MACHINE LEARNING Assignment 5

Answers:

1. R-squared is a better measure of how well a regression model fits the data because it gives a single number that tells us how much of the variation in the data is explained by the model. On the other hand, Residual Sum of Squares (RSS) is just a calculation of how far off the model predictions are from the actual data, but it's not as easy to understand. R-squared is usually preferred because it provides a clearer picture of the model's performance.
2. TSS (Total Sum of Squares) is the total variability in the response variable, which is calculated as the sum of the squared differences between the observed values and the average of all the values.
   ESS (Explained Sum of Squares) is the amount of variability in the response variable that is explained by the regression model, calculated as the sum of the squared differences between the predicted values from the model and the average of all the values.
   RSS (Residual Sum of Squares) is the amount of variability in the response variable that is not explained by the regression model, calculated as the sum of the squared differences between the observed values and the predicted values from the model.
   These three metrics are related as follows:
   TSS = ESS + RSS
3. Regularization is a technique used in machine learning to prevent a model from overfitting the data. Overfitting occurs when a model is too complex and fits the training data too well, but performs poorly on new, unseen data.Regularization adds a penalty term to the model's loss function to discourage overly complex models and encourage the model to be more simple and generalizable. This helps balance the model's ability to fit the training data and its ability to perform well on new data.
4. Gini impurity is a measure used in decision tree algorithms to evaluate the quality of a split in the data. It is a number between 0 and 1 that indicates how mixed the class labels are in a set of data. A value of 0 means all the data belongs to the same class, while a value of 1 means the class labels are evenly distributed.
   In decision tree models, the goal is to make splits in the data that result in sets with low Gini impurity, meaning the class labels are mostly the same in each set. This helps to make accurate predictions for new, unseen data.
5. Yes, unregularized decision trees are prone to overfitting.
   Because they fit the training data too well and perform poorly on new data. This happens because unregularized decision trees can keep splitting the data until they fit the training data perfectly, resulting in a complex model that is not generalizable.
   To avoid overfitting, regularization techniques, such as pruning or limiting the depth of the tree, can be used.
6. Ensemble technique in machine learning is a method of combining multiple models to produce a better prediction result. The idea behind ensemble is that the combination of multiple models can lead to a better prediction than any individual model, as the weaknesses of one model can be compensated by the strengths of another.

7. Bagging and Boosting are both methods used to improve the accuracy of a machine learning model.

Bagging creates multiple copies of the same model and combines their results to reduce overfitting. This approach results in a more stable prediction.

Boosting creates a series of models, each correcting the mistakes of the previous model, to improve overall accuracy. This approach results in a more accurate prediction.

Bagging is good for improving stability, while Boosting is good for improving accuracy.

8. The out-of-bag error in random forests is a measure of how well the model performs on unseen data. It is calculated by using a portion of the training data that was not used to build the individual trees in the random forest. The prediction error of each tree is calculated using these unused data points, and the average error is used as an estimate of the overall accuracy of the model. The out-of-bag error is a convenient way to evaluate the model without the need for additional validation data.
9. K-fold cross-validation is a method used to test the accuracy of a machine learning model. It involves splitting the training data into k parts, using k-1 parts for training and 1 part for testing. This process is repeated k times, with each part used for testing once. The average accuracy from all k iterations is used as the final evaluation of the model. K-fold cross-validation helps to avoid overfitting by using all of the training data for both training and testing.
10. Hyper parameter tuning is the process of finding the best settings for a machine learning model. These settings, called hyper parameters, control how the model behaves and affect its accuracy. By finding the best hyper parameters, the model can perform better and avoid overfitting. The process of hyper parameter tuning can be time-consuming and requires training the model multiple times with different hyper parameter combinations.

    Examples of hyperparameters include the learning rate for gradient descent, the number of trees in a random forest, or the number of hidden layers in a neural network.

11. A large learning rate in Gradient Descent can cause several issues during the training of a machine learning model. The following are some of the problems that can occur:

    i. **Oscillation**: Yhe model bounces back and forth instead of gradually approaching the optimal solution. This happens when the learning rate is set too high and can result in slow convergence and the model never reaching the best solution.
    ii. **Overstepping the optimum**: A large learning rate can cause the model to overstep the optimum and end up on the other side, resulting in slower convergence and a suboptimal solution.
    iii. **Divergence**: In some cases, a large learning rate can cause the model to diverge and never converge to the optimum.
    iv. **Instability**: A large learning rate can also cause the model to be unstable and jump around randomly, leading to slow convergence and poor performance.

12. No, Logistic Regression is not suitable for classification of non-linear data. Logistic Regression is a linear model and assumes a straight-line relationship between the

input features and target variable. If the relationship between the input features and target variable is non-linear, a straight line will not be able to capture the underlying relationship, leading to a poor fit. In these cases, other machine learning models such as decision trees or support vector machines are better suited for the task as they can handle non-linear relationships.

13. AdaBoost gives more importance to the data points that were classified incorrectly in the previous models. It combines many simple models to make a better overall model.

    Gradient Boosting is a more complex method. It tries to fix the mistakes of the previous models by fitting new models in each iteration. It can handle more complex data and is more flexible.

    Both methods are used to improve model accuracy, but Gradient Boosting is usually more powerful and flexible.

14. Bias-variance tradeoff is about finding the right balance between a model being too simple or too complex. A model that is too simple (high bias) will not fit the data well, and a model that is too complex (high variance) will fit the data too closely and not generalize well to new data. The goal is to find a model that strikes the right balance, giving good results on both the training data and new data.
15. In SVM, a kernel is a way to transform the data so a good decision boundary can be made. There are three types of kernels:

    i. Linear kernel: This is the simplest kernel and works well when the data is already split into two clear groups.
    ii. RBF (Radial basis function) kernel: This is used when the data is not easily split into two groups. It measures how similar the data points are to each other.
    iii. Polynomial kernel: This transforms the data into a higher dimension and can create a non-linear decision boundary. It is used when the data cannot be split into two groups in a straight line.