STATISTICS WORKSHEET-5

Q1 to Q10 are MCQs with only one correct answer. Choose the correct option.

1. Using a goodness of fit, we can assess whether a set of obtained frequencies differ from a set of frequencies.
a) Mean b) Actual c) Predicted d) Expected
ANS. d
2. Chisquare is used to analyse a) Score b) Rank c) Frequencies d) All of these
ANS. c
3. What is the mean of a Chi Square distribution with 6 degrees of freedom? a) 4 b) 12 c) 6 d) 8
ANS. c
4. Which of these distributions is used for a goodness of fit testing?
a) Normal distribution b) Chisqared distribution c) Gamma distribution d) Poission distribution
ANS. b
5. Which of the following distributions is Continuous
a) Binomial Distribution b) Hypergeometric Distribution c) F Distribution d) Poisson Distribution
ANS. c
6. A statement made about a population for testing purpose is called?
a) Statistic b) Hypothesis c) Level of Significance d) TestStatistic
ANS. b
7. If the assumed hypothesis is tested for rejection considering it to be true is called?
a) Null Hypothesis b) Statistical Hypothesis c) Simple Hypothesis d) Composite Hypothesis
ANS. a
8. If the Critical region is evenly distributed then the test is referred as?
a) Two tailed b) One tailed c) Three tailed d) Zero tailed
ANS.
9. Alternative Hypothesis is also called as?
a) Composite hypothesis b) Research Hypothesis c) Simple Hypothesis d) Null Hypothesis
ANS. b
10. In a Binomial Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by a) np b) n
ANS. a

MACHINE LEARNING WORKSHEET-5

Q1 to Q15 are subjective answer type questions, Answer them briefly.

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?
ANS.
2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.
ANS. TSS (Total Sum of Squares) = The total sum of squares refers to a statistical technique used in regression analysis to determine the dispersion of data points. The sum of squares can be used to find the function that best

fits by varying the least from the data. In a regression analysis, the goal is to determine how well a data series can be fitted to a function that might help to explain how the data series was generated.

The sum of squares measures the deviation of data points away from the mean value.

For a set *X* of *n* items:

Sum of squares=*i*=0∑*n* (*Xi* −*X*)2

Xi =The *ith* item in the set

X=The mean of all items in the set

(*Xi* −*X*)=The deviation of each item from the mean

RSS (Residual Sum of Squares) = The residual sum of squares (RSS) measures the level of variance in the error term, or residuals, of a regression model. The RSS measures the amount of error remaining between the regression function and the data set after the model has been run. A smaller RSS figure represents a regression function that is well-fit to the data

How to Calculate the Residual Sum of Squares

$RSS = \sum^{n}_{i=1} (y^{i} - f(x_{i}))^{2}$

y_{i} = *the* i^{th} *value of the variable to be predicted*

$f(x_{i})$ = *predicted value of* y_{i}

n = upper limit of summation

Explained sum of square (ESS) is a statistical quantity used in modeling of a process. ESS gives an estimate of how well a model explains the observed data for the process. It tells how much of the variation between observed data and predicted data is being explained by the model proposed. Mathematically, it is the sum of the squares of the difference between the predicted data and mean data

$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

ESS(Explained sum of square) = total sum of squares – residual sum of squares

3. What is the need of regularization in machine learning?

ANS. Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting.

Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

4. What is Gini–impurity index?

ANS. Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure

5. Are unregularized decision-trees prone to overfitting? If yes, why?

ANS. Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions

6. What is an ensemble technique in machine learning?

ANS. Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model.

7. What is the difference between Bagging and Boosting techniques?

8. What is out-of-bag error in random forests?

ANS. The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RandomForest Classifier to be fit and validated whilst being trained

9. What is K-fold cross-validation?

ANS. K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability When given new data. K refers to the number of groups the data sample is split into. For example, if you see that the k-value is 5, we can call this a 5-fold cross-validation.

10. What is hyper parameter tuning in machine learning and why it is done?

ANS. Hypermeter tuning consists of findings a set of optimal hyperparameter values for a learing algorithm while applying this optimized algorithm to any data set.

That combination of hyperparameters maximizes the model's performance.

11. What issues can occur if we have a large learning rate in Gradient Descent?

ANS. In Gradient Descent, we must set the learning rate to an appropriate value. If the learning rate is very large we will skip the optimal solution. If we take too small learning rate then we will need too many iterations to converge to the best value

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

ANS. Logistic Regression does not perform well on Non-Linear data because it has a linear decision surface

13. Differentiate between Adaboost and Gradient Boosting.

ANS. Gradient Boosting = The way Gradient Boosting its result is to find the difference between prediction and actual value of an instance. Then, the target label of that instance will be replaced with this subtraction in the next round. Difference between actual and prediction comes from the derivative of the mean squared error as a loss function.

In gradient boosting, each tree has a same weight. To make a final decision, we will find the sum of the predictions of those sequential trees.

AdaBoost= It builds a decision tree, then it will increase the target label for incorrectly predicted ones, and it will decrease the target label value for correctly predicted ones. In this way, predictions with high error will be more important in the next rounds.

trees have weights in adaboost. Each tree will contribute to the prediction with respect to its weight.

14. What is bias-variance trade off in machine learning?

ANS. While building the machine learning model, it is really important to take care of bias and variance in order to avoid overfitting and underfitting in the model. If the model is very simple with fewer parameters, it may have low variance and high bias. Whereas, if the model has a large number of parameters, it will have high variance and low bias. So, it is required to make a balance between bias and variance errors, and this balance between the bias error and variance error is known as the Bias-Variance trade-off.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM

ANS. Linear kernel = Linear Kernel is used when the data is linearly seperable, that is it can be seperated using a single line

RBF kernel =

Polynomial kernel=

WORKSHEET 5   SQL

1. Write SQL query to show all the data in the Movie table.

```
 SELECT * FROM movie
```

2. Write SQL query to show the title of the longest runtime movie.

```
  SELECT title FROM movie WHERE runtime= (SELECT MAX(runtime) FROM movie);
```

3. Write SQL query to show the highest revenue generating movie title.

```
  SELECT title FROM movie WHERE revenue= (SELECT MAX(revenue) FROM movie);
```

4. Write SQL query to show the movie title with maximum value of revenue/budget.

```
 SELECT title, revenue FROM movie WHERE revenue/budget = (SELECT MAX(revenue/budget) FROM movie);
```

5. Write a SQL query to show the movie title and its cast details like name of the person, gender, character name, cast order.

```
  SELECT movie.title, mc.character_name,mc.cast_order FROM
            INNER JOIN movie_cast as mc ON movie.movie_id= mc.movie_id
INNER JOIN gender as G ON movie_cast.gender_id= G.gender_id
            INNER JOIN person on movie_cast.person_id= person.person_id;
```

6. Write a SQL query to show the country name where maximum number of movies has been produced, along with the number of movies produced.

7. Write a SQL query to show all the genre_id in one column and genre_name in second column.

```
  select movie_genre.genre_id, genre.genre_name from movie_genre
     inner join genre on movie_genre.genre_id= movie_genre.genre_id;
```

8. Write a SQL query to show name of all the languages in one column and number of movies in that particular column in another column.

```
  SELECT L.country_name, count(m_l.movie_id) AS numbers_of_movie FROM languages as L
     INNER JOIN movie_language AS m_l ON L.language_id= m_l.language_id
     GROUP BY m_l.movie_id;
```

9. Write a SQL query to show movie name in first column, no. of crew members in second column and number of cast members in third column.

10. Write a SQL query to list top 10 movies title according to popularity column in decreasing order.

```
   SELECT title FROM movie ORDER BY popularity DESC LIMIT 10;
```

11. Write a SQL query to show the name of the 3rd most revenue generating movie and its revenue.

```
    SELECT title, revenue FROM movie ORDER BY revenue DESC LIMIT 2,1;
```

12. Write a SQL query to show the names of all the movies which have “rumoured” movie status.

```
  SELECT movie_status, title FROM movie WHERE movie_status= 'rumoured';
```

13. Write a SQL query to show the name of the “United States of America” produced movie which generated maximum revenue.

```
 SELECT c.country_name ,movie.title FROM country AS c
      inner join production_country AS pc ON c.country_id=pc.country_id
      inner join movie ON pc.movie_id=movie.movie_id WHERE c.country_name= 'United States of America'
     ORDER BY movie.revenue DESC LIMIT 1;
```

14. Write a SQL query to print the movie_id in one column and name of the production company in the second column for all the movies.

```
 SELECT movie.movie_id, p.company_name FROM movie
INNER JOIN movie_company ON movie.movie_id= movie_company.movie_id
            INNER JOIN production_company AS p ON movie_company.company_id=
            p.company_id;
```

15. Write a SQL query to show the title of top 20 movies arranged in decreasing order of their budget.

```
 SELECT title FROM movie ORDER BY budget DESC LIMIT 20;
```