learning. The various regularization techniques such as –

1. Lasso (L1 form) 2. Ridge (L2 form) 3. Dropout regularization techniques 4. Early stopping 5. Data augmentation

Among these techniques the Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge methods are important.

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

Regression is a maximum estimation where we find parameters of relation between dependent and independent variables which maximize the likelihood of getting such samples from the population. Since regression is an estimation, we cannot be completely correct. So, the error term is a catch-all for what we miss out in this estimation because in reality.

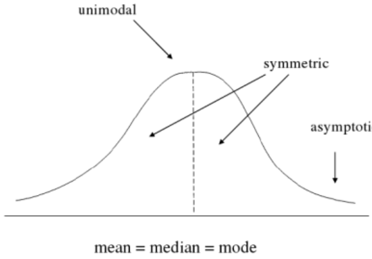
The error term is the difference between the expected price at a particular time and the price that was actually observed. n statistics, an error term is the sum of the deviations of each actual observation from a model regression line.

STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0 [**True**] – **(A)**
2. The following theorem states that the distribution of averages of id variables, properly normalized, becomes that of a standard normal as the sample size increases is [**Central mean theorem**] – **(B)**
3. The following is incorrect with respect to use of Poisson distribution. [**Modeling bounded count data**] – **(B)**
4. The correct statement is [**All of the mentioned**] – **(D)**
5. Random variables are used to model rates. [**Poisson**] – **(C)**
6. Usually replacing the standard error by its estimated value does change the CLT [**True**] – **(A)**
7. The following testing is concerned with making decisions using data is [**Hypothesis**] – **(B)**
8. Normalized data are centered at ---- end have units equal to standard deviations of the original data. [**0**] – **(A)**
9. The following statement is incorrect with respect to outliers? [**Outliers cannot conform to the regression relationship]** – **(C)**

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

The normal distribution is the most important probability distribution(Pd) in statistics because many continuous data in nature and psychology displays the bell-shaped curve when compiled and graphed.Normal distributions are symmetric, unimodal, and asymptotic, and the mean, median, and mode are all equal. A normal distribution is perfectly symmetrical around its center. That is, the right side of the center is a mirror image of the left side. There is also only one mode, or peak, in a normal distribution as shown the figure below-



Normal distributions have key characteristics that are easy to spot in graphs: The mean, median and mode are exactly the same. The distribution can be described by two values: the mean and the standard deviation.

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

Missing data is an unavoidable part of the process. As data researchers, have a lot of resources, time and energy into making sure the data set is as accurate as possible. However, data inevitably goes missing. As someone who has been handling data analytics and overseen dozens of research projects for several years, missing data is just one of those "It sucks, but it's no one's fault" scenarios. Sometimes, data sets come up short, no matter how many times data scientists clean and prepare it. The best way to handle such situations is to develop contingency plans to minimize the damage.

The easiest and used method to handle the missing data is to simply delete the records with the missing value. If the dataset contains a huge number of a sample as corresponding to the missing value this approach is quite feasible and can be implemented on the dataset.

The various imputation techniques to recommend are – 1. Mean or median Imputation. 2. Multi-variate imputations by Chained Equations. 3. Random forest Imputation methods.

12. What is A/B testing?

An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not. A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

There are several benefits of A/B testing. A/B testing will increase user engagement, reduce bounce rates, increase conversion rates, minimize risk, and effectively create content. Running an A/B test can have significant positive effects on a website or mobile app.

13. Is mean imputation of missing data acceptable practice?

Mean as a imputation method is a good choice for series which randomly fluctuate around a certain value/level. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution. The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is typically considered a complex practice since it ignores feature correlation.

Consider an example: We have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does. Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Thus, it is not an acceptable practice.

14. What is linear regression in statistics?

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The main uses of regression analysis are forecasting, time series modeling and finding the cause-and-effect relationship between variables. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

15. What are the various branches of statistics?

Statistics is the branch of mathematics that deals with data. Data is a collection of values. There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

1. ***Data collection is all about how the actual data is collected. Sometimes, data is harder to collect. Sometimes, data is harder to collect. So there are issues in the collection of the data; you need to make sure that the data has been collected fairly before you go on to deal with it, and try to present it and make conclusions.***
2. ***Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically. The basic aim of descriptive statistics is to 'present the data' in an understandable way. It deals with the presentation and collection of data. This is usually the first part of a statistical analysis.***
3. ***Inferential statistics involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics. Inferential statistics is the aspect that deals with making conclusions about the data.***

PYTHON – WORKSHEET 1

1. The following operators is used to calculate remainder in a division is [**%]** – **(C)**

2. In python 2//3 is equal to? **[0**] – **(B)**

3. In python, 6< 2 is equal to? [**24**] – **(C)**

4. In python, 6&2 will give which of the following as output? [**2**]- **(A)**

5. In python, 6|2 will give which of the following as output? [**6**]- **(D)**

6. What does the finally keyword denotes in python? **[It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block**.] – **(B)**

7. What does raise keyword is used for in python? [**It is used to raise an exception**]- **(A)**

8. Which of the following is a common use case of yield keyword in python? [**in defining a generator**]- **(C)**

9. Which of the following are the valid variable names? [**_abc**]- **(A)**

10. Which of the following are the keywords in python? [**All the above**] -**(D)**