1. A study is conducted to examine the relationship between alcohol consumption and school performance. Participants in this study were classified as 'abstainer', 'light drinker', or 'heavy drinker'. The same participants were also classified as 'good', 'moderate', or 'worst' of their class. Results showed that students in the 'good' or the 'worst' of the class were more likely to be 'heavy drinkers' with a p-value of $\leq 0.01$. Which of the following statistical tests were most likely used to generate this results?
   a) Analysis of variance
   b) Chi-square test
   c) Paired t-test
   d) Pearson correlation analysis
   e) Pooles sample t-test

Ans : **(b) Here Chi-square test was used to determine if two categorical attributes alcoholic ( 'abstainer', 'light drinker', or 'heavy drinker') and class performance ('good', 'moderate', or 'worst') are dependent.**

2. Which of the following statements are true and false. Mark 'T' for the true statement and 'F' for the false statement.
   a) Charles Spearman correlation analysis is applicable to only ordinal data.
      **False** - **It is applicable to both interval scale and ordinal data**

   b) Pearson correlation coefficient between two attributes X an Y are high if the covariance of X and Y are high.
      **True - Pearson correlation coefficient is given by r = Cov(X,Y)/$\sigma_x\sigma_y$**

   c) Given a pair of 500 observations with this attributes A, and B, the degree of freedom of data is 499.
      **False - The degree of freedom should be calculated from the contingency table. if A has m and B has n categories of variables, then degree of freedom is (m-1)x(n-1)**

   d) Paired t-test is applicable if a pair of observations are collected from the same participants and the differences of observations are normally distributed.
      **True**

3. The quality of fit is denoted as $R^2$ and coefficient of determination is $r^2$. Mark the following statements as true (T) or false (F).
   a) Given a sample bivariate data both $R^2$ and $r^2$ give the same results.
      **False - $R^2$ measures how good a regression model is, whereas $r^2$ measure what percentage of observations are correlated**.

b) Both $R^2$ and $r^2$ are in the range 0 to 1 both inclusive.

**False** - $R^2$ **= 1 - (SSE/SS). If SSE > SS, then** $R^2$ **can be -ve value. On the other hand,** $r^2$ **is square of e, which has the range of values -1 to +1, both inclusive and hence always +ve values**

c) $R^2$ is used in the context of regression analysis, whereas $r^2$ is used in the context of correlation analysis.

**True**

d) $\chi^2$-value with Chi-square test of correlation analysis is always a positive quantity and its value > 0.

**True**

4. During the Covid-19 pandemic, a study was conducted to see the effect of three vaccines Sputnik, Madena, and Covishield and two drugs Hexaspatomia and Fluorfivis on the blood oxygen saturated level and convulsion disorder. A number of participants across the globe participated in the data collection and test of variances with multiple populations were conducted. Answer the following questions with respect to this study.

a) Identify the dependent variable in the test of variance.

**Dependent variables are blood oxygen saturated level and convulsion disorder.**

b) What are the factors in this study?

**Two factors are vaccine and drug.**

c) How many groups (or labels) were involved in this data analysis?

**Three vaccines and two drugs were considered, so the number of groups was 3x2 = 6**

d) This analysis of variance comes under (tick the correct answer)

i) One-way ANOVA
ii) Tow-way ANOVA
iii) M-way ANOVA
iv) MANOVA

5. Following Table. 1 includes different methods (in column A) and the different task (in column B) to solved. Mark a link from item in column A to item in column B.

| Column A | Column B |
|---|---|
| a) Maximum likelihood estimation | i) Support vector machine |
| b) Error estimation | ii) Naïve Bayes' classification |
| c) Cramer's V-rule | iii) Test of normality |
| d) M-estimation | iv) Logistic regression |
| e) Shapiro-wilk test | v) Degree of correlation between categorial attribute |

Ans : **(a) -> (iv) | (b) -> (i) | (c) -> (v) | (d) -> (ii) | (e) -> (iii)**

6. Table. 2 list some statements as major differences between c4.5 and CART algorithm. Underline the parts of statements which are not stated appropriately.

| C4.5 | CART |
|---|---|
| • Follows information **gain estimation (uses gain ratio)** for selecting splitting attribute | • Follows **gain ratio (uses Gini Index)** to select the splitting attribute |
| • It gives an **n-ary ($n \geq 2$) (Binary)** decision tree | • It gives a **binary (n-ary ($n \geq 2$)** decision tree |
| • Decision trees with this algorithm have **longest (smallest)** height and hence testing time is **small (high)** | • Decision trees with this algorithm are with **shortest (Least)** height and hence testing time is **high (low)** |
| • Suitable for training data with **categorical (numerical)** attributes | • Suitable for classifying data with **numerical (categorical)** attributes |

7. Table. 3 shows some concepts (in column A) and some data analysis activities (in column B). Find a map from an item in column A to an item in column B. Note some items in each column may not be mappable.

| Concept | Data analysis |
|---|---|
| A. Goodness of fit | i) Kernel trick |
| B. Entropy | ii) Sensitivity of classifier |
| C. Predictive accuracy | iii) Decision tree induction |
| D. Random sampling | iv) Auto correlation |
| E. ROC plot | v) Validation of ML algorithm |
| F. Supervised learning | vi) Logistic regression |

**Ans: A -> iv | B -> iii | D -> v | E -> ii**

8. Which of the following specification is true for a perfect classifier? Choose the correct option from the following choices

   a. **TPR =1, FPR=0, Precision=1, $F_1$Score =1 (Answer)**

   b. TPR =0, FPR=1, Precision=0, $F_1$Score =1

c. TPR =1, FPR=1, Precision=0, $F_1$Score =0

d. TPR =0, FPR=0, Precision=1, $F_1$Score =0

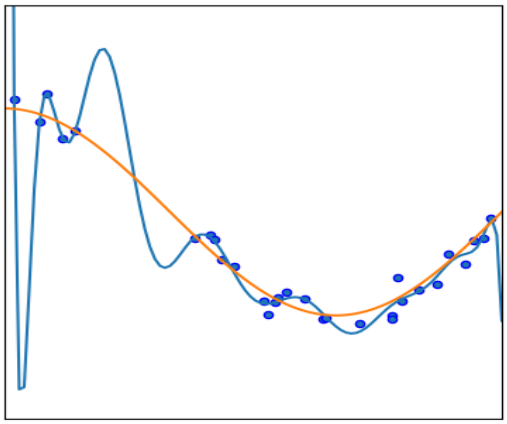
e. None of These

9. In Table-4, Rightmost column is blank and the leftmost column includes some Kernel function. All symbols bear their usual meanings. Fill the blank in the rightmost column

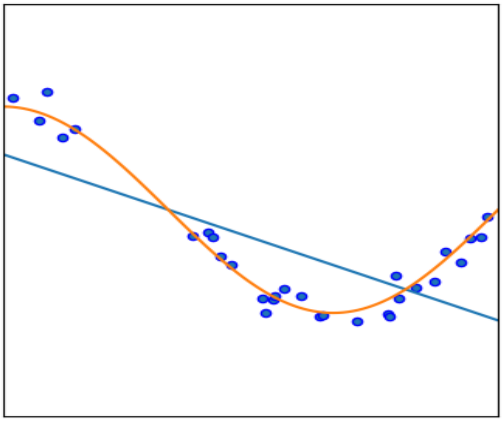
| Kernel Function | Name of the function |
|---|---|
| A. $K(x,y) = \tanh(\beta_1 X^T Y + \beta_0)$ | **Sigmoid Kernel** |
| B. $K(x,y) = \exp[\, c \,\|\| x - y \|\|^2 / 2\sigma^2 \,]$ | **Gaussian RBF Kernel** |
| C. $K(x,y) = \exp[-(x-y)^T A (x-y)]$ | **Mahalanobis Kernel** |
| D. $K(x,y) = (X^T Y + 1)^p$ | **Polynomial Kernel** |

10. Following four diagrams show model fitting under different situation. The figures are not labelled. Some labels are given below (for your hint). Select a label from the list and tag it into a figure which you think appropriate.

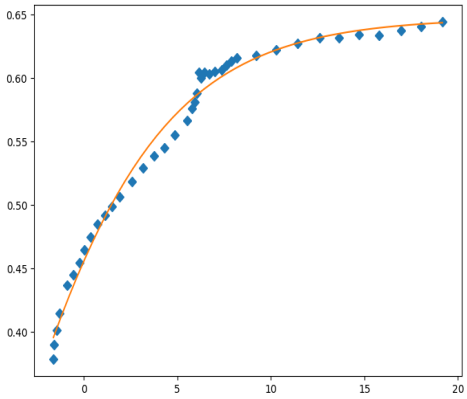
List - Underfitting, Overfitting, Optimal fitting, MMH of linear SVM, MMH of non-linear SVM, Correlation of degree 1, Non-linear regression, Logistic Regression.



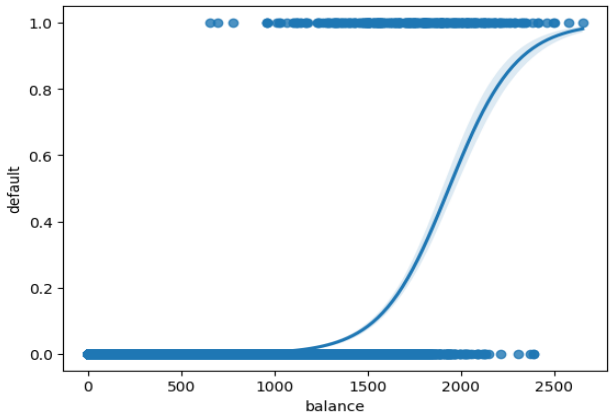
**Overfitting**



**Underfitting**




**Non-linear Regression**




**Logistic Regression.**