Assignment 10

Machine Learning

Q-1. In the linear regression equation y = θ0 + θ1x, θ0 is the:

Ans: y intercept(option C)

Q-2. **True or False: Linear Regression is a supervised learning algorithm.**

Ans: True (option A)

Q-3. **In regression analysis, the variable that is being predicted is:**

Ans: the dependent variable(option B)

Q-4. **Generally, which of the following method(s) is used for predicting continuous dependent variables?**

Ans: Linear Regression(option-B)

Q-5. **The coefficient of determination is**:

Ans: the correlation coefficient squared D) equal to zero(option c)

Q-6**. If the slope of the regression equation is positive, then**:

Ans: y increases as x increases(option-B)

Q-7. **Linear Regression works best for:**

Ans: Linear Data (option A)

Q-8. **The coefficient of determination can be in the range of:**

Ans:- 0 to 1 (option A)

Q-9. **Which of the following evaluation metrics can be used for linear regression?**

Ans: RMSE and MAE (option B and D)

Q-10. **Which of the following is true for linear regression?**

Ans: A)Linear regression is a supervised learning algorithm

B) Linear regression supports multi-collinearity.

C) Shape of linear regression's cost function is convex.(option A,B,C are correct)

Q-11. **Which of the following regularizations can be applied to linear regression?**

Ans: ridge and lasso(option A and B)

Q-12**. Linear regression performs better for**

Ans: Large amount of training samples with small number of features.(option A)

Large number of features (option c)

Q-13. **Which of the following assumptions are true for linear regression?**

Ans: Linearity and Non-independent (option A and C)

Q-14. **Explain Linear Regression?**

Ans: Linear regression is a statistical technique used to establish a relationship between a dependent variable and one or more independent variables. In its simplest form, linear regression involves finding the best-fit line through a set of data points. This line can then be used to make predictions about the values of the dependent variable based on the values of the independent variables.

In a linear regression model, the dependent variable is typically denoted as "y", while the independent variable(s) are denoted as "x". The equation for a simple linear regression model can be written as:

y = a + bx

Where "a" represents the intercept, or the point at which the line intersects the y-axis, and "b" represents the slope of the line. The slope indicates how much y changes for every unit change in x.

In a multiple linear regression model, there are multiple independent variables, and the equation is expanded to include a coefficient for each variable:

y = a + b1x1 + b2x2 + ... + bnxn

The goal of linear regression is to find the values of "a" and the "b"s that minimize the difference between the predicted values of y and the actual values of y for a given set of data. This is typically done using a method called "least squares," which involves minimizing the sum of the squared differences between the predicted and actual values of y.

Once the line of best fit has been established, it can be used to make predictions about the values of y for new values of x. This makes linear regression a useful tool for forecasting and trend analysis in many different fields, including economics, finance, and science.

Q-15. **What is difference between simple linear and multiple linear regression?**

Ans: The main difference between simple linear regression and multiple linear regression is the number of independent variables used in the model.

In simple linear regression, there is only one independent variable, denoted as "x", and a single dependent variable, denoted as "y". The goal of the model is to establish a linear relationship between x and y. The equation for a simple linear regression model is:

y = a + bx

Where "a" is the intercept, and "b" is the slope of the line.

In multiple linear regression, there are two or more independent variables, denoted as x1, x2, x3, etc., and a single dependent variable, denoted as "y". The goal of the model is to establish a linear relationship between the independent variables and the dependent variable. The equation for a multiple linear regression model is:

y = a + b1x1 + b2x2 + b3x3 + ...

Where "a" is the intercept, and "b1", "b2", "b3", etc. are the coefficients that represent the effect of each independent variable on the dependent variable.

STATISTICS WORKSHEET-10

Q-1. Rejection of the null hypothesis is a conclusive proof that the alternative hypothesis is?

Ans: Neither(option C)

Q-2. Parametric test, unlike the non-parametric tests, make certain assumptions about

Ans: The underline distribution (option B)

Q-3. The level of significance can be viewed as the amount of risk that an analyst will accept when making a decision

Ans: True (option A)

Q-4. By taking a level of significance of 5% it is the same as saying

Ans: We are 5% confident the results have not occurred by chance (option A)

Q-5. One or two tail test will determine

Ans: If the region of rejection is located in one or two tails of the distribution.(option C)

Q-6. Two types of errors associated with hypothesis testing are Type I and Type II. Type II error is committed when

Ans: We reject a null hypothesis when it is true (option B)

Q-7. A randomly selected sample of 1,000 college students was asked whether they had ever used the drug Ecstasy. Sixteen percent (16% or 0.16) of the 1,000 students surveyed said they had. Which one of the following statements about the number 0.16 is correct?

Ans: It is a sample proportion.(option A)

Q-8. In a random sample of 1000 students, $p\hat{}$ = 0.80 (or 80%) were in favour of longer hours at the school library. The standard error of $p\hat{}$ (the sample proportion) is

Ans: 013 (option A)

Q-9. For a random sample of 9 women, the average resting pulse rate is x = 76 beats per minute, and the sample standard deviation is s = 5. The standard error of the sample mean is

Ans: 1.667 (option C)

Q-10. . Assume the cholesterol levels in a certain population have mean μ= 200 and standard deviation σ = 24. The cholesterol levels for a random sample of n = 9 individuals are measured and the sample mean x is determined. What is the z-score for a sample mean x = 180?

Ans: -2.5 (option C)

Q-11. . In a past General Social Survey, a random sample of men and women answered the question "Are you a member of any sports clubs?" Based on the sample data, 95% confidence intervals for the population proportion who would answer "yes" are .13 to .19 for women and .247 to .33 for men. Based on these results, you can reasonably conclude that

Ans: At least 16% of American women belong to sports clubs.(option B)

Q-12. Suppose a 95% confidence interval for the proportion of Americans who exercise regularly is 0.29 to 0.37. Which one of the following statements is FALSE?

Ans: It is reasonable to say that more than 40% of Americans exercise regularly (option B)

Q-13. How do you find the test statistic for two samples?

Ans: The method for finding the test statistic for two samples depends on the specific hypothesis test being performed. However, the general approach involves calculating a standardized test statistic that measures the difference between the sample statistics under the null hypothesis, relative to the expected variability of the sampling distribution.

For example, in a two-sample t-test, the test statistic can be calculated as follows:

t = (x1 - x2) / (s_pool * sqrt(1/n1 + 1/n2))

where: x1 and x2 are the sample means of the two groups being compared s_pool is the pooled standard deviation, which is an estimate of the common standard deviation of the two populations being sampled n1 and n2 are the sample sizes of the two groups being compared

The t-value is then compared to a critical value from the t-distribution with degrees of freedom calculated using the Welch-Satterthwaite equation, or alternatively, from a table or software. If the calculated t-value is greater than the critical value, the null hypothesis is rejected in favor of the alternative hypothesis.

Q-14. How do you find the sample mean difference?

Ans: To find the sample mean difference between two samples, you subtract the mean of one sample from the mean of the other sample.

The formula for the sample mean difference is:

sample mean difference = x1 - x2

where: x1 is the mean of the first sample x2 is the mean of the second sample

For example, suppose you want to find the sample mean difference for two samples, A and B, with sample means of 10 and 15, respectively. The sample mean difference would be:

sample mean difference = x1 - x2 = 10 - 15 = -5

In this case, the sample mean for sample B is larger than the sample mean for sample A by 5 units, which is represented by a negative value for the sample mean difference.

Q-15. What is a two sample t test example?

Ans: A two-sample t-test is a hypothesis test used to compare the means of two independent samples. Here is an example:

Suppose a researcher is interested in comparing the average height of men and women in a population. The researcher collects two independent samples of height measurements: one sample of 50 men and another sample of 50 women.

The null hypothesis for this two-sample t-test is that there is no difference between the mean heights of men and women in the population. The alternative hypothesis is that there is a significant difference.

The researcher conducts a two-sample t-test and calculates a t-value of 2.34 with 98 degrees of freedom and a p-value of 0.022. This means that the difference between the means of the two samples is statistically significant at a significance level of 0.05, and there is evidence to reject the null hypothesis.

Based on this result, the researcher concludes that the average height of men is statistically different from the average height of women in the population. The researcher also calculates a 95% confidence interval for the difference in means, which can help provide an estimate of the range of likely differences in population means.

WORKSHEET 3 PYTHON

Q-1. Which of the following will raise a value error in python?

Ans: int(-3.2) (option C)

Q-2. What will be the output of round(3.567)?

Ans: 4 (option C)

Q-3. How is the function pow(a,b,c) evaluated in python?

Ans: (a**b)%c (option B)

Q-4. What will be the output of print(type(type(int))) in python 3?

Ans: <class 'type'> (option A)

Q-5. What will be the output of ord(chr(65))?

Ans: 65 (option-C)

Q-6. What is called when a function is defined inside a class?

Ans: Method (option D)

Q-7. What will be the output of all([1, 0, 5 ,7])?

Ans: False (option-B)

Q-8. Is the output of the function abs() the same as that of the function math.fabs()?

Ans: Sometimes(option B)

Q-9. Select all correct float numbers in python?

Ans: A) -68.7e100 C) 4.2038 D) 3.0

Q-10. Which of the following is(are) correct statement(s) in python?

Ans: A) You can pass positional arguments in any order.

B) You can pass keyword arguments in any order.

C) You can call a function with positional and keyword arguments.