Statistics worksheet-5

1. d (Expected)
2. c (Frequencies)
3. c (6)
4. b (Chisquared distribution)
5. c (F-distribution)
6. b (Hypothesis)
7. a (Null hypothesis)
8. a (Two-tailed)
9. b (Research hypothesis)
10. a (np)

Machine learning-5

1. R-squared is a better measure of the goodness of fit of a regression model because it represents the proportion of the total variance in the respobnse variable that is explained by the model, with values between 0 and 1. Higher values of R-squared indicates a better fit of model.However Residual sum of squares (RSS) does not provide informatin on the proportion of explained variance, making R-squared a more useful measure of the goodness of fit in regression analysis.

2. **TSS** (Total Sum of squares) is the sum of squares of the difference between the actual response(or dependent variable) values and their mean. It represents the total variance in the response variable.
ESS (Explained Sum of squares) is the sum of squares of the difference between the predicted response values (obtained from

regression model) and the mean of the response variable. It represents the amount of variance in the response variable that is explained by predictor variables.

RSS (Residual Sum of squares) is the sum of squares of the difference between the actual response values and the predicted response values. It represents the amount of variance in the response variable that is not explained by predictor variables.

Equation that relate these three metrics i.e.- "TSS=ESS+RSS" .

3. The need of regularization technique in machine learning is to prevent overfitting which occurs when a model is too complex and fits the training data to well, but performs poorly on new unseen data.The main purpose of regularization is to add a penalty to the loss function that helps to prevent overfitting and improve the generalization ability of the model.(Overfitting can results in a model that has high variance and low generalization ability) There are two common type of regularization techniques in machine learning i.e. **L1 (Lasso) and L2 (Ridge).**

Both technique add penalty to the loss function but L1 add a penalty proportional to the absolute value of the coefficients while L2 add a penalty proportional to the square of the coefficients.

Regularization is especially important in situations where the number of features is large, as in these cases, most likely the model will overfit and the data is left unregularized.

4. The Gini-impurity index is a simple yet effective measure of impurity or disorder of a set of classes, used in decision tree-based algorithms such as CART(classification and regression trees) for binary classification. It is a measure, also for evaluating the quality of a split in decision trees and other algorithms that use decision tree as a base.It is used to determine the best features to split the data at each nodes in the tree. It shows how well a given features separates the classes in the data. It ranges from 0(perfect separation,no impurity) to 1(complete disorder, maximum impurity).

5. Yes, unregularized decision-trees are prove to overfitting. Because the decision-tree algorithm splits the data into smaller

and smaller subgroups, creating a highly complex tree structure that fits a training data very well but may not generalize well to new data.
The tree continues to split the data until it reaches a minimum size or until a certain criteria is met. This results in a tree with large numbers of leaves, each of which may correspond to a very small subset of the data. As a result, the tree may capture and fit noise in the training data , leading to overfitting.

6. Ensembles techniques in machine learning are methods that combine multiple models to create a single, more robust model. The idea behind these techniques is to exploit the strengths of multiple models and mitigate the weakness of individual models. There are several types of ensemble techniques i.e.

a. Bagging (Bootstrap Aggregating) : This involves training multiple instances of a single model on different subsets of the training data. The prediction of the individual models are then combined to produce a single prediction.

b. Boosting: This involves training multiple models sequentially, where each model is trained to correct the mistakes made by the previous model.

c. Stacking: This involves training multiple models on the same data, and then using their predictions as features to train a higher level model.
Ensemble techniques can often result in improved performance compared to a single model, as the combination of multiple models can reduce overfitting and improve generalization.

7. Bagging and boosting are two ensemble learning methods used in machine learning to improve the accuracy of a model by combining the outputs of multiple models. The factor which differ both of these techniques are the category and its function. Bagging is a parallel ensemble technique which reduces overfitting by reducing the variance of the model through averaging predictions. While boosting is a sequential ensemble technique which reduces overfitting by reducing the bias of the medal through iteratively focusing on the instances that are misclassified.

8. Out-of-bag error in a random forest is a measure of the accuracy of a random forest model, calculate without using the samples that are used in building individual trees. The OOB errors can be used as a measure of the generalization performance of the random forest model, as it provides an estimate of the model's accuracy on new and unseen data. It can also be used for model selection and feature selection. The OOB error provides a convenient way to access the performance of random forest models without the need for cross-validation or separate validation sets, as the OOB samples are used as an internal validation set.

9. K-fold cross validation is a resampling procedure used in machine learning to access the performance of a model on a dataset. It involves dividing the original sample into k-subsets, or folds of approximately equal size. It is a useful tool for estimating the performance of a model when a limited sample size is available. The value of k is typically set to 5 or10, but can be chosen based on sample size and the desired level of precision.

The steps of k-fold cross validation are:-

a. Partition of the original sample into k-subsets, or folds of equal size.

b. Train the model on k-1 of the folds, and evaluate it on the remaining one.

c. Repeat the process k times, using a different fold as the evaluation set each time.

d. Average the performance across all k iterations.

The performance measure used in it is often the mean squared error(MSE) or accuracy.

10. Hyperparameter tuning is the process of finding the best set hyperparameters for a machine learning model, so as to optimize its performance on a given task. (Hyperparameters are parameters that are set before training a machine learning model, and they control the learning process and the complexity of the model) Hyperparameter tuning is necessary because the performance of the machine learning model is heavily influenced by the choice of hyperparameters. Hyperparameter

tuning can be performed using grid search, random search, or gradient based optimization. The goal of hyperparameter tuning is to find the hyperparameters that yield the best performance, as measured by the performance metrics such as accuracy,F1 score or AUC.

11. A large learning rate can lead to instability and poor convergence in the optimization process, making it difficult to find the optimal solution.
Issues occured when we have a large learning rate in Gradient Descent are:
a. Oscillations
b. Slow convergence
c. Divergence
d. Overshooting the minimum

12. Logistic regression can be used for classification of non-linear data, but its performance may be limited in cases where the relationship between the input features and the target variable is highly non-linear. In these cases more complex models such as decision trees, random forest, or neural networks might be more works.

13. Adaboost and Gradient boosting are both ensemble techniques used for improving the performance of weak machine learning models, but they differ in their approach and implementation
Where Adaboost focuses on adjusting the weights of the training samples, while gradient boosting focuses on improving the prediction of a weak learner by adding a new weak learner that corrects the error made by previous weak learner.

14. The bias-variance trade-off is a fundamental concept in a machine learning that refers to the tension between the desire for a model to fit in a training data well(low bias) and the desire for the model to be able to generalize well to new, unseen data(low-variance).
A model with low bias oversimplifies the relationship between the input features and the target variable, while a model with high variance overfits the training data. Finding the right

balance between these two factors is crucial for good performance in machine learning and that is the goal of machine learning.

15. Support vector machine(SVM) are a popular machine learning algorithm for classification and regression problems. The choice of kernel function in SVM is important for determining the nature of the decision boundary between the classes and it depends on the nature of the problem and the relationship between the input features and the target variable. Linear, Radial basis function, and Polynomial kernels are different functions that can be used in SVM. The linear kernel is well-suited for linear problems, the RBF kernel is a popular choice for non-linear problems, and the polynomial kernel can be used for problem where the relationship between the input features and the target variable is well approximated by polynomial function.

Linear kernel: It is the simplest kernel function. It computes the dot product between the input features, which is equivalent to transforming the data into a higher dimensional space and finding a linear decision boundary in that space.

RBF kernel: It is a popular choice for non-linear problems. It transforms the data into high-dimensional feature space, where a non-linear decision boundary can be found. It computes the Euclidean distance between the input features and a set of basis function, which are centeral at various points in the feature space.

Polynomial kernel: It is the another kernel function that can be used for non-linear problems. It computes dot product between the input features raised to a specified power, which can capture complex, non-linear relationships between the input features and the target variable.