# ***AnswerSheet 1***

## Ans 1

R-squared is a better measure of goodness of fit model in regression. Because we feel that large negative residuals (i.e. points far below the line) are as bad as large positive ones (i.e., points that are high above the line). By squaring the residual values, we treat positive and negative discrepancies in the same way.

## Ans 2

**TSS** ( The Total Sum of Square)describe as the dispersion of observed variables around the mean, or the variance. So, the goal of TSS is to measure the total variability of the dataset.

Total SS = $\Sigma(Yi - \text{mean of } Y)^2$.

Where Sigma (Σ) is a mathematical term for summation or "adding up." It's telling us to add up all the possible results from the rest of the equation.

**ESS (**The Explained Sum of Squares) 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.

**RSS** (The Residual Sum of Squares) 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.

Residuals Sum of Squares= $\Sigma$ e2

**Ans 3**

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.

The need of regularization for fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

**Ans 4**

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

If all the elements belong to a single class, then it can be called pure. The degree of Gini Index varies between 0 and 1,

where,
'0' denotes that all elements belong to a certain class or there exists only one class (pure)
'1' denotes that the elements are randomly distributed across various classes (impure).

A Gini Index of '0.5 'denotes equally distributed elements into some classes

**Ans 5**

Yes, unregularized decision-trees prone to overfitting because decisions- 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.

**Ans6**

Ensemble techniques are 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

**Ans 7**

| **Bagging** | **Boosting** |
|---|---|
| Various training data subsets are randomly drawn with replacement from the whole training dataset. | Each new subset contains the components that were misclassified by previous models. |
| Bagging attempts to tackle the over-fitting issue. | Boosting tries to reduce bias. |
| Objective to decrease variance, not bias. | Objective to decrease bias, not variance. |
| If the classifier is unstable (high variance), then we need to apply bagging. | If the classifier is steady and straightforward (high bias), then we need to apply boosting. |
| Every model is constructed independently. | New models are affected by the performance of the previously developed model. |
| Every model receives an equal weight. | Models are weighted by their performance. |

| | |
|---|---|
| It is the easiest way of connecting predictions that belong to the same type. | It is a way of connecting predictions that belong to the different types. |

**Ans 8**

An error estimate is made for cases that were not used when constructing the tree. This is called an out-of-bag(OOB) . Out-of-bag(OOB)error estimate mentioned as a percentage. The decision trees are prone to overfitting, and this is the main drawback of it.

**Ans 9**

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.

**Ans 10**

Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. The prefix 'hyper_' suggests that they are 'top-level' parameters that control the learning process and the model parameters that result from it.

**Ans 11**

A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution, whereas a learning rate that is too small can cause the process to get stuck.

## Ans12

Logistic regression is indeed non linear in terms of Odds and Probability, however it is linear in terms of Log Odds.

The target label has no linear correlation with the features. In such cases, logistic regression (or linear regression for regression problems) can't predict targets with good accuracy (even on the training data).

## Ans 13

### Adaboost

AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

### Gradient Boosting

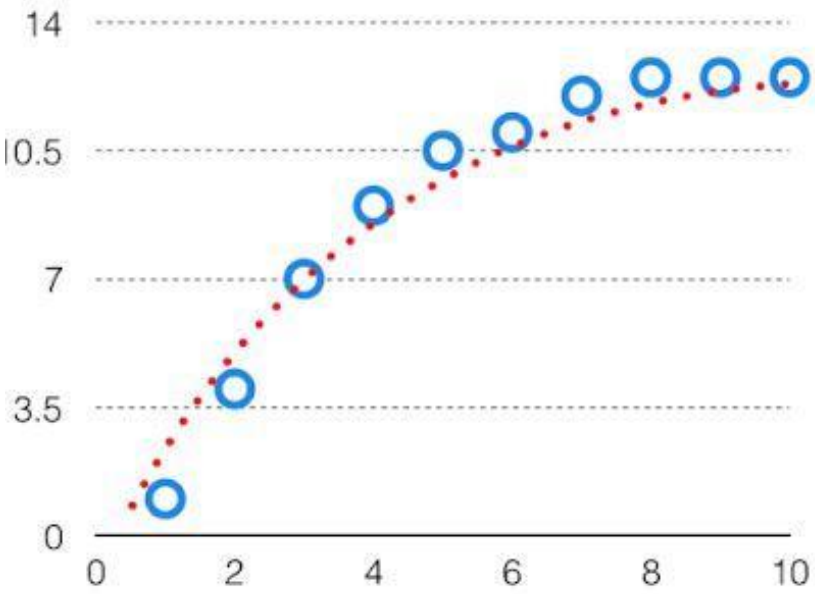
Gradient Boost is a robust machine learning algorithm made up of Gradient descent and Boosting. The word 'gradient' implies that we can have two or more derivatives of the same function. Gradient Boosting has three main components: additive model, loss function and a weak learner.

## Ans 14

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.

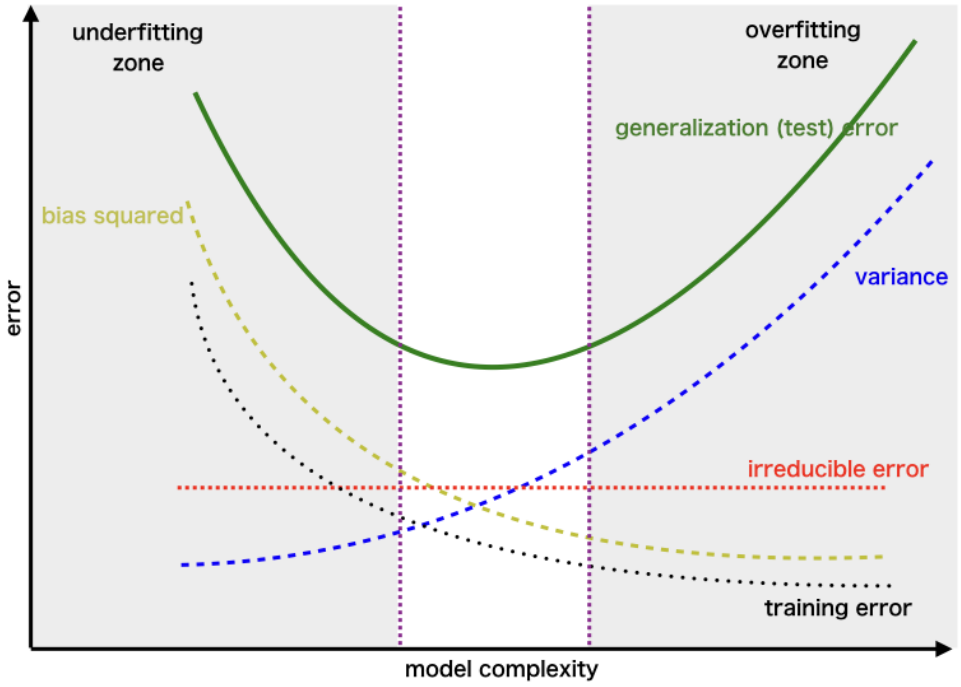
If the algorithm is too simple (hypothesis with linear eq.) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex ( hypothesis with high degree eq.) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time. For the graph, the perfect tradeoff will be like.




The best fit will be given by hypothesis on the tradeoff point.

The error to complexity graph to show trade-off is given as




This is referred to as the best point chosen for the training of the algorithm which gives low error in training as well as testing data.

**Ans 15**

**Linear Kernal**:

The linear kernel is typically used on data sets with large amounts of features as increasing the dimensionality on these data set does not necessarily improve separability. Text classification is a typical example of this kind of 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.

## POLYNOMIAL KERNAL:

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

## *ANSWERSHEET 1 (STATISTICS)*

Ans1 (d) Expected

Ans2(C) Frequencies

Ans3(c) 6

Ans4(b) Chisqared distribution

Ans5(c) c) F Distribution

Ans 6 (b) Hypothesis

Ans7(a) Null Hypothesis

Ans8 (a) Two tailed

Ans9(b) Research Hypothesis

Ans 10(a) np