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.

Expected

2. Chisquare is used to analyse

Frequencies

3. What is the mean of a Chi Square distribution with 6 degrees of freedom?

6

4. Which of these distributions is used for a goodness of fit testing?

Chisqared distribution

5. Which of the following distributions is Continuous

Binomial Distribution

6. A statement made about a population for testing purpose is called?

Hypothesis

7. If the assumed hypothesis is tested for rejection considering it to be true is called?

Null Hypothesis

8. If the Critical region is evenly distributed then the test is referred as?

Two tailed

9. Alternative Hypothesis is also called as?

Research Hypothesis

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

np

MACHINE LEARNING

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?

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable.So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model.

In investing, R-squared is generally interpreted as the percentage of a fund or security's movements that can be explained by movements in a benchmark index.

An R-squared of 100% means that all movements of a security (or other dependent variable) are completely explained by movements in the index (or the independent variable(s) you are interested in).

A residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term.

A residual sum of squares (RSS) measures the level of variance in the error term, or residuals, of a regression model.

Ideally, the sum of squared residuals should be a smaller or lower value than the sum of squares from the regression model's inputs.

The RSS is used by financial analysts in estimating the validity of their econometric models

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.

TSS-The coefficient of determination is used as a measure of how well a regression line explains the relationship between a dependent variable (Y) and an independent variable (X). The closer the coefficient of determination is to 1, the more closely the regression line fits the sample data.

ESS-Explained sum of square (ESS) or Regression sum of squares or Model sum of squares 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.

RSS-A residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term.

TSS = ESS + RSS, where TSS is Total Sum of Squares, ESS is Explained Sum of Squares and RSS is Residual Sum of Suqares.

The aim of Regression Analysis is explain the variation of dependent variable Y.

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

This is a form of regression that constrains / regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of over fitting.

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

The commonly used regularization techniques are :

L1 regularization.

L2 regularization.

Regularization basically adds the penalty as model complexity increases.

Regularization parameter (lambda) penalizes all the parameters except intercept so that model generalizes the data and won't over fit.

A regression model which uses L1 Regularization technique is called LASSO(Least Absolute Shrinkage and Selection Operator) regression.

A regression model that uses L2 regularization technique is called Ridge regression.

Lasso Regression adds “absolute value of magnitude” of coefficient as penalty term to the loss function(L).

Need for Regularization Technique

Over fitting: Over fitting results in the model failing to generalize on the unseen dataset

Multicollinearity: Model suffering from multicollinearity effect

Computationally Intensive: A model becomes computationally intensive

4. What is Gini–impurity index?

Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.But what is actually meant by ‘impurity’? If all the elements belong to a single class, then it can be called pure.

The degree of Gini index varies between 0 and 1, where 0 denotes that all elements belong to a certain class or if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes. A Gini Index of 0.5 denotes equally distributed elements into some classes.

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

Decision trees are prone to over fitting, 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.

In decision trees, over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. Thus it ends up with branches with strict rules of sparse data. Thus this affects the accuracy when predicting samples that are not part of the training set

6. What is an ensemble technique in machine learning?

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 would.

A machine learning technique that combines several base models in order to produce one optimal predictive model.

Ensemble learning methods are popular and the go-to technique when the best performance on a predictive modeling project is the most important outcome

Ensembles offer two specific benefits on a predictive modeling project, and it is important to know what these benefits are and how to measure them:

A minimum benefit of using ensembles is to reduce the spread in the average skill of a predictive model.

A key benefit of using ensembles is to improve the average prediction performance over any contributing member in the ensemble.

The mechanism for improved performance with ensembles is often the reduction in the variance component of prediction errors made by the contributing models.

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

Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.

Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

In Bagging the result is obtained by averaging the responses of the N learners (or majority vote).

However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates.

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

Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees,and other machine learning models utilizing bootstrap aggregating (bagging). Bagging uses sub sampling with replacement to create training samples for the model to learn from.

The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in theirrespective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained 1.

9. What is K-fold cross-validation?

K-fold cross-validation is one of the most commonly used model evaluation methods. Even though this is not as popular as the validation set approach, it can give us a better insight into our data and model.

While the validation set approach is working by splitting the dataset once, the k-Fold is doing it five or ten times.

K-Fold CV is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point.

Evaluating a Machine Learning model can be quite tricky. Usually, we split the data set into training and testing sets and use the training set to train the model and testing set to test the model. We then evaluate the model performance based on an error metric to determine the accuracy of the model.

This method however, is not very reliable as the accuracy obtained for one test set can be very different to the accuracy obtained for a different test set.

K-fold Cross Validation (CV) provides a solution to this problem by dividing the data into folds and ensuring that each fold is used as a testing set at some point.

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

Hyper parameters tuning is crucial as they control the overall behavior of a machine learning model.

Every machine learning models will have different hyper parameters that can be set.

A hyper parameter is a parameter whose value is set before the learning process begins.

Approaches to Hyper parameter tuning:

Manual Search

Random Search

Grid Search

Hyper parameters control the over-fitting and under-fitting of the model. Optimal hyper parameters often differ for different datasets.

To get the best hyper parameters the following steps are followed:

1. For each proposed hyper parameter setting the model is evaluated

2. The hyper parameters that give the best model are selected.

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

When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error.

Most of the time the reason for an increasing cost-function when using gradient descent is a learning rate that's too high.

The learning rate controls how quickly the model is adapted to the problem. Smaller learning rates require more training epochs given the smaller changesmade to the weights each update, whereas larger learning rates result in rapid changes and require fewer training epochs.

A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution, whereas a learning rate that is too small can cause the process to get stuck.

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

No, Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters.

Or in other words, the output cannot depend on the product of its parameters.

Logistic regression is known and used as a linear classifier.

It is used to come up with a hyper plane in feature space to separate observations that belong to a class from all the other observations that do not belong to that class. The decision boundary is thus linear. Robust and efficient implementations are readily available to use logistic regression as a linear classifier.

13. Differentiate between Adaboost and Gradient Boosting.

In Adaboost, ‘shortcomings’ are identified by high-weight data points.

In Gradient Boosting, ‘shortcomings’ (of existing weak learners) are identified by gradients.Adaboost is more about ‘voting weights’ and Gradient boosting is more about ‘adding gradient optimization.

Adaboost increases the accuracy by giving more weightage to the target which is misclassified by the mode.

Gradient boosting calculates the gradient (derivative) of the Loss Function with respect to the prediction (instead of the features).

Gradient boosting increases the accuracy by minimizing the Loss Function (error which is difference of actual and predicted value) and havingthis loss as target for the next iteration.

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

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.

Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but have high error rates on test data.If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without over fitting and under fitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

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

Linear Kernel is used when the data is linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a large number of Features in a particular Data Set. Training a SVM with a Linear Kernel is faster than with any other Kernel.

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.RBF Kernel is popular because of its similarity to K-Nearest Neighborhood Algorithm. It has the advantages of K-NN and overcomes the space complexity problem as RBF Kernel Support Vector Machines just needs to store the support vectors during training and not the entire dataset.

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

Usually linear and polynomial kernels are less time consuming and provides less accuracy than the rbf or Gaussian kernels

SQL WORKSHEET-5

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/budget as revenue_budget_ratio from movie
order by revenue_budget_ratio desc
limit 1;
```

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 title , person_name , gender, character_name, cast_order

from movie as m inner join movie_cast AS c

on m.movie_id = c.movie_id

inner join gender as g

on c.gender_id = g.gender_id

inner join person as p

on c.person_id = p.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.

```
select country_name, count(movie_id) as no_of_movies

from production_country as a inner join country as b

on a.country_id = b.country_id

group by country_name

order by no_of_movies desc

limit 1;
```

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

```
select * from genre;
```

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 language_name, count(movie_id) as no_of_movies

from movie_languages as a inner join language as b

on a.language_id = b.language_id`

group by.language_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.

```
select title, count(person_id) as no_of_cast

from movie as a inner join movie_cast as b

on .movie_id = b.movie_id

group by  b.movie_id;
```

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 from movie
```

```
order by revenue desc
limit 1
offset 2;
```

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

```
select 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 title, revenue
from movie as a inner join production_country as b
on a.movie_id = b.movie_id
inner join country as c
on c.country_id = b.country_id
where country_name ="United States Of America"
order by 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_id, company_name
```

```
from movie_company as a inner join production_company as b

on a.company_id = b.company_id;
```

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

```
select title, budget from movie

order by budget desc

limit 20;
```