MACHINE LEARNING

1 D

2 A

3 B

4 C

5 C

6 B

7 D

8 D

9 A

10 B

11 B

12 A,B,C

13 While training our model, it can be overfitted or under fitted at times, to avoid the problem of overfitting and to get optimal model regularization is used. The model is said to be overfitted when it tries to lean patterns including unwanted ones (noise) leading to the model not predicting properly. Regularization minimizes the adjusted loss function. There are two regularization techniques lasso regularization and ridge regularization.

14 (a) *Ridge regularization*

It modifies the overfitted model by adding a penalty equivalent to the sum of squares of the magnitude of the coefficients. The cost function consists of two-term one being the sum of squared residuals (Loss) and other sum of squares of the magnitude of the coefficient multiplied by the shrinkage factor (λ). Here βj is the slope of the curve/line.

$$\text{cost function} = Loss + \sum|\beta j^2| \times \lambda$$

The higher the value of the shrinkage factor the greater the reduction in the magnitude of the coefficient.

It shrinks the coefficient of those predictors which contribute less and have a weight close to zero but not zero. It contains all predictors for predictions.

(b) Lasso regularization

It modifies the overfitted model by adding a penalty equivalent to the sum of the absolute value of the coefficients. Implies sum of the coefficients can be zero due to negative coefficients (which may be present).

The cost function consists of two-term one being the sum of squared residuals (Loss) and other sum of the absolute value of the coefficients multiplied by the shrinkage factor (λ). Here βj is the slope of the curve/line.

$$\text{cost function} = Loss + \sum||\beta j|| \times \lambda$$

The higher the value of the shrinkage factor the greater the reduction in the magnitude of the coefficient.

It eliminates those features that do not contribute by placing their weights at zero. It doesn't contain all predictors for predictions (at times).

15 The Linear Regression model finds the best fit linear line between independent and dependent variables.

The equation for linear regression is $y = \beta_0 + \beta_1 X_i$
where y is the dependent variable
β_0 intercept
β_1 slope
X_i independent variables

The goal is to get the best value for β_0 and β_1 to obtain the best-fit line having the least error, but in reality, it's difficult to get a line in which all points align there will be points that will be outside the line the least distance between the point and the best-fit line is called the residue. One of the motives of linear regression is to minimize the residue value. Residue can also be defined as the difference between the predicted and actual value.

sum of residue $r = \sum(Y_{pred} - Y_{actual})$

sum of squares of residue $r^2 = \sum (Y_{pred} - Y_{actual})^2$

where Y_{pred} is predicted value
Y_{actual} is actual value
r is the residue

Gradient descent is an optimization algorithm to optimize the cost function (mean squared error of the sum of residue). The cost function is minimized by updating values of β_0 and β_1 till an optimum solution is obtained such that the cost function becomes minimum. In gradient descent, the parameters considered are learning rate based on this the algorithm converges to minima.

The model can then be evaluated using R squared or adjusted R squared.

STATISTICS WORKSHEET-1

1 A

2 A

3 B

4 D

5 C

6 B

7 B

8 A

9 C

10 A distribution is said to be normally distributed also known as Gaussian distribution if the mean is 0 and the standard deviation is 1, it will have no skew and it will be symmetrically distributed. In graphical form, it appears as a bell hence it's called a bell curve.

11 Missing data occurs when no value is available in one or multiple columns of an individual. The missing data can be picked put by using describe function or by using isnull function. As the datas are missing it can impact on results produced by the model. It occurs when the person doesn't know how to fill the particular data.

If a large amount of data is missing in that particular column then we can delete that column as imputing it could cause the model to predict in the wrong way.

There are many imputation techniques involved

- by replacing it with a constant value
- statistical approach (mean, mode, median)
  mean is used if it's numeric and data is not skewed
  the mode can be used for categorical ones.
- Advanced method
  the advantage compared to other methods is we don't need to specify the columns in which the missing data are present and detected automatically hence the amount of code needed is reduced
  1. Simple imputer - it fills all the missing data by mean of that particular column in which the missing data is present.
  2. KNN imputer - imputes the missing data by finding the closest neighbours using Euclidean distance and imputes the value based on the value of the neighbour. It doesn't work with categorical data it can be overcome by encoding.

3. Iterative imputer- where each feature is modelled as a function of other features, Each feature is imputed sequentially one after the other allowing prior imputed values to be used in predicting subsequent features.

12 A/B testing is statistical hypothesis testing, used for making decisions based on sample statistics.

The following steps are done

- by making a hypothesis
- launch test to get statistical evidence to accept or reject

two hypothesis are considered

1. null hypothesis - the condition assumed
2. alternative hypothesis- challenges the null hypothesis assumed

the pvalue is obtained by doing any test like ANOVA, one tail, or two tail tests etc. If pvalue <0/05 null hypothesis is rejected else null hypothesis is accepted.

13 When data is missing the easiest way to fill them is by using the mean for the numerical features. Doing this reduces the model's accuracy and bias. Assume a dataset in which the salary of a person missing who earns 80 lakhs but on taking the mean of the salary in the dataset comes to 35 lakhs which causes an error in prediction as the value used for analysis is wrong.

Mean reduces the variance of the data causing narrow confidence in the probability distribution. Alternatives can be to use KNN imputation or iterative imputation.

14 Linear Regression models the relationship between independent and dependent features. If it uses multiple independent features it is called multiple linear regression and if it uses one independent feature it is called simple linear regression.

The Linear Regression model finds the best fit linear line between independent and dependent variables.

The equation for linear regression is $y = \beta_0 + \beta_1 X_i$
where y is the dependent variable
β_0 intercept
β_1 slope
X_i independent variables

The goal is to get the best value for β_0 and β_1 to obtain the best-fit line having the least error, but in reality, it's difficult to get a line in which all points align there will be points that will be outside the line the least distance between the point and the best-fit line is called the residue. One of the motives of linear regression is to minimize the residue value. Residue can also be defined as the difference between the predicted and actual value.

15 Two types of statistical methods are used for analyzing the data descriptive statistics and inferential statistics.

Descriptive statistics describe the properties of sample and population data. It focuses on central tendency, variability, and distribution of sample data. To describe the dataset, mean, mode and median are used. Variability refers to a set of statistics that shows how much difference there is among the elements and is measured using range, variance, and standard deviation. The distribution refers to the depiction on charts such as histograms or dot plots and uses probability function, skewness, and kurtosis. Descriptive statistics helps to understand the collective property of the data samples and helps in making predictions for inferential statistics.

Inferential statistics are used to come to conclusion about the characteristics of the sample population. It is used to generalize about large groups based on the distribution, variability, and relationship between characteristics within the data sample. Correlation is used to determine the strength and nature of relationships between the dependent and independent features. Regression analysis and hypothesis testing are used for statistical inference.

PYTHON – WORKSHEET 1

1 C

2 B

3 C

4 A

5 D

6 C

7 A

8 C

9 A,C

10 A,B