MACHINE LEARNING

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

Linear : 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.

Advantages of using Linear Kernel:

Training a SVM with a Linear Kernel is Fasterthan with any other Kernel.

When training a SVM with a Linear Kernel, only the optimisation of the C Regularisation parameter is required. On the other hand, when training with other kernels, there is a need to optimise the Y parameter which means that performing a grid search will usually take more time.

RBF : Gaussian RBF(Radial Basis Function) is another popular Kernel method used in SVM models for more. RBF kernel is a function whose value depends on the distance from the origin or from some point. Gaussian Kernel is of the following format;

||X1 — X2 || = Euclidean distance between X1 & X2

Polynomial : In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

2) R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit of model in regression and why?

R Squared is used to determine the strength of correlation between the predictors and the target. In simple terms it lets us know how good a regression model is when compared to the average. R Squared is the ratio between the residual sum of squares and the total sum of squares.

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

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variable explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

3) 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.

The Total SS (TSS or SST) tells us how much variation there is in the dependent variable .

Total SS = Σ(Yi – mean of Y)2

Sum of squares is a measure of how a data set varies around a central number (like the mean).

The Explained SS tells us how much of the variation in the dependent variable our model explained.

Explained SS = Σ(Y-Hat – mean of Y)2.

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:

Residual Sum of Squares = Σ e2

4) What is Gini -impurity index?

The Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution.
The Gini index measures the area between the Lorenz curve and the hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line.
A Gini index of zero represents perfect equality and 100, perfect inequality.

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

6) What is an ensemble technique in machine learning?

In the world of Statistics and Machine Learning, Ensemble learning techniques attempt to make the performance of the predictive models better by improving their accuracy. Ensemble Learning is a process using which multiple machine learning models (such as classifiers) are strategically constructed to solve a particular problem.
Although there are several types of Ensemble learning methods, the following three are the most-used ones in the industry:

Bagging based Ensemble learning

Boosting-based Ensemble learning

Voting based Ensemble learning

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

- While they are built independently for Bagging, Boosting tries to add new models that do well where previous models fail.
- Only Boosting determines weights for the data to tip the scales in favor of the most difficult cases.
- It is an equally weighted average for Bagging and a weighted average for Boosting, more weight to those with better performance on training data.
- Only Boosting tries to reduce bias. On the other hand, Bagging may solve the over-fitting problem, while Boosting can increase it.

8) what is out-of-bag error in random forests?

Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging) to sub-sample data samples used for training.

9) What is K-fold cross-validation?

Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

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

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned .The same kind of machine learning model can require different constraints, weights or learning rates to generalize different data patterns. These measures are called hyperparameters, and have to be tuned so that the model can optimally solve the machine learning problem. Hyperparameter optimization finds a tuple of hyperparameters that yields an optimal model which minimizes a predefined loss function given independent data.The objective function takes a tuple of hyperparameters and returns the associated loss. Cross-validation is often used to estimate this generalization performance.

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

When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error.

12) What is bias-variance trade off in machine learning?

Bias is the tendancy of an estimator to pick a model for the data that is not structurally correct. A biased estimator is one that makes incorrect assumptions on the model level about the dataset. For example, suppose that we use a linear regression model on a cubic function. This model will be biased: it will structurally underestimate the true values in the dataset, always, no matter how many points we use.

Variance is error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting)

As we increase the complexity of our model, we can see a reduction in error due to lower bias in the model. However , this is only happen until a particular point. As we continue to make our model more complex, we end up overfitting our model and hence our model will start suffering from high variance.

13) What is the need of regularization in machine learning?

This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting. A simple relation for linear regression looks like this.

14) Differentiate between Adaboost and Gradient Boosting ?

In simple terms, GB works with residual errors based on decision tree. And Adaboost works with combination of weak learners.

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

Logistic regression has traditionally been used to come up with a hyperplane that separates the feature space into classes. But if we suspect that the decision boundary is nonlinear we may get better results by attempting some nonlinear functional forms for the logit function.