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?

A1. The residual sum of squares (RSS) is the absolute amount of explained variation, whereas R-squared is the absolute amount of variation as a proportion of total variation.

RSS is 'scale-variant' i.e., it's value changes with the scale of the variable while R-squared is 'scale-invariant' i.e., values will not change with the change in scale.

Hence R-squared is a better measure of goodness of fit model in 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.

A2. TSS: Total Sum of Squares :- The coefficient of determination is used as a measure of how well a regression line explains the relationship between a dependent variable (Y) and an independent variable (X).

RSS: Residual Sum of Squares :- It is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model.

ESS: Explained Sum of Squares:- It is a statistical technique used to measure the amount of variance in a data set that is explained by a regression model.

TSS = ESS +RSS

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

A3. Overfitting is a phenomenon that occurs when a Machine Learning model is constraint to training set and not able to perform well on unseen data.

Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
The commonly used regularization techniques are :

1. L1 regularization: LASSO

2. L2 regularization: RIDGE

4.What is Gini–impurity index?

A4. 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

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

A5. Unregularized decision trees can learn a training set to a point of high granularity that makes them easily overfit.

6. What is an ensemble technique in machine learning?

A6. Ensemble techniques combine the decisions from multiple models to improve the overall performance.

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

A7.

S.NO	**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 elements 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) the apply boosting.**

S.NO	Bagging	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?

A8. The out-of-bag error is the average error for each predicted outcome calculated using predictions from the trees that do not contain that data point in their respective bootstrap sample. This way, the Random Forest model is constantly being validated while being trained.

9. What is K-fold cross-validation?

A9. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation

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

A10. Hyper parameter tuning is a procedure of changing values of some paramaters in order to increase the accuracy of the mode.

Eg: GridSearchCV

Randomised CV

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

A11. The learning rate can seen as step size, n. As such, gradient descent is taking successive steps in the direction of the minimum. If the step size n is too large, it can (plausibly) "jump over" the minima we are trying to reach, ie. we overshoot. This can lead to osculations around the minimum or in some cases to outright divergence.

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

A12. We cannot solve non-linear problems with logistic regression. All non-linear features must be transformed before they can be used with logistic regression model.

13. Differentiate between Adaboost and Gradient Boosting.

A13

Features	Gradient boosting	Adaboost
Model	It identifies complex observations by huge residuals calculated in prior iterations	The shift is made by up-weighting the observations that are miscalculated prior
Trees	The trees with week learners are constructed using a greedy algorithm based on split points and purity scores. The trees are grown deeper with eight to thirty-two terminal nodes. The week learners should stay a week in terms of nodes, layers, leaf nodes, and splits	The trees are called decision stumps.
Classifier	The classifiers are weighted precisely and their prediction capacity is constrained to learning rate and increasing accuracy	Every classifier has different weight assumptions to its final prediction that depend on the performance.
Prediction	It develops a tree with help of previous classifier residuals by capturing variances in data.	It gives values to classifiers by observing determined variance with data. Here all the week learners possess equal weight and it is usually fixed as the rate for learning which is too minimum in magnitude.

	The final prediction depends on the maximum vote of the week learners and is weighted by its accuracy.	
Shortcomings	Here, the gradients themselves identify the shortcomings.	Maximum weighted data points are used to identify the shortcomings.
Loss value	Gradient boosting cut down the error components to provide clear explanations and its concepts are easier to adapt and understand	The exponential loss provides maximum weights for the samples which are fitted in worse conditions.

Applications	This method trains the learners and depends on reducing the loss functions of that week learner by training the residues of the model	Its focus on training the prior miscalculated observations and it alters the distribution of the dataset to enhance the weight on sample values which are hard for classification

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

A14. 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 eq.) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well

The need to find a good balance between the bias and variance of the model we have used. This trade-off in complexity is what is referred to as bias and variance trade-off. An optimal balance of bias and variance should never overfit or underfit the model

This trad-eoff applies to all forms of supervised learning: classification, regression, and structured output learning

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

A15. 1.RBF kernel: The RBF kernel is the most widely used kernel concept to solve the problem of classifying datasets that cannot be separated linearly. This kernel is known to have good performance with certain parameters, and the results of the training have a small error value compared to other kernels. The equation formula for the RBF kernel function is:

```
K(x,xi) = exp(-gamma * sum((x – xi^2))
```

2.Ploynomial kernal**:** A Polynomial Kernel is more generalized form of the linear kernel. In machine learning, the polynomial kernel is a kernel function suitable for use in support vector machines (SVM) and other kernelizations, where the kernel represents the similarity of the training sample vectors in a feature space. Polynomial kernels are also suitable for solving classification problems on normalized training datasets. The equation for the polynomial kernel function is:

K(x,xi) = 1 + sum(x * xi)^d

This kernel is used when data cannot be separated linearly.

The polynomial kernel has a degree parameter (d) which functions to find the optimal value in each dataset. The d parameter is the degree of the polynomial kernel function with a default value of d = 2. The greater the d value, the resulting system accuracy will be fluctuating and less stable. This happens because the higher the d parameter value, the more curved the resulting hyperplane line

3.Linear Kernel: A linear kernel can be used as normal dot product any two given observations. The equation for the kernel function is:

K(x, xi) = sum(x * xi)