MACHINE LEARNING

ASSIGNMENT - 5

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?

Ans:- R-squared is a better measure of goodness of fit model in regression .

R- Squared :- R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

R-squared measures the strength of the relationship between your linear model and the dependent variables on a **0 - 100% scale**

Residual sum of square (RSS) The residual sum of squares (RSS) measures the level of variance in the error term, or residuals, of a regression model. 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.

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 ...

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.

Ans:- **Total sum of squares =** TSS represents the total sum of squares. It is the squared values of the dependent variable to the sample mean. In other words, the total sum of squares measures the variation in a sample.

The sum of squares total, denoted SST, is the squared differences between the observed dependent variable and its mean. You can think of this as the dispersion of the observed variables around the mean – much like the variance in descriptive statistics.

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Explained Sum of Squares = The explained sum of squares (ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model

Residual Sum of Squares = 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.

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

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

Ans:- Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

In the context of machine learning, regularization is the process which regularizes or shrinks the coefficients towards zero. In simple words, regularization discourages learning a more complex or flexible model, to prevent overfitting.

4. What is Gini–impurity index?

Ans:- Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree.

Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure.

The Gini Coefficient or Gini Index measures the inequality among values of a variable. Higher the value of an index, more dispersed is the data. Alternatively, the Gini coefficient can be looked like half of the relative mean absolute difference. Gini index is the most commonly used measure of inequality.

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

Ans:- 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.

Overfitting refers to the condition when the model completely fits the training data but fails to generalize the testing unseen data. Overfit condition arises when the model memorizes the noise of the training data and fails to capture important patterns

6. What is an ensemble technique in machine learning?

Ans:- Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. This has boosted the popularity of ensemble methods in machine learning

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

Ans:- **Bagging** :- Bagging stands for Bootstrap aggregating, which combines several models for better predictive results. In statistical classification and regression, bagging improves the stability and accuracy of machine learning algorithms by decreasing the variance and reducing the chances of overfitting.

Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of the original data. Bagging will mainly focus at getting an ensemble model with less variance than its components. Bagging algorithms that aim to reduce the complexity of models that overfit the training data.

Boosting :- Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification. Boosting and stacking will mainly try to produce strong models less biased than their components (even if variance can also be reduced). Boosting is an approach to increase the complexity of models that suffer from high bias, that is, models that underfit the training data.

Difference between bagging and boosting

Bagging > **1.** The original dataset is divided into multiple subsets, selecting observations with replacement.
2. This method combines predictions that belong to the same type.
3. Bagging decreases variance.
4. Base classifiers are trained parallelly.
5. The models are created independently.

Boosting > **1.** The new subset contains the components mistrained by the previous model.
2. This method combines predictions that belong to the different types.
3. Boosting decreases bias.
4. Base classifiers are trained sequentially.
5. The model creation is dependent on the previous ones.

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

Ans:- The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained

Every data point is passed for prediction to trees where it would be behaving as OOB and an aggregated prediction is recorded for each row. The OOB_score is computed as the number of correctly predicted rows from the out-of-bag sample. OOB Error is the number of wrongly classifying the OOB Sample

9. What is K-fold cross-validation?

Ans:- 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.

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

Ans:- Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters 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?

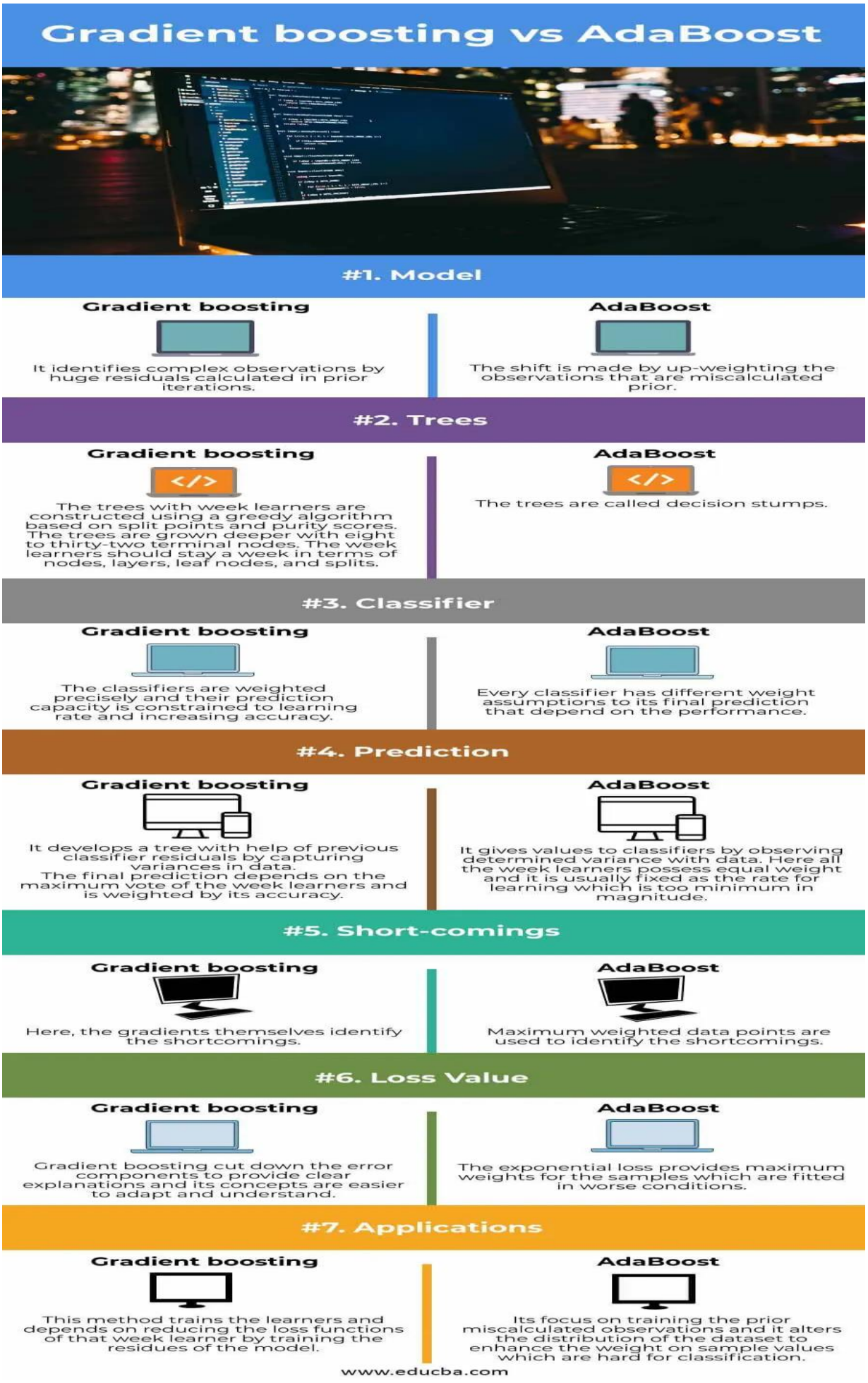
Ans:- If learning rate is too large, gradient descent can overshoot the minimum. It may fail to converge and even diverge.

In order for Gradient Descent to work, we must set the learning rate to an appropriate value. This parameter determines how fast or slow we will move towards the optimal weights. If the learning rate is very large we will skip the optimal solution.

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

Ans:- It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set. It is very fast at classifying unknown records. Non-linear problems can't be solved with logistic regression because it has a linear decision surface

13. Differentiate between Adaboost and Gradient Boosting.



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

Ans:- In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

A good model is the one which neither have high variance nor high bias but as variance and bias are inversely propotional to each other ,so to make both minimum possible there must be a tradeoff(or a balance) between them .This is what we call **bias-variance tradeoff** .

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

Ans:- **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.

Rbf kernel : RBF short for Radial Basis Function Kernel is a very powerful kernel used in SVM. Unlike linear or polynomial kernels, RBF is more complex and efficient at the same time that it can combine multiple polynomial kernels multiple times of different degrees to project the non-linearly separable data into higher dimensional space so that it can be separable using a hyperplane.

Poly kernel : In a polynomial kernel for SVM, the data is mapped into a higher-dimensional space using a polynomial function. The dot product of the data points in the original space and the polynomial function in the new space is then taken. The polynomial kernel is often used in SVM classification problems where the data is not linearly separable. By mapping the data into a higher-dimensional space, the polynomial kernel can sometimes find a hyperplane that separates the classes.