**STATISTICS WORKSHEET-5**

Answer No. 1: D

Answer No. 2: C

Answer No. 3: C

Answer No. 4: B

Answer No. 5: C

Answer No. 6: B

Answer No. 7: A

Answer No. 8: A

Answer No. 9: B

Answer No. 10: A

**MACHINE LEARNING**

Answer No. 1: Goodness-of-fit is a statistical hypothesis test to see how well sample data fit a distribution from a population with a normal distribution.

Investors have found success using Residual Sum of Squares over R-squared because of its ability to make a more accurate view of the correlation between one variable and another. RSS does this by taking into account how many independent variables are added to a particular model against which the stock index is measured. Therefore it totally depends on individual which one of them perform better in measuring of goodness of fit.

Answer No. 2: The coefficient of determination, R2, is a statistical measure that shows the proportion of variation explained by the estimated regression line. Variation refers to the sum of the squared differences between the values of Y and the mean value of Y

*R*2 always takes on a value between 0 and 1. The closer *R*2 is to 1, the better the estimated regression equation fits

Is also known as the total number of squares (TSS).

Now this sum can be divided into the following two categories:

* **Explained sum of squares (ESS):** The ESS is the portion of total variation that measures how well the regression equation explains the relationship between *X* and *Y*.
* **Residual sum of squares (RSS):** This expression is also known as unexplained variation and is the portion of total variation that measures discrepancies (errors) between the actual values of Y and those estimated by the regression equation.

The smaller the value of RSS relative to ESS, the better the regression line fits.

The sum of RSS and ESS equals TSS.

Answer No. 3: Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.  
The commonly used regularisation techniques are:

L1 regularisation

L2 regularisation

Dropout regularisation

Answer No. 4: Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree. More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

Gini =

Answer No. 5: Overfitting can be one problem that describes if your model no longer generalizes well.

Overfitting happens when any learning processing overly optimizes training set error at the cost test error. While it’s possible for training and testing to perform equality well in cross validation, it could be as the result of the data being very close in characteristics, which may not be a huge problem. In the case of decision tree’s, they can learn a training set to a point of high granularity that makes them easily overfit. Allowing a decision tree to split to a granular degree, is the behaviour of this model that makes it prone to learning every point extremely well to the point of perfect classification — ie: overfitting.  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.

Answer No. 6: 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. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account, and average those models to produce one final model.

Answer No. 7: Bagging and Boosting are two types of Ensemble Learning. These two decrease the variance of single estimate as they combine several estimates from different models. So, the result may be a model with higher stability.

|  |  |
| --- | --- |
| Bagging | Boosting |
| Simplest way of combining predictions that belong to the same type. | A way of combining predictions that belong to different types. |
| Each model receives equal weight | Models are weighted according to their performance. |
| Bagging tries to solve over-fitting | Boosting tries to reduce bias |
| If the classifier is having high variance, then apply bagging. | If the classifier is stable and high bias then apply boosting. |
| Random Forest | Gradient Boosting |
| Each model is built independently. | New models are influenced by performance of previously built models. |

Answer No. 8: The RandomForestClassifier is trained using bootstrap aggregation, where each new tree is fit from a bootstrap sample of the training observations zi=(xi,yi). The out-of-bag (OOB) error is the average error for each zi calculated using predictions from the trees that do not contain zi in their respective bootstrap sample.

Answer No. 9: Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. It runs on a single parameter called k that refers to the number of groups that a given data sample is to be split into. The procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Answer No. 10: In machine learning, *hyperparameter optimization* or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

Answer No. 11: When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error. When the learning rate is too small, training is not only slower, but may become permanently stuck with a high training error.

 A large learning rate allows the model to learn faster, at the cost of arriving on a sub-optimal final set of weights. A smaller learning rate may allow the model to learn a more optimal or even globally optimal set of weights but may take significantly longer to train.

At extremes, a learning rate that is too large will result in weight updates that will be too large and the performance of the model (such as its loss on the training dataset) will oscillate over training epochs. Oscillating performance is said to be caused by weights that diverge.

Answer No. 12: Logistic regression cannot be used for classification of non-linear data as the decision boundary is linear.

Answer No. 13: Adaboost requires users to specify weak learners before the learning process. At each iteration adaboost changes the sample distribution by modifying the weights attached to each of the instances. It increases the weights of the wrongly predicted instances and decreases the weights of the correctly predicted instances. The weak learners thus focus on the difficult instances.

Gradient boosting is based on minimising the loss function. Gradient boosting generates learners during the learning process. It builds first learner to predict the values and calculate the loss, it will build the second learner to predict the loss after the first step and so on.

Answer No. 14: A bias variance trade-off is a way to diagnose the performance of an algorithm by breaking down its prediction error. Bias and variance are the prediction errors. If our model is too simple and has a 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 will have high variance and low bias. We need to find a good balance between bias and variance without overfitting and underfitting the data. We can say an algorithm can’t be more complex or less complex at the same time.

Answer No. 15: Linear kernel is used when the data is linearly separable i.e., it can be separated using a single line. It is used when there are a large number of features in the dataset. It is one of the most common used kernels. RBF or radial basis function is used when the boundaries are hypothesized to be curve shaped.

**SQL WORKSHEET-5**

Answer No. 1: select \* from movie;

Answer No. 2: select title from movie order by runtime desc limit 1;

Answer No. 3: select title from movie order by revenue desc limit 1;

Answer No. 4: select title from movie order by budget desc limit 1;

Answer No. 5: select title, person\_name, gender, character\_name, cast\_order

-> from movie

->inner join movie\_cast on movie.movie\_id=movie\_cast.movie\_id

-> inner join gender on movie\_cast.gender\_id=gender.gender\_id

->inner join person on movie\_cast.person\_id=person.person\_id;

Answer No. 6: select country\_name, count(movie\_id) as count\_movie

-> from country

-> inner join production\_country on country.country\_id=production\_country.country\_id

-> group by country\_name order by count(movie\_id) desc limit 1;

Answer No. 7: select \* from genre;

Answer No. 8: select language\_name, count(movie\_id) as movie\_count

-> from language

-> inner join movie\_language on language.language\_id=movie\_language.language\_id

-> group by language\_name;

Answer No. 9:

Answer No. 10: select title from movie order by popularity desc limit 10;

Answer No. 11: select title, revenue

-> from (select title, revenue from movies order by revenue desc limit 3) as comp

-> order by revenue limit 1;

Answer No. 12: select title from movie where movie\_status=’rumoured’;

Answer No. 13: select title from movie

-> inner join production\_country on movie.movie\_id=production\_country.movie\_id

-> inner join country on production\_country.country\_id=country.country\_id

-> where country\_name='United States of America'

-> order by revenue desc limit 1;

Answer No. 14: select movie\_id, company\_name

-> from movie\_company inner join production\_company

-> on movie\_company.company\_id=production\_company.company\_id

Answer No. 15: select title from movie order by budget desc limit 20;