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?**

Answer- R-squared is better because it tells us how much of our data the model explains, from 0 to 1. It's easier to compare models with R-squared, and it considers if the model is too complex. RSS doesn't do these things.

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.**

Answer-

1. **Total Sum of Squares (TSS):** Measures the total variation in y without considering the model. It's like looking at all the differences from the average y value.
2. **Explained Sum of Squares (ESS):** Measures how much of y's variation the model explains. It's like the part of TSS that the model can account for.
3. **Residual Sum of Squares (RSS):** Measures the differences between actual y values and predicted y values. It's like the part of TSS that the model can't explain.
    - The equation relating these is: $TSS=ESS+RSSTSS=ESS+RSS$
    - This shows that the total variation in y (TSS) equals what the model can explain (ESS) plus what it can't explain (RSS).

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

Answer- Regularization helps to reduce the variance of the model, without a substantial increase in the bias. If there is variance in the model that means that the model won't fit well for dataset different than training data. The tuning parameter λ controls this bias and variance tradeoff. When the value of λ is increased up to a certain limit, it reduces the variance without losing any important properties in the data. But after a certain limit, the model will start losing some important properties which will increase the bias in the data. Thus, the selection of good value of λ is the key. The value of λ is selected using cross-validation methods. A set

of λ is selected and cross-validation error is calculated for each value of λ and that value of λ is selected for which the cross-validation error is minimum.

4. **What is Gini–impurity index?**

   Answer-

   - Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset. 'It is calculated by multiplying the probability that a given observation is classified into the correct class and sum of all the probabilities when that particular observation is classified into the wrong class.
   - Ginni impurity value lies between 0 and 1, 0 being no impurity and 1 denoting random distribution. The node for which the Ginni impurity is least is selected as the root node to split.

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

   Answer-

   - Unregularized decision trees are prone to overfitting due to their high sensitivity to training data, potential for creating complex and deep structures, tendency to memorize the data, and lack of generalization to new, unseen data. Regularization techniques like pruning, minimum samples per leaf, and maximum depth constraints are used to address these issues and improve generalization.

6. **What is an ensemble technique in machine learning?**

   Answer-

   - **Bagging:** It creates multiple copies of the same model, each trained on a different subset of the training data (chosen randomly with replacement). Bagging reduces variance and helps prevent overfitting by averaging the predictions of these models.
   - **Boosting:** Boosting involves training a sequence of weak learners (models slightly better than random guessing) where each subsequent model focuses on correcting errors made by the previous ones. Algorithms like AdaBoost, Gradient Boosting, and XGBoost are popular examples of boosting techniques.
   - **Random Forest:** Random Forest combines multiple decision trees, each trained on a random subset of features and data samples. It's effective at handling high-dimensional data, reducing overfitting, and providing feature importance rankings.

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

Answer-

Bagging

1. Training data subsets are drawn randomly with replacement from the entire training dataset.
2. Bagging attempts to tackle the over-fitting issue.
3. Every model receives an equal weight.
4. Objective to decrease variance, not bias.
5. Every model is built independently.

Boosting

1. Each new subset contains the components that were misclassified by previous models.
2. Boosting tries to reduce bias.
3. Models are weighted by their performance.
4. Objective to decrease bias, not variance.
5. New models are affected by the performance of the previously developed model.

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

Answer- The out-of-bag (OOB) error in random forests is an estimate of how well the model will perform on new, unseen data. It's calculated using the data points that were not included in the training of each tree. This helps evaluate the model's accuracy without needing a separate validation set.

9. **What is K-fold cross-validation?**

Answer-

- K-fold cross-validation divides data into K parts, trains the model K times using K-1 parts for training and 1 part for validation, then averages the results for a more reliable performance estimate.
- **Data Splitting:** Divide data into K equal parts.
- **Training and Validation:** Train model K times, using K-1 parts for training and 1 part for validation each time.
- **Performance Evaluation:** Assess model performance on each validation set.
- **Average Score:** Average the performance scores to get a reliable estimate of model performance.

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

Answer-

- Hyperparameter tuning in machine learning involves adjusting the settings or configurations of a model that are set before the training process begins. These settings are not learned from the data but are specified by the user.
- Hyperparameters can include the learning rate in neural networks, the depth of decision trees, the number of clusters in clustering algorithms, and the regularization strength in regression models, among others.
- It is done to optimize the model's performance, improve accuracy, and find the best configuration for generalization to new data. Hyperparameter tuning helps in finding the right balance between model complexity and performance, leading to better results and more robust models.

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

Answer-

- **Overshooting:** The algorithm might skip over the minimum point, leading to slower convergence or oscillations.
- **Divergence:** It can cause the algorithm to move away from the optimal solution, resulting in inaccurate results.
- **Instability:** Large learning rates can make the optimization process unstable, with parameters bouncing around and not settling.
- **Longer Training Time:** Due to issues like oscillations and instability, convergence may take longer or may not occur at all, increasing training time and computational costs.

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

Answer- No, Logistic Regression is not suitable for classifying non-linear data because it assumes linear relationships between features and the target variable. It cannot capture the complexities of non-linear relationships.

13. Differentiate between Adaboost and Gradient Boosting.

Answer-

- **AdaBoost (Adaptive Boosting):**

1. Sequential training of weak learners.
2. Adjusts weights of misclassified data points to focus on difficult instances.
3. Combines weak learners using weighted majority voting.
4. Prone to noise and outliers.

- **Gradient Boosting:**

5. Sequential training of decision trees (typically).
6. Fits each new tree to the residual errors of the previous ones.

7. Combines trees using a weighted sum approach.
8. Can handle noisy data and outliers better due to its iterative nature.

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

Answer-

Bias-

- Bias represents the error introduced by a model's assumptions and simplifications about the data.
- A model with high bias tends to be too simplistic, making it unable to capture the true underlying patterns in the data. It often leads to underfitting, meaning the model performs poorly both on the training data and new, unseen data because it oversimplifies the problem.

Variance:

- Variance, on the other hand, refers to the model's sensitivity to the fluctuations or noise in the training data. A model with high variance is too complex and flexible, essentially "memorizing" the noise in the training data rather than generalizing well to new, unseen data.
- High variance leads to overfitting, where the model performs excellently on the training data but poorly on new data because it hasn't learned the true patterns but instead has adapted to the specific examples in the training set.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Answer-

1. **Linear Kernel:**

- Creates a linear decision boundary.
- Suitable for linearly separable data.
- Computes inner products between data points.

2. **RBF (Radial Basis Function) Kernel:**

- Uses a non-linear decision boundary.
- Maps data into high-dimensional space using a Gaussian function.
- Suitable for non-linearly separable data.

3. **Polynomial Kernel:**

- Creates a decision boundary using polynomial functions.
- Can capture more complex relationships than linear kernels.
- Has a parameter (degree) that determines the polynomial's order.