MACHINE LEARNING

Q1 - Q11 have only one correct answer

Answer

Q1: A) Least Square Error

Q2: A) Linear regression is sensitive to outliers

Q3: B) Negative

Q4: A) Regression

Q5: C) Low bias and high variance

Q6: B) Predictive Model

Q7: D) Regularization

Q8: D) SMOTE

Q9: A) TPR and FPR

Q10: B) False

Q11: B) Apply PCA to project high dimensional data

In Q12, more than one options are correct.

Q12: A) We don't have to choose the learning rate. & B) It becomes slow when number of features is very large.

Q13-Q15 are subjective answer type questions

Q13. Explain the term regularization?

Ans. we use regularization techniques to moderate learning so that a model can learn instead of memorizing training data. It solves the problem of overfitting. It normalizes and moderates weights attached to a feature, so that algorithms do not rely on just a few features to predict the result.

Q14. Which particular algorithms are used for regularization?

Ans. There are two main regularization techniques:

1: Lasso Regression (least absolute shrinkage and selection operator) (L1): This technique uses absolute weight values for normalization.

2: Ridge Regression (L2): When using this technique, we add the sum of weight's square to a loss function and thus create a new loss function.

Q15. Explain the term error present in linear regression equation

Ans. Error is the difference between actual value and the value which is predicted with the help of regression line.

PYTHON-WORKSHEET 1

Q1-Q8 have only one correct answer.

Answers

Q1: C) %

Q2: B) 0

Q3: C) 24

Q4: A) 2

Q5: D) 6

Q6: C) the finally block will be executed no matter if the try block raises an error or not.

Q7: A) It is used to raise an exception.

Q8: C) In defining a generator.

Q9 and Q10 have multiple correct answers.

Q9: A) _abc & C) abc2

Q10: A) yield & B) raise

STATISTICS WORKSHEET-1

Q1-Q9 have only one correct answer.

Q1: a) True

Q2: a) Central Limit Theorem

Q3: b) Modeling bounded count data

Q4: d) All of the mentioned

Q5: c) Poisson

Q6: b) False

Q7: b) Hypothesis

Q8: a) 0

Q9: c) Outliers cannot conform to the regression relationship

Q10-Q15 are subjective answer type questions

Q10. What do you understand by the term Normal Distribution?

Ans. A normal Distribution is a symmetric, bell-shaped distribution. It is used to represent the distribution of data, when plot on graph, its curve looks like a bell therefore it is also known as bell-shaped curve. Distributions are always centered on the average values, it has no skewness. And the width of the curve is defined by standard deviation. The width of the curve determines how tall the curve is i.e. the wider the curve, the shorter it is, the narrower the curve, the taller it is

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans. We handle missing data by using various imputation techniques. These techniques depends on the nature of missing data.

Some of the techniques are : 1)Listwise deletion: It directly removes the rows that have missing data provided you don't have many columns with missing data or else we could lose big amount of data. 2) Mean/median/mode imputation: It replaces missing data with the mean/median/mode of the column, it is used for numerical missing value only. 3) Using machine learning models to predict the missing data (probably the best approach)

Q12. What is A/B testing?

Ans. When we have two versions of something and we want to find out which version performs better, then we use A/B testing to find out which is better. Let's say we have two versions (eg. version A & version B) of a website with different signup forms and we want to find which will attract more users, then we can use A/B testing. We split the traffic (in our case let's say equally) to both the versions of websites and then we find out how many users signed up for newsletter through version A & how many through version B and then we can compare them and we will know which version is more attractive.

Q13. Is mean imputation of missing data acceptable practice?

Ans. It is an acceptable practice, but mean imputation really ought to be a last resort. It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy to implement. Some of the drawbacks are: 1) Mean imputation does not preserve the relationships among variables. 2) Mean Imputation Leads to an Underestimate of Standard Errors

Q14. What is linear regression in statistics?

Ans. Linear regression is the most basic and commonly used predictive analysis. It is used to predict a dependent variable value based on independent variable value. It uses least-square method to find the best fit line for data. It shows relationship between dependent variable and independent variable.

Q15. What are the various branches of statistics?

Ans. The two main branches of statistics are:

i) Descriptive statistics
ii) Inferential statistics

Descriptive statistics: In this we have limited data, we can analyze it and describe it easily with the help of pie charts, graphs etc. The basic aim of descriptive statistics is to 'present the data' in an understandable way.

Inferential statistics: In this we have tons of data and it's hard to analyze it, so we take sample of data and use it to infer

the requirements. The basic aim of inferential statistics is to make conclusions from the data