Statistical worksheet.

1. d) expected.
2. c) frequencies.
3. c) 6
4. b) chi-square 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 Assignment

Q. 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) **R-squared** is a better measure of goodness of fit because it provides a proportion of variance explained by the model, with values closer to 1 indicating a better fit. RSS alone doesn't provide context on the total variance; it just indicates the error magnitude.

Q. 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 variation in data.

ESS: Variation explained by the model.

RSS: Variation not explained by the model.

The relationship is **TSS = ESS + RSS**.

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

Ans) Regularization is needed in machine learning to prevent overfitting, ensuring the model generalizes well to unseen data by penalizing complex models.

Q.4) What is Gini–impurity index?

Ans) The Gini impurity index is a metric used in decision trees that measures the probability of a randomly chosen element being incorrectly classified, with lower values indicating better purity.

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

Ans) Yes, unregularized decision-trees are prone to overfitting because they can create overly complex trees that perfectly fit the training data but fail to generalize to new data.

Q.6) What is an ensemble technique in machine learning?

Ans) An ensemble technique in machine learning combines multiple models to improve prediction performance, often leading to better results than any single model alone.

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

Ans) Bagging reduces variance by averaging predictions from multiple models trained on different subsets of data, while Boosting reduces bias by sequentially training models to correct the errors of previous ones.

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

Ans) Out-of-bag error is an estimate of prediction error for random forests, calculated using predictions from trees on data not included in their bootstrap sample.

Q.9) What is K-fold cross-validation?

Ans) K-fold cross-validation is a technique where the dataset is divided into K subsets, and the model is trained and validated K times, each time using a different subset as the validation set and the remaining as the training set.

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

Ans) Hyperparameter tuning in machine learning is the process of finding the optimal set of parameters for a model to improve its performance on unseen data. It's done to enhance the model's ability to generalize.

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

Ans) A large learning rate in Gradient Descent can cause the algorithm to overshoot the minimum, leading to divergence or erratic convergence behavior.

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

Ans) Logistic Regression is a linear model and may not perform well on non-linear data without transformations or kernel tricks to capture non-linearity.

Q.13) Differentiate between Adaboost and Gradient Boosting.

Ans) Adaboost focuses on misclassified points by adjusting weights, while Gradient Boosting optimizes a loss function by adding models that correct previous errors.

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

Ans) The bias-variance tradeoff is the balance between a model's ability to generalize well (low variance) and its accuracy on training data (low bias), with the goal of minimizing overall error.

Q.15) The bias-variance tradeoff is the balance between a model's ability to generalize well (low variance) and its accuracy on training data (low bias), with the goal of minimizing overall error.

Ans) Yes, that's correct. The bias-variance tradeoff involves finding the right balance to minimize the total error in a model