**Assignment – II**

1. **R-squared or Residual Sum of Squares (RSS)**: R-squared (coefficient of determination) is typically preferred as a measure of goodness of fit because it represents the proportion of the variance in the dependent variable that is predictable from the independent variables. RSS, while informative, doesn't account for the scale of the data and is not normalized like R-squared.
2. **TSS (Total Sum of Squares), ESS (Explained Sum of Squares), RSS (Residual Sum of Squares)**:
   * **TSS**: Total Sum of Squares measures the total variance in the dependent variable.
   * **ESS**: Explained Sum of Squares measures the variance explained by the regression model.
   * **RSS**: Residual Sum of Squares measures the unexplained variance. The equation relating them is: TSS = ESS + RSS.
3. **Need of regularization in machine learning**: Regularization helps prevent overfitting by adding a penalty term to the model's loss function, discouraging complex models that fit the training data too closely. It improves generalization on unseen data.
4. **Gini-impurity index**: It is a measure of impurity or disorder used in decision trees for classification tasks. It quantifies how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the set.
5. **Unregularized decision-trees prone to overfitting**: Yes, decision trees without regularization can grow deep and fit the training data very closely, capturing noise and leading to poor performance on unseen data. Regularization techniques like pruning are used to mitigate this.
6. **Ensemble technique in machine learning**: Ensemble techniques combine predictions from multiple machine learning models to improve the accuracy and robustness over a single model. Examples include Random Forests, Gradient Boosting, and Bagging.
7. **Difference between Bagging and Boosting**:
   * **Bagging**: Constructs multiple models in parallel, each trained on a random subset of the data, and aggregates their predictions (e.g., Random Forests).
   * **Boosting**: Builds models sequentially, where each model corrects the errors of its predecessor, focusing more on difficult cases (e.g., AdaBoost, Gradient Boosting).
8. **Out-of-bag error in random forests**: It is the average error for each training sample calculated using predictions from the trees that do not contain the sample in their respective bootstrap sample during training. It serves as an internal validation measure for random forests.
9. **K-fold cross-validation**: It is a technique used to assess the performance of a predictive model by partitioning the data into K subsets (folds) of approximately equal size. The model is trained on K-1 folds and tested on the remaining fold iteratively, and the performance metric is averaged over all folds.
10. **Hyperparameter tuning in machine learning**: It is the process of choosing a set of optimal hyperparameters for a learning algorithm. Hyperparameters are parameters whose values are set before the learning process begins. Tuning is done to improve model performance and generalization.
11. **Issues with large learning rate in Gradient Descent**: With a large learning rate, gradient descent steps can be too large, causing the algorithm to diverge or overshoot the minimum of the loss function. This can prevent convergence and lead to unstable training.
12. **Logistic Regression for Non-Linear Data**: Logistic Regression is a linear classifier and cannot model complex nonlinear decision boundaries effectively. To handle non-linear data, nonlinear transformations of features or more complex models like SVMs with nonlinear kernels are typically used.
13. **Difference between Adaboost and Gradient Boosting**:

* **Adaboost**: Adaboost sequentially fits multiple weak learners to weighted versions of the data, adjusting weights to focus on misclassified samples.
* **Gradient Boosting**: Gradient Boosting builds sequentially by fitting new models to the residual errors of the previous model, optimizing a differentiable loss function.

1. **Bias-variance tradeoff**: It refers to the dilemma in machine learning where increasing model complexity (reducing bias) typically leads to an increase in variance (model's sensitivity to small fluctuations in the training data). Finding a balance is crucial for achieving good generalization.
2. **Linear, RBF, Polynomial kernels used in SVM**:

* **Linear kernel**: K(x,x′)=xTx′K(x, x') = x^T x'K(x,x′)=xTx′ computes the dot product of the input vectors.
* **RBF (Radial Basis Function) kernel**: K(x,x′)=exp⁡(−γ∥x−x′∥2)K(x, x') = \exp(-\gamma \| x - x' \|^2)K(x,x′)=exp(−γ∥x−x′∥2) computes similarity based on the Euclidean distance between the samples, scaled by parameter γ\gammaγ.
* **Polynomial kernel**: K(x,x′)=(xTx′+c)dK(x, x') = (x^T x' + c)^dK(x,x′)=(xTx′+c)d computes similarity based on polynomial terms up to degree ddd, with optional coefficient ccc.

**STATISTICS WORKSHEET**

1. Using a goodness of fit, we can assess whether a set of obtained frequencies differ from a set of frequencies.

**- d) Expected**

2. Chi-square is used to analyze

**- c) Frequencies**

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

**- c) 6**

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

**- b) Chi-squared distribution**

5. Which of the following distributions is Continuous?

**- c) F Distribution**

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

**- b) Hypothesis**

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

- **a) Null Hypothesis**

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

**- a) Two tailed**

9. Alternative Hypothesis is also called as?

**- b) 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

**- a) np**