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** is generally better as it provides a normalized measure of the proportion of variance explained by the model, making it easier to interpret and compare across different 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** is the total variance in the dependent variable.
    - **ESS** is the variance explained by the regression model.
    - **RSS** is the variance not explained by the model (error).
    - **Equation**: TSS=ESS+RSSTSS = ESS + RSSTSS=ESS+RSS.
3. **What is the need for regularization in machine learning?**
    - Regularization helps prevent overfitting by penalizing large coefficients, leading to a simpler model that generalizes better to unseen data.
4. **What is Gini-impurity index?**
    - Gini impurity is a measure of the probability of a randomly chosen element being incorrectly classified if it was randomly labeled according to the distribution of labels in the subset.
5. **Are unregularized decision-trees prone to overfitting? If yes, why?**
    - Yes, because they can create overly complex trees that perfectly fit the training data, capturing noise rather than the underlying pattern.
6. **What is an ensemble technique in machine learning?**
    - An ensemble technique combines multiple models to improve performance, reduce overfitting, and enhance generalization.
7. **What is the difference between Bagging and Boosting techniques?**
    - **Bagging** builds multiple independent models on different subsets of data and averages their predictions.
    - **Boosting** builds models sequentially, each new model correcting errors made by previous ones.
8. **What is out-of-bag error in random forests?**
    - It's an estimate of the prediction error obtained by aggregating the errors for each data point using only the trees that did not include that data point in their bootstrap sample.
9. **What is K-fold cross-validation?**
    - It's a technique where the dataset is divided into K equal parts, and the model is trained K times, each time using a different part as the validation set and the remaining K-1 parts as the training set.
10. **What is hyperparameter tuning in machine learning and why is it done?**
    - Hyperparameter tuning involves selecting the best set of hyperparameters for a model to improve its performance on validation data.

11. **What issues can occur if we have a large learning rate in Gradient Descent?**
    - A large learning rate can cause the model to overshoot the optimal solution, leading to divergence or highly oscillating updates.
12. **Can we use Logistic Regression for classification of Non-Linear Data? If not, why?**
    - Logistic Regression by itself cannot handle non-linear data because it creates a linear decision boundary. However, using feature transformations or kernel methods can allow it to handle non-linear relationships.
13. **Differentiate between Adaboost and Gradient Boosting.**
    - **Adaboost** assigns weights to each instance and focuses on misclassified instances in subsequent models.
    - **Gradient Boosting** builds models sequentially, with each new model trying to reduce the residuals of the previous models using gradient descent.
14. **What is bias-variance trade-off in machine learning?**
    - It is the balance between the error introduced by bias (underfitting) and variance (overfitting). Ideally, we aim to find a model with low bias and low variance.
15. **Give a short description of Linear, RBF, and Polynomial kernels used in SVM.**
    - **Linear Kernel**: Uses the dot product of input features to separate data with a linear decision boundary.
    - **RBF (Radial Basis Function) Kernel**: Maps input features into higher-dimensional space to handle non-linear relationships, focusing on the distance between points.
    - **Polynomial Kernel**: Computes the similarity of input features raised to a specified power, useful for capturing polynomial relationships in the data.