WORKSHEET 4

MACHINE LEARNING

Q1 answer : C)High R-squred value for train-set and Low R-squared value for test-set.

Q2 answer : B) Decision trees are highly prone to overfitting.

Q3 answer : D) Decision tree

Q4 answer : C)Precision

Q5 answer : B)Model B

Q6 answer : A) Ridge D)Lasso

Q7 answer : B)Decision Tree

Q8 answer : A) Pruning C)Restricting the max depth of the tree

Q9 answer :

A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points

B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

Q10 answer : Models with tons of predictors tend to perform better in sample than when tested out of sample. The adjusted R 2 "penalizes" you for adding the extra predictor variables that don't improve the existing model. It can be helpful in model selection. Adjusted R 2 will equal R 2 for one predictor variableQ11 answer :

Q11 answer : The main difference between Ridge and LASSO Regression is that if ridge regression can shrink the coefficient close to 0 so that all predictor variables are retained. Whereas LASSO can shrink the coefficient to exactly 0 so that LASSO can select and discard the predictor variables that have the right coefficient of 0.

Q12 answer : VIF measures the strength of the correlation between the independent variables in regression analysis. This correlation is known as multicollinearity, which can cause problems for regression models. While a moderate amount of multicollinearity is acceptable in a regression model, a higher multicollinearity can be a cause for concern.it is always greater than or equal to 1.

WORKSHEET 4

Q13 answer : To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale will help the gradient descent converge more quickly towards the minima.

Q14 answer : 1) Mean squared error or Mean absolute error and 2) R2_score

Q15 answer : Sensitivity or recall =0.95 ,

Specificity =0.82

Precision = 0.80

Accuracy =0.88

WORKSHEET 4

SQL Worksheet 4

Q1 answer : A. commit B.Rollback D.Savepoint

Q2 answer : A.Create C.Drop D.Alter

Q3 answer : B. select name from sales;

Q4 answer : C. Authorizing Access and other control over Database

Q5 answer : B. Column Alias

Q6 answer : B. commit

Q7 answer : A. Parenthesis - (...).

Q8 answer : C.Table

Q9 answer : D. All of the mentioned

Q10 answer : A. ASC

Q11 answer : Denormalization is a database optimization technique in which we add redundant data to one or more tables. This can help us avoid costly joins in a relational database. It is an optimization technique that is applied after normalization.

Q12 answer: A database cursor is an object used to pinpoint records in a database.

Q13 answer : 1)Basic SQL Queries. 2)complex SQL queries. 3)Sub queries

Q14 answer : SQL constraints are used to specify rules for the data in a table. Constraints are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the table.

Q15 answer : Auto Increment is a function that operates on numeric data types. It automatically generates sequential numeric values every time that a record is inserted into a table for a field defined as auto increment.

WORKSHEET 4

Statistics worksheet

Q1 answer : d) All of the mentioned

Q2 answer : a)Discrete

Q3 answer : a) pdf

Q4 answer : c)mean

Q5 answer : c) empirical mean

Q6 answer : a)variance

Q7 answer : c)0 and 1

Q8 answer : b)bootstrap

Q9 answer : b)summarized

Q10 answer : Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.

Q11 answer :

Q12 answer : Hypothesis testing is guided by statistical analysis. Statistical significance is calculated using a p-value, which tells you the probability of your result being observed, given that a certain statement (the null hypothesis) is true. If this p-value is less than the significance level set (usually 0.05), the experimenter can assume that the null hypothesis is false and accept the alternative hypothesis. Using a simple t-test, you can calculate a p-value and determine significance between two different groups of a dataset.

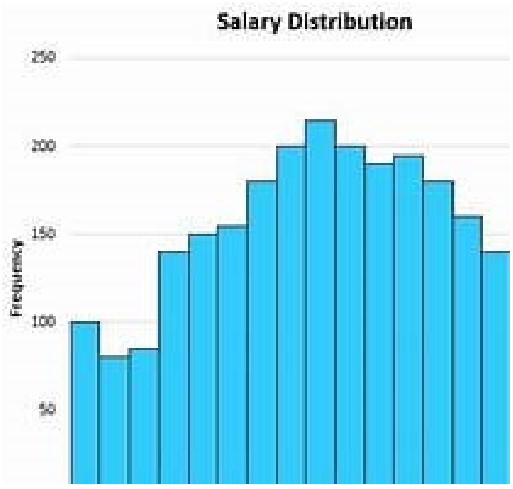
Q13 answer :

1)Allocation of wealth among individuals and

2)Values of oil reserves among oil fields (many small ones, a small number of large ones)

WORKSHEET 4

Q14 answer : consider the following distribution of salaries for residents in a certain city: The median does a better job of capturing the “typical” salary of a resident than the mean.



Q15 answer: The likelihood function (often simply called the likelihood) represents the probability of random variable realizations conditional on particular values of the statistical parameters. Thus, when evaluated on a given sample, the likelihood function indicates which parameter values are more likely than others, in the sense that they would have made the observed data more probable