Machine learning Worksheet 5- by sukhpal singh

1. Answer - R2 it represents the proportion of the variance in our data which is explained by our model; the closer to one, the better the fit. The residual sum of squares (RSS) is the sum of the squared distances between actual versus predicted values:

   RSS=Σi=1 to n [(yi – y'i)^2]

   Where yi is a given data point and y'i is the fitted value for yi. The actual number we get depends largely on the scale of our response variable. Taken alone, the RSS isn't so informative. Therefore, R2 is a better measure.
2. Answer - The residual sum of squares (RSS) is the sum of the squared distances between actual versus predicted values:

   RSS=Σi=1 to n [(yi – y'i)^2]

   ESS: The explained sum of squares (ESS) 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.

   ESS=Σi=1 to n [(y'i – ymean)^2]

   TSS: Total sum of squares (TSS) = explained sum of squares (ESS)+ residual sum of squares (RSS).

   TSS = Σi=1 to n [(y'i – ymean)^2] + Σi=1 to n [(yi – y'i)^2]

   The relation between the above 3 could be linearly expressed as:

   TSS = RSS + ESS

3. Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting. Regularisation is a penalty faced in case of regressions. Regularisation constraints or shrinks the coefficient towards zero. This means that this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting. Regularisation significantly reduces the variance of the model, without substantial increase in its bias

4. Gini index or Gini impurity measures the probability of a particular variable to be wrongly classified when chosen randomly. This measure is calculated where the modelling contains Tree Algorithms like Decision Trees or random forest. If we have C total classes and p(i) is the probability of picking a data point with class i, then the Gini Impurity is calculated as

$$G=i=1\Sigma C\ p(i)*(1-p(i))$$

Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure.

5. Yes, unregularized decision trees are prone to overfitting. Decision 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 samples of events that meet the previous assumptions. This small sample could lead to unsound conclusions. But unlike other algorithms, the decision tree does not use regularisation to fight against overfitting. Instead it uses pruning. There are mainly two types of pruning performed. Pre-pruning that stops growing the tree earlier, before it perfectly classifies the training set. Post-pruning allows the tree to perfectly classify the training set, and then post prune the tree.
6. Ensemble techniques combine the decisions from multiple models to improve the overall performance. Bagging and Boosting are two of the most used techniques in machine learning.
7. Bagging is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average. Boosting 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.
8. Out-of-bag 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.
9. In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 sub samples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation.
10. While defining the parameters, often the default values are not the ones that give the best result. In machine learning, hyper parameter optimization or tuning is the problem of choosing a set of optimal hyper parameters for a learning algorithm. A hyper parameter is a parameter

whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

11. A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution.
12. No, logistic regression only forms linear decision surface. Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries
13. Gradient Boosting is a generic algorithm to find approximate solutions to the additive modelling problem, while AdaBoost be a special case with a particular loss function. Hence, Gradient Boosting is much more flexible. On the other hand, AdaBoost can be interpreted from a much more intuitive perspective and can be implemented without the reference to gradients by reweighting the training samples based on classifications from previous learners.
14. If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and under fitting the data.
15. 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 many Features in a particular Data Set. 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 In the polynomial kernel, we simply calculate the dot product by increasing the power of the kernel.