MACHINE LEARNING – 5

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. Residual Sum of Squares (RSS) is a better measure of goodness of fit model in regression than R-squared because generally, a higher r-squared indicates more variability is explained by the model. However, it is not always the case that a high r-squared is good for the regression model. The quality of the statistical measure depends on many factors, such as the nature of the variables employed in the model, the units of measure of the variables, and the applied data transformation. Thus, sometimes, a high r-squared can indicate the problems with the regression model. But the residual sum of squares (RSS) measures the level of variance in the error term, or residuals, of a regression model and 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. RSS value of zero means your model is a perfect fit.

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. The residual sum of squares (RSS) or sum of squared estimate of errors (SSE) is the sum of the squares of residuals (deviations predicted from actual empirical values of data). It is a measure of the discrepancy between the data and an estimation model, such as a linear regression. A small RSS indicates a tight fit of the model to the data.
   The explained sum of squares (ESS) or sum of squares due to regression (SSR) is a quantity used in describing how well a model, often a regression model, represents the data being modelled. It tells how much of the variation between observed data and predicted data is being explained by the model proposed.
   Total Sum of Squares (TSS or SST) is defined as the sum of overall observations, of the squared differences of each observation from the overall mean.
   Total SS = Explained SS (ESS) + Residual Sum of Squares (RSS)

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. Regularizations are used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting/underfitting.
   A scenario where the machine learning model tries to learn from the details along with the noise in the data and tries to fit each data point on the curve is called Overfitting.
   A scenario where a machine learning model can neither learn the relationship between variables in the testing data nor predict or classify a new data point is called Underfitting.
   Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.
   There are two main types of regularization techniques: Ridge Regularization and Lasso Regularization.

4. What is Gini–impurity index?
   Ans. Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. Gini index varies between values 0 and 1, where 0 expresses the purity of classification, i.e. All the elements belong to a specified class or only one class exists there. And 1 indicates the random distribution of elements across various classes. The value of 0.5 of the Gini Index shows an equal distribution of elements over some classes. Gini Impurity is 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?
   Ans. Unregularized decision-trees are prone to overfitting 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. In other words, decision-tree will make the decision among a subset of all the features(columns), so when it reaches a final decision, it is a complicated and long decision chain. Only if a data point satisfies all the rules along this chain, the final decision can be made. This kind of specific rules on training dataset make it very specific for the training set, on the other hand, cannot generalize well for new data points that it has never seen. Especially when your dataset has many features (high dimension), it tends to overfit more.

6. What is an ensemble technique in machine learning?
   Ans. 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.
   Two types of Ensemble Methods are: -
   i. Bagging
   ii. Boosting

7. What is the difference between Bagging and Boosting techniques?
   Ans. Bagging: It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average. It aims to decrease variance, not bias.
   Boosting: It is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm. It aims to decrease bias, not variance.

8. What is out-of-bag error in random forests?
   Ans. The Random Forest Classifier is trained using bootstrap aggregation, where each new tree is fit from a bootstrap sample of the training observations. 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 Random Forest Classifier to be fit and validated whilst being trained.

9. What is K-fold cross-validation?
   Ans. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. 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. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.
   Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.
   It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

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. In order to 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. The algorithm will take too big of steps and continuously miss the optima. Because of this effect, a learning rate that is too large takes longer to train, because it is continually overshooting its objective and "unlearning" what it has learned, thus requiring expensive backtracking, or causing unproductive oscillations.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
    Ans. No, Logistic Regression cannot be used for classification of Non-Linear Data because 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. It is suitable in cases where a straight line can separate the different classes. However, in cases where a straight line does not suffice then nonlinear algorithms are used to achieve better results.

13. Differentiate between Adaboost and Gradient Boosting.
    Ans. Adaboost increases the performance of all the available machine learning algorithms, and it is used to deal with weak learners. It gains accuracy just above the arbitrary chances of classifying the problem. The adaptable and most used algorithm in AdaBoost is decision trees with a single level. The adaptive boosting method minimizes the exponential loss function which changes the algorithm more profound to its outliers.

    The gradient boosting depends on the intuition, which is the next suitable possible model, when get combined with prior models that minimize the cumulative predicted errors. The crucial idea of gradient boosting is to fix the targeted outcomes for the next model to reduce the error. In gradient boosting, the differentiable loss function makes more sensitive to outliers when compared to AdaBoost.

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. The bias–variance dilemma is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set.
    The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
    The variance is an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting).

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.
    Ans. Linear Kernel: It is the most basic type of kernel, usually one dimensional in nature. It proves to be the best function when there are lots of features. The linear kernel is mostly preferred for text-classification problems as most of these kinds of classification problems can be linearly separated. Linear kernel functions are faster than other functions.
    Polynomial kernel: in the polynomial kernel, we simply calculate the dot product by increasing the power of the kernel. It is a more generalized representation of the linear kernel. It is not as preferred as other kernel functions as it is less efficient and accurate.
    Radial basis function kernel (RBF): Gaussian RBF (Radial Basis Function). RBF kernel is a function whose value depends on the distance from the origin or from some point. The value of gamma varies from 0 to 1. You must manually provide the value of gamma in the code. The most preferred value for gamma is 0.1.