MACHINE LEARNING

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?

A-Both R-squared and Residual Sum of Squares (RSS) are measures of goodness of fit in regression analysis, but they capture different aspects of the model's performance.

R-squared (also known as the coefficient of determination) measures the proportion of variation in the dependent variable that is explained by the independent variables in the model. In other words, it indicates how well the model fits the data, with values ranging from 0 to 1. Higher R-squared values indicate a better fit, as they mean that a larger proportion of the variation in the dependent variable is explained by the independent variables in the model.

On the other hand, RSS measures the total sum of squared differences between the actual values of the dependent variable and the predicted values by the model. It represents the amount of unexplained variation in the data, and lower RSS values indicate a better fit, as they mean that the model is able to explain more of the variation in the data.

Therefore, both measures are useful in evaluating the goodness of fit of a model, but they serve different purposes. R-squared is a useful measure to assess the overall fit of the model and to compare different models, while RSS is useful to identify the degree of the error in the model's predictions.

In general, a good model should have both a high R-squared value and a low RSS value, indicating that it explains a large proportion of the variation in the dependent variable and has a low degree of error in its predictions. However, in some cases, one measure may be more important than the other, depending on the research question and the nature of the data being analyzed.

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.

A-The Total SS (TSS or SST) tells you how much variation there is in the dependent variable.
Total SS = $\Sigma(Yi - \text{mean of } Y)^2$.
The Explained SS tells you how much of the variation in the dependent variable your model explained.
Explained SS = $\Sigma(Y\text{-Hat} - \text{mean of } Y)^2$.
The residual sum of squares tells you how much of the dependent variable's variation your model **did not explain**. It is the sum of the squared differences between the actual Y and the predicted Y:
Residual Sum of Squares = Σ e2

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

A-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.

4.What is Gini–impurity index?

A-Gini Index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

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

A-How prone they are to overfitting depends on your stopping rule and your pruning rule.

But, in general, they are prone to this because they are *very* data intensive - that is, they examine the data in a lot of ways. At each node, they look at every possible split of every independent variable (sometimes they impose a rule of monotonicity - if the variable is continuous or ordinal).

Even with a relatively small number of variables, that can be a *lot* of things to examine, especially if one of them is a categorical variable with more than a few levels. For instance, suppose you have a data set on voting in the USA. Suppose one variable is state - with 50 levels. There are $250 \approx 1.1 * 1015$ ways to split the data. That's more than all the people on Earth, much less voters in the USA, much less whatever sample you have.

6.What is an ensemble technique in machine learning?

A-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.

Main Types of Ensemble Methods

Bagging

Bagging, the short form for bootstrap aggregating, is mainly applied in classification and regression. It increases the accuracy of models through decision trees, which reduces variance to a large extent. The reduction of variance increases accuracy, eliminating overfitting, which is a challenge to many predictive models.

Boosting

Boosting is an ensemble technique that learns from previous predictor mistakes to make better predictions in the future. The technique combines several weak base learners to form one strong learner, thus significantly improving the predictability of models. Boosting works by arranging weak learners in a sequence, such that weak learners learn from the next learner in the sequence to create better predictive models.

Stacking

Stacking, another ensemble method, is often referred to as stacked generalization. This technique works by allowing a training algorithm to ensemble several other similar learning algorithm predictions. Stacking has been successfully implemented in regression, density estimations, distance learning, and classifications. It can also be used to measure the error rate involved during bagging.

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

A-Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance

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

A-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 [1].

9.What is K-fold cross-validation?

A- Cross validation is an evaluation method used in machine learning to find out how well your machine learning model can predict the outcome of unseen data. It is a method that is easy to comprehend, works well for a limited data sample and also offers an evaluation that is less biased, making it a popular choice.The data sample is split into 'k' number of smaller samples, hence the name: K-fold Cross Validation. You may also hear terms like four fold cross validation, or ten fold cross validation, which essentially means that the sample data is being split into four or ten smaller samples respectively.

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

A-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.When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error.When the learning rate is too small, training is not only slower, but may become permanently stuck with a high training error.

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

A-When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error. When the learning rate is too small, training is not only slower, but may become permanently stuck with a high training error

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

A-Logistic regression is known and used as a linear classifier. It is used to come up with a hyper*plane* in feature space to separate observations that belong to a class from all the other observations that do *not* belong to that class. The decision boundary is thus *linear*. Robust and efficient implementations are readily available (e.g. scikit-learn) to use logistic regression as a linear classifier.

While logistic regression makes core assumptions about the observations such as IID (each observation is independent of the others and they all have an identical probability distribution), the use of a linear decision boundary is *not* one of them. The linear decision boundary is used for reasons of simplicity following the Zen mantra – when in doubt simplify. In those cases where we suspect the decision boundary to be nonlinear, it may make sense to formulate logistic regression with a nonlinear model and evaluate how much better we can do.

13. Differentiate between Adaboost and Gradient Boosting.

A-

AdaBoost

AdaBoost or Adaptive Boosting is the first Boosting ensemble model. The method automatically adjusts its parameters to the data based on the actual performance in the current iteration. Meaning, both the weights for re-weighting the data and the weights for the final aggregation are re-computed iteratively. In practice, this boosting technique is used with simple classification trees or stumps as base-learners, which resulted in improved performance compared to the classification by one tree or other single base-learner.

Gradient Boosting

Gradient Boost is a robust machine learning algorithm made up of Gradient descent and Boosting. The word 'gradient' implies that you can have two or more derivatives of the same function. Gradient Boosting has three main components: additive model, loss function and a weak learner. The technique yields a direct interpretation of boosting methods from the perspective of numerical optimisation in a function space and generalises them by allowing optimisation of an arbitrary loss function.

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

A-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.

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

- **Linear Kernel:** used when data is linearly separable.

- **Radial Basis Function Kernel (RBF):** The similarity between two points in the transformed feature space is an exponentially decaying function of the distance between the vectors and the original input space as shown below. RBF is the default kernel used in SVM.

- **Polynomial Kernel:** The Polynomial kernel takes an additional parameter, 'degree' that controls the model's complexity and computational cost of the transformation