ASSIGNMENT-5

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?**

Residual Sum of Squares(RSS):

- The residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term.
- The smaller the residual sum of squares, the better your model fits your data; the greater the residual sum of squares, the poorer your model fits your data.
- A value of zero means your model is a perfect fit.
- RSS in short will determines how well the model explains or represents the data.

R-squared:

- A statistical measure that determines the proportion of variance in the dependent variable that can be explained by the independent variable.
- How well the data fit the regression model is explained by R-squared.
- Formula : R-squared = Sum of squares due to regression / Total sum of squares.

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 (Total Sum of Squares): TSS tells us how much variation is there in the dependent variable.

$TSS = \Sigma(Yi - \text{mean of } Y)^2$

ESS (Explained Sum of Squares):

ESS tells us how much of the variation in the dependent variable our model explained.

ESS = Σ(Y-Hat – mean of Y)^2

<u>RSS (Residual Sum of Squares)</u>:

The residual sum of squares tells us how much of the dependent variable's variation our model did not explain. It is the sum of the squared differences between the actual Y and the predicted Y:

RSS = Σ e2

Equation relating these three metrics is given by :

TSS = ESS + RSS

where ,

TSS = Total Sum of Squares

ESS = Explained Sum of Squares

RSS = Residual Sum of Squares

3. **What is the need of regularization in machine learning?**
   While training the machine learning model the model can easily be overfitted and underfitted. To remove this, we use regularization in machine learning to properly fit the model onto our test set. We can say that regularization technique helps to reduce the chances of overfitting and underfitting and helps to get the optimal model.

4. **What is Gini–impurity index?**
   Gini impurity is a function that determines how well a decision tree was split. Basically, it helps us to determine which splitter is best so that we can build a pure decision tree. Gini impurity ranges values from 0 to 0.5. It is one of the methods of selecting the best splitter.

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

Yes, unregularized decision trees are prone to overfitting. Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.

But unlike other algorithms decision tree does not use regularization to fight against overfitting. Instead it uses pruning. There are mainly to types of pruning performed:

- Pre-pruning that stop growing the tree earlier, before it perfectly classifies the training set.
- Post-pruning that allows the tree to perfectly classify the training set, and then post prune the tree.

6. What is an ensemble technique in machine learning?

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. Voting and averaging are two of the easiest ensemble methods. They are both easy to understand and implement. Voting is used for classification and averaging is used for regression.

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

	Bagging	**Boosting**
1)	The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types
2)	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3)	Each model receives equal weight.	Models are weighted according to their performance.

4)	Each model is built independently.	New models are influenced by the performance of previously built models.
5)	Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.	Every new subset contains the that were misclassified by previous models.
6)	Bagging tries to solve the over-fitting problem.	Boosting tries to reduce bias.
7)	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) then apply boosting.
8)	In this base classifiers are trained parallelly.	In this base classifiers are trained sequentially.
9)	Example: The Random forest model uses Bagging.	Example: The AdaBoost uses Boosting techniques

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

Out-of-bag error is one of these methods for validating the machine learning model.
This approach utilizes the usage of bootstrapping in the random forest. Since the bootstrapping samples the data with the possibility of selecting one sample multiple times, it is very likely that we won't select all the samples from the original data set. Therefore, one smart decision would be to exploit somehow these unselected samples, called out-of-bag samples.

Correspondingly, the error achieved on these samples is called out-of-bag error. What we can do is to use out-of-bag samples for each decision tree to measure its performance. This strategy provides reliable results in comparison to other validation techniques such as train-test split or cross-validation.

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

   K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data. K refers to the number of groups the data sample is split into. For example, if you see that the k-value is 5, we can call this a 5-fold cross-validation. Each fold is used as a testing set at one point in the process.

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

    Hyperparameter tuning is basically referred to as tweaking the parameters of the model, which is basically a prolonged process. Hyper-parameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyper-parameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning. Hyper-parameter tuning is done to increase the accuracy of the model. Hyperparameter tuning maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

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

    If the learning rate is too large in Gradient Descent, it can cause the model to converge too quickly to a suboptimal solution.

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

    Non-linear problems can't be solved with logistic regression because it has a linear decision surface. The decision boundary is a line or a plane that separates the target variables into different classes that can be either linear or nonlinear. In the case of a Logistic Regression model, the decision boundary is a straight line.

13. **Differentiate between Adaboost and Gradient Boosting.**

Adaboost	Gradient Boosting
1) An additive model where shortcomings of previous models are identified by high-weight data points.	An additive model where shortcomings of previous models are identified by the gradient.

2)	The trees are usually grown as decision stumps.	The trees are grown to a greater depth usually ranging from 8 to 32 terminal nodes.
3)	Each classifier has different weights assigned to the final prediction based on its performance.	All classifiers are weighed equally and their predictive capacity is restricted with learning rate to increase accuracy.
4)	It gives weights to both classifiers and observations thus capturing maximum variance within data.	It builds trees on previous classifier's residuals thus capturing variance in data.

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

- If the algorithm is too simple (hypothesis with linear eq.) then it may be on high bias and low variance condition and thus is error-prone.
- If algorithms fit too complex (hypothesis with high degree eg:.) then it may be on high variance and low bias.
- In the latter condition, the new entries will not perform well.
- Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off.
- This trade-off in complexity is why there is a trade-off between bias and variance.
- An algorithm can't be more complex and less complex at the same time.

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

Polynomial Kernel: It represents the similarity of vectors in the training set of data in a feature space over polynomials of the original variables used in the kernel.

Linear Kernel: Linear Kernel is used when the data is linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Data Set. One of the examples where there are a lot of features, is Text Classification, as each alphabet is a new feature. So we mostly use Linear Kernel in Text Classification.

RBF: RBF is the default kernel used within the sklearn's SVM classification algorithm and can be described with the following formula: where gamma can be set manually and has to be >0.